2010 25th International Symposium on Defect and Fault Tolerance in VLSI Systems

Incorporating Heterogeneous Redundancy in a Nanoprocessor for Improved Yield and Performance

Priyamvada Vijayakumar, Pritish Narayanan, Israel Koren,
C. Mani Krishna and C. Andras Moritz
Department of Electrical and Computer Engineering
University of Massachusetts Amherst Amherst MA 01003 USA
$\{vijayakumar, koren\}@ecs.umass.edu$

Abstract

Emerging nano-device based architectures are expected to experience high defect rates associated with the manufacturing process. In this paper, we introduce a novel built-in heterogeneous fault-tolerance scheme, which incorporates redundant circuitry into the design to provide fault tolerance. A thorough analysis of the new scheme was carried out for various system level metrics. The implementation and analysis were carried out on WISP-0, a stream processor implemented on the Nanoscale Application Specific Integrated Circuits (NASIC) fabric. We show that intelligent assignment of redundancy levels and nanoscale-voting strategies across WISP-0 greatly improves area, effective yield and performance for the nano-processor. The new scheme outperforms homogeneous schemes for a defect range of 3% to 9.75% where the metric used is the product of performance and effective yield.

Index Terms

Heterogeneous, Homogeneous, NASICs, nanowires, Effective Yield, Performance

I. Introduction

Semiconductor nanowires [1][2], carbon nanotubes [3] and molecular devices [4] are some of the emerging nano-materials and devices proposed for novel computational fabrics. However, reliable manufacturing of nanoscale computational architectures is quite challenging. With the very high defect rates associated with nanoscale manufacturing, various strategies need to be applied for reliable manufacturing of the particular nano computational fabric. Different approaches such as built-in defect tolerance [5][6] and reconfiguration [7][8] have been explored for emerging nano-computational fabrics to achieve fault tolerance [9][10]. Built-in fault tolerance techniques do not need complex micro-nano interfacing, special reconfigurable devices or defect map extraction.

In most of the previously published built-in fault tolerant designs, redundancy has been uniformly applied across the entire nanoscale design. While this makes for simplicity, we show in this paper that, for certain defect levels, a heterogeneous application of redundancy has definite advantages in terms of the tradeoff between the additional yield achieved to the additional area and performance overhead by the fault-tolerance circuitry.

In a heterogeneous design, this would translate into different components being provided with differing levels of redundancy, with built-in techniques introduced intelligently based on component requirement and system level metrics.

In this paper, we explore various heterogeneous schemes and compare them against homogeneous application of redundancy. We show that careful assignment of redundancy levels and nanoscale voting strategies across a nano-processor design achieves a balance among area, effective yield and performance for the processor.This heterogeneous redundancy scheme is generic and can be incorporated into any design in nano-computational fabric. However, the evaluations here were carried out for a processor design based on the NASIC fabric [5][6].

The main contributions of the paper are: i) Introduction of new heterogeneous redundancy schemes for nanoscale computing fabrics; and ii) Detailed evaluation of key system-level metrics including effective yield, normalized performance and composite product metrics for the implemented schemes that quantify the benefits of heterogeneous redundancy schemes.

The rest of the paper is organized as follows: Section 2 provides an overview of the implementation methodology for heterogeneity in nano fabrics. The design and implementation opportunities of novel heterogeneous schemes in nano-computational fabrics are also discussed. Section 3 and Section 4 present the experiments conducted and results obtained. Section 5 concludes the paper.

II. Heterogeneity in Nano fabrics

A. Fabric and Design Overview

The new heterogeneous scheme is an implementation of built-in fault tolerance for designs in nano-computational fabric. In this paper, the heterogeneous schemes have been extensively explored on WISP-0 processor implemented on NASICs fabric.

NASICs [5][6][11][12] is a computational fabric based on a 2D grid of semiconductor nanowires with external dynamic controls for data streaming and cascading. WISP-0 is a stream processor with a five-stage pipelined streaming architecture using

1550-5774/10 $26.00 © 2010 IEEE
DOI 10.1109/DFT.2010.40

IEEE computer society

Table I
DELAY COMPARISON OF WISP-0 TILES

Tiles	Timing Delay (ps)
INC	47.42
ROM	55.22
DEC	13.52
IDE8	12.18
IDE14	13.22
MUX21	52.26
MUX41	56.77
ALU	220.49

eight nanotiles: Inc, Rom, Dec, Ide8, Ide14, Mux41, Mux21, ALU [5][6]. Adjacent nanotiles communicate using nanowires, with each nanotile being driven by surrounding microwires.

Before applying heterogeneous redundancy, WISP-0 was further balanced with respect to timing and delay. The nominal time delay of the various tiles of WISP0 is shown in Table 1. Since the pipeline frequency is determined by a small number of high fan-in data-paths, the delays are asymmetric. As seen in Table 1, the ALU is the slowest stage in WISP-0 and therefore, it was further partitioned into two stages to achieve a more balanced pipeline. The frequency of operation of the resulting 9 tiles has been then revaluated. As can be seen in Fig. 1(a), the frequency of operation of the stages has been made more balanced. The frequencies of operation plotted in Fig. 1(a) are for the nine tiles with no redundancy.

B. Opportunity for heterogeneous redundancy schemes

Self-assembly based manufacturing processes are expected to have high defect rates that are orders of magnitude larger than those for conventional CMOS. Typically, a 5%-10% device level defect rate is expected [13], which in conjunction with the high densities of nanoscale fabrics translates into 10^8-10^9defects per cm^2. Comprehensive fault-tolerance strategies are therefore necessary to achieve acceptable yield.

It should be noted that heterogeneous redundancy schemes can be applied to any design implemented on nano-computational fabric. The application of heterogeneous redundancy as against homogeneous redundancy, to any design would also help in preserving the density advantage of the nano-computational fabric by imposing the least possible area penalty. Thus, the promising feature of heterogeneous scheme is to deal with high defect rate while still keeping the density advantage of the chosen nano-fabric over CMOS technology. To investigate the application of heterogeneous redundancy schemes to achieve fault tolerance, architectural simulations were carried out on WISP-0, the test case that was chosen for this implementation.Without built-in redundancy, the yield of WISP-0 goes down to 0 when defect rate is 3

Different techniques have been proposed to incorporate fault tolerance in NASIC fabrics. For example, Biased Voting Scheme and FastTrack are explored in [14]. While the Biased Voting scheme leverages the property of NASIC circuits that logic ‘0’ faults are much less likely than logic ‘1’ faults, the ‘FastTrack’ scheme attempts to leverage the fact that path delays may differ significantly. More information regarding this scheme has been provided in the later part of this paper. These two techniques were developed targeting various manufacturing criteria and system level requirements.

Careful inspection of the timing profile of the WISP-0 architecture (see Fig. 1(a)) reveals the opportunity of applying heterogeneous redundancy by introducing higher levels of redundancy into the faster tiles. Appling more redundancy to faster tiles generally entails a lower performance penalty since they have a larger inherent time slack. Hence, rather than having uniform redundancy, it may be beneficial to apply an asymmetric or heterogeneous scheme. Simulations were run on individual tiles to obtain the timing profile of each tile after the introduction of redundancy. The timing profiles of the tiles being used in the heterogeneous scheme, with some tiles being duplicated (2w, i.e., two way redundancy) and others being triplicated (3w) are shown in Fig. 1(b). In these cases, the timing slack available in faster designs is taken advantage of to implement a higher level of redundancy. This implies that the performance of the overall system does not degrade due to the higher levels of redundancy. It can be seen that the performance penalty due to the introduction of redundancy is utilized by the heterogeneous scheme to bridge the differences in the timing profile of the various units.

III. EXPERIMENTS

This section describes different heterogeneous fault tolerance schemes for the NASIC fabric and quantifies the resulting effective yield, performance and other system level metrics.

A. Fault Model

A generic fault model with uniform distribution of defects has been assumed. Defects in NASIC fabrics would depend on the manufacturing pathway used. One possible manufacturing pathway has been outlined in [11]. In it, stuck-on transistors are

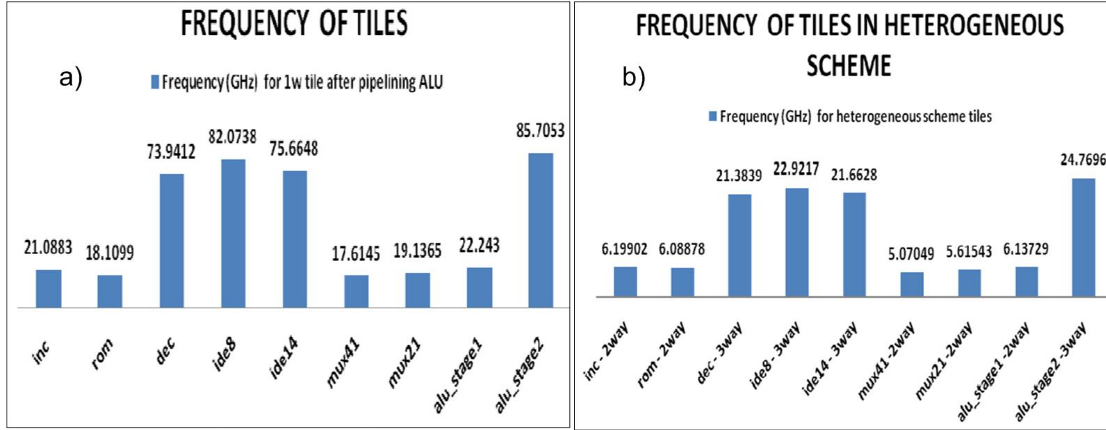

Figure 1. Maximum operating frequency of WISP-0 tiles a) after ALU pipelining and b) with heterogeneous scheme and ALU pipelining

the most prevalent type of defects due to the ion implantation and metallization processes involved. Reliable manufacturing of nanowires up to a few microns in length has been demonstrated [1][2], so the frequency of broken nanowires is assumed to be negligible. Given the logic style and prevalent defect types, it is expected that high fan-in tiles are less likely to produce faulty '0's. A nanowire output may evaluate to zero if all devices are turned on. So in a high fan-in NAND logic, even if only one of the devices is correctly turned off, the combination of logic and circuit style would automatically mask stuck-on defects. Consequently, high fan-in gates are expected to require a lower level of redundancy than low fan-in gates.

B. Simulation Setup

A custom designed simulator called FTSIM was used to run the simulations [14]. The inputs to the simulator are i) the NASIC circuit to be analyzed, ii) the gate timing characterization file, and, iii) the fault model. FTSIM is capable of simulating any tile designed on the NASIC fabric and it simulates the working of the circuit for the number of cycles specified. The simulator can also inject various types of defects into the circuit and identify their impact on the logical functioning of the circuit. The user specifies the defect and the number of defect patterns to be injected. FTSIM simulates the system with random different defect patterns and outputs the yield.

Timing faults can also be detected by the simulator. Delay characterization of NASIC circuits was done using HSPICE [15] and the data incorporated into the simulator. For each applied test pattern, FTSIM checks whether a timing fault occurs. For each run, the fastest operating frequency that produces the correct output is determined.

We have used the following three metrics to capture the impact of the added redundancy on performance and yield: performance, effective yield, and the normalized performance * effective yield product (PEY) for defect levels rates up to 15% [13].

The normalized performance represents the average frequency across all the simulations, which is then normalized to the mean operating frequency for the slowest technique. This metric hence captures the effective performance improvement of a technique as compared to the slowest scheme.

Effective yield is defined as (Overall Yield)*(Area of no redundant design/ Area of redundant design). This metric takes into account the tradeoff between yield and area overhead and represents the number of functional chips obtained from a given area.

The PEY product attempts to encapsulate the above two metrics and hence can help us in selecting a scheme that provides a good tradeoff between the two objectives of performance and effective yield. It is the product of the normalized performance with the effective yield which gives us an idea on the performance cost of the incorporated redundancy. It does not only consider the area overhead but also the performance penalty suffered by the architecture due to the incorporation of redundancy.

For a given defect rate, 1,000 trial runs with different defect maps and circuit delays were executed to achieve stability and sufficiently accurate estimation of the performance distribution and effective yield.

C. Redundancy techniques : Nomenclature and scheme conventions

The various redundancy techniques explored and analyzed are as follows:

1) Homogeneous redundancy: This is used as a baseline against which to compare more tailored techniques. As the term implies, homogeneous redundancy involves providing the same level of redundancy to all tiles. If a tile is replicated n times, we represent this scheme by "*nw*". Thus, duplication and triplication would be represented by *2w* and *3w*, respectively.

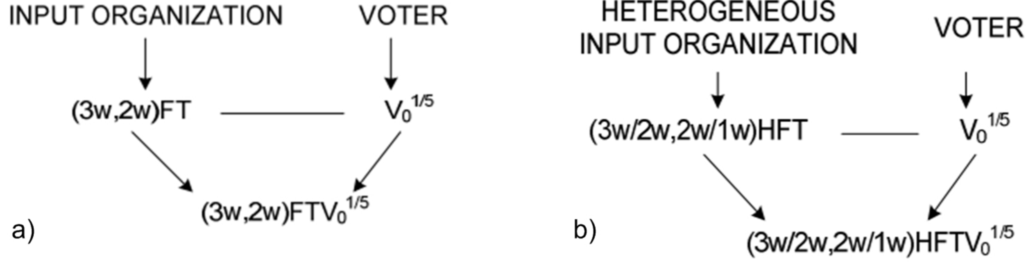

Figure 2. Notations for built-in fault tolerance schemes a) FastTrack with biased voters and unbalanced inputs b) Heterogeneity in conjunction with FastTrack

2) FastTrack redundancy: The FastTrack scheme is based on the following observation: i) some inputs (in some of the tiles) arrive sooner than others, ii) it is a property of the NASIC circuit that logic '0' faults are considerably less likely than logic '1' related faults. Thus, the voters used in this scheme are biased toward zero. Here, a voter denoted by $V_0{}^{2/5}$ indicates that it is biased to '0' and requires only 2 of the 5 inputs to be '0' to produce a result of '0'. This is in contrast to a majority voter where at least 3 out of the 5 inputs are required to be zero in order for the voter output to be zero. Other nano-computing fabrics may require different biasing schemes based on the underlying fault models.

Leveraging these asymmetric delay paths (resulting from some inputs being faster than the other) combined with biased voting schemes results in a redundancy scheme with better performance but at the cost of a lower effective yield. The notation used for FastTrack schemes indicates what input redundancy levels are combined with a particular type of a biased voter (see Fig. 2(a))[14]. For example, (3w,2w)FT$V_0{}^{2/5}$means that the architecture includes two sets of pipelined stages; the first set consists entirely of 3-way redundant tiles and the second set consists entirely of 2-way redundant tiles, with voter biased to zero.

3) Heterogeneous redundancy: The heterogeneous redundancy scheme implies that the blocks have asymmetric redundancy. Thus, (3w/2w)H is a heterogeneous scheme with certain stages in the design employing 3-way redundancy and the rest 2-way redundancy. As explained earlier, faster tiles can be provided with greater redundancy without affecting the overall performance of the circuit. The speed of a tile is a key parameter in our decision as to the appropriate level of redundancy to be used.

4) Heterogeneous redundancy to FastTrack: Heterogeneous redundancy can also be incorporated into FastTrack schemes. For example a (3w/2w, 2w/1w)HFT$V_0{}^{2/5}$ means that two sets of pipelined stages exist, with the first set being (3w,2w)H and the second set (2w,1w)H in conjunction with zero-biased voters (see Fig. 2(b)). This scheme helps to gain greater performance benefits due to the application of heterogeneity to FastTrack.

In the next section, yield–area-performance tradeoffs are discussed for the above schemes, followed by the experiments and results.

IV. Results and Analysis

Architectural simulations were carried out by using the FTSIM simulator described in Section III.B.

A. Homogeneous vs. Heterogeneous redundancy

Comparison of 2w homogeneous, 3w homogeneous and (2w/3w)H schemes has been performed. Fig. 3(a) compares the effective yield of the heterogeneous scheme to that of two homogeneous redundancy schemes. The plot of effective yield can be divided into three regions.

Region I favors the implementation of the Homogeneous 2w scheme. At lower fault rates, less redundancy is sufficient to take care of the faults and obviously, less redundancy implies lower area overhead. Thus, a higher effective yield is obtained by the homogeneous 2w scheme in Region I.

Region III favors the homogeneous 3w scheme. This is the region of high fault rates where a greater fault tolerance is required. The high area overhead due to the implementation of homogeneous 3w scheme is justified for achieving a reasonable yield.

The heterogeneous schemes are most beneficial in Region II where the expected fault rates are in the range of 3% to 7%. The homogeneous 2w scheme fails in Region II as it cannot provide the redundancy required to combat the large number of faults in this region. The homogeneous 3w scheme fails as it has a too high area overhead. Hence the (3w/2w)H heterogeneous scheme wins in this region striking a balance between area overhead and yield.

Fig. 3(b) shows the mean processor performance for the heterogeneous and homogeneous schemes. A homogeneous 3-way redundancy scheme is the slowest of the three schemes considered due to the triplication of all signals and the increased fan-in. The heterogeneous scheme employs high levels of redundancy only in non-timing-critical portions of the design. Performance

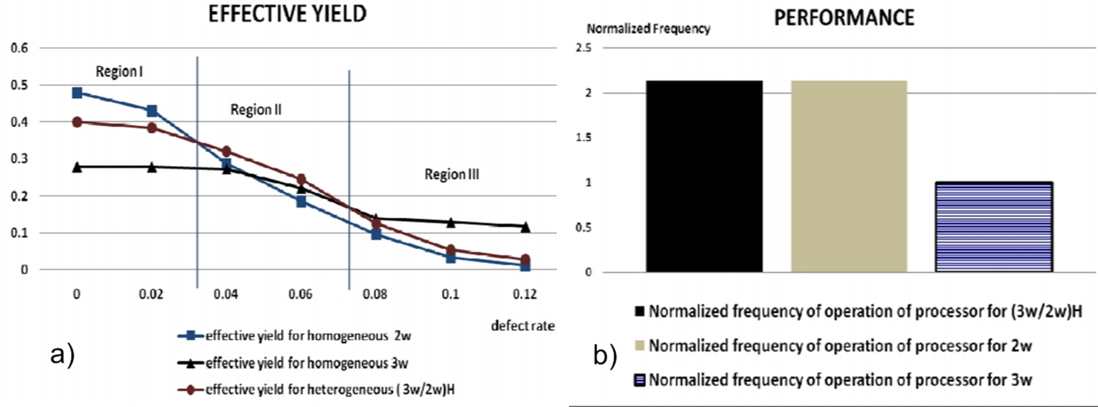

Figure 3. Simulation results showing tradeoffs between redundancy schemes (a) effective yield comparison (b) performance comparison

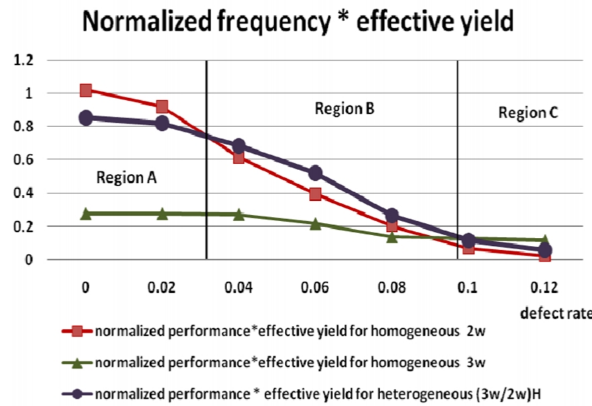

Figure 4. Comparison of performance * effective yield (PEY) product for different redundancy schemes

critical tiles employ only a 2-way redundancy. Therefore, the performance of the heterogeneous scheme is comparable to that of the 2-way redundancy homogeneous schemes (7.589GHz).

Analysis of the (2w/3w)H, and the homogeneous 2w and 3w schemes was also done with respect to the performance * effective yield product. Fig. 4 shows the performance * effective yield plot for the above schemes. The analysis of the plots leads us to the following conclusions. The 2w homogeneous scheme is best in Region A (up to 3% defect rate). This is identical to the effective yield case since the performance of 2w and (3w/2w)H schemes is identical. Also both 2w and (3w/2w)H schemes have at least 4X improvement over the 3w scheme in this region due to better performance and effective yields.

The heterogeneous scheme provides best results in Region B. Furthermore, the tradeoff point between the heterogeneous and 3w schemes is shifted further to the right (9.75%) due to the performance trends. This implies that when considering both the effective yield and the performance, the heterogeneous schemes are the best across a wider range (3%-9.75%) of defect rates.

B. Heterogeneous redundancy applied to FastTrack

The primary purpose of the FastTrack technique is to improve performance by exploiting the asymmetry in the various path delays. It can be seen from Fig. 5(b) that the performance of (3w,2w)$FTV_0{}^{1/5}$ is the same as that of the 2w homogeneous scheme ascertaining the performance benefit.

This opens a new avenue for introducing heterogeneous scheme in FastTrack. This can yield a redundancy technique that would give us the highest performance benefit. It can be seen in Fig. 5(b) that (3w/2w,2w/1w)$HFTV_0{}^{1/5}$gives us about a 3X performance benefit compared to (3w,2w)$FTV_0{}^{1/5}$. It should be noted that such a large performance benefit comes at the cost of a lower effective yield. Hence, the FastTrack schemes are recommended only when the performance of the processor is the most critical requirement. It can be seen from Fig. 5(a) that the incorporation of heterogeneity into FastTrack suffers from low effective yield.

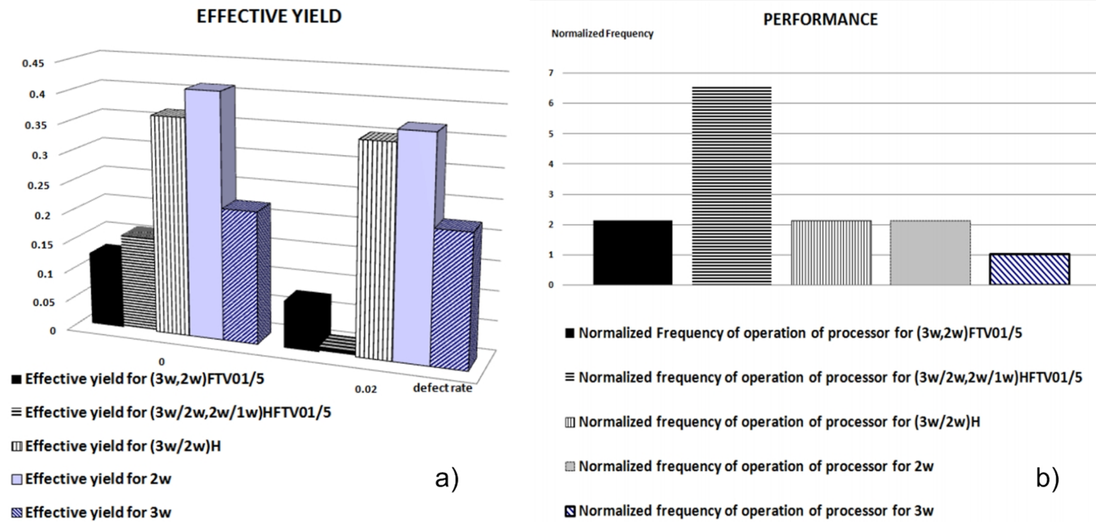

Figure 5. Results for redundancy and Fast Track Schemes a) Effective Yield and b) Performance

V. Conclusions

In this paper, we have studied the application of heterogeneous redundancy to a nanoarchitecture.While heterogeneous schemes were explored here for the NASIC fabric, the principle of heterogeneity is applicable to other nano-computing fabrics as well. The implementation was carried out on WISP-0, a stream processor implemented on a 2D Nanowire NASIC fabric. The scheme was carefully applied based on component requirements and system level metrics. The timing profile of the WISP-0 architecture was studied and the implementation of the heterogeneous scheme was carried out by introducing higher levels of redundancy into the faster tiles.

Intelligent application of redundancy to obtain greater yield and performance benefits was achieved by the implementation of the heterogeneous scheme. The (3w/2w)H scheme was further shown to be the best across a wide range (3%-9.75%) of defect rates, when considering both the effective yield and the performance. Greater performance benefits can be obtained by the incorporation of this scheme into the FastTrack technique. Thus, with appropriate nano-fabric architectural design and built-in heterogeneous fault tolerance it is possible to achieve higher yield and performance benefits for a given nano-computational fabric design implementation.

Acknowledgment

This work was supported in part by the Center for Hierarchical Manufacturing (CHM) at UMass Amherst, and NSF awards CCR:0105516, NER:0508382, CCR:051066 and CCF-0915612.

References

[1] W. Lu and C. M. Lieber, "Semiconductor nanowires," *Journal of Physics D: Applied Physics*, vol. 39, no. 21, pp. R387–R406, 2006.

[2] Y. Cui, X. Duan, J. Hu, and C. M. Lieber, "Doping and electrical transport in silicon nanowires," *The Journal of Physical Chemistry B*, vol. 104, no. 22, pp. 5213–5216, Jun. 2000.

[3] Z. Chen, J. Appenzeller, Y. Lin, J. Sippel-Oakley, A. G. Rinzler, J. Tang, S. J. Wind, P. M. Solomon, and P. Avouris, "An integrated logic circuit assembled on a single carbon nanotube," *Science*, vol. 311, no. 5768, p. 1735, Mar. 2006.

[4] C. P. Collier, E. W. Wong, M. Belohradsk, F. M. Raymo, J. F. Stoddart, P. J. Kuekes, R. S. Williams, and J. R. Heath, "Electronically configurable Molecular-Based logic gates," *Science*, vol. 285, no. 5426, pp. 391–394, Jul. 1999.

[5] T. Wang, P. Narayanan, and C. A. Moritz, "Heterogeneous Two-Level logic and its density and fault tolerance implications in nanoscale fabrics," *IEEE Transactions on Nanotechnology*, vol. 8, no. 1, pp. 22–30, 2009.

[6] C. Moritz, T. Wang, P. Narayanan, M. Leuchtenburg, Y. Guo, C. Dezan, and M. Bennaser, "Fault-tolerant nanoscale processors on semiconductor nanowire grids," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 11, pp. 2422 –2437, nov. 2007.

[7] D. B. Strukov and K. K. Likharev, "Reconfigurable hybrid CMOS/Nanodevice circuits for image processing," *IEEE Transactions on Nanotechnology*, vol. 6, pp. 696–710, Nov. 2007.

[8] G. S. Snider and R. S. Williams, "Nano/CMOS architectures using a field-programmable nanowire interconnect," *Nanotechnology*, vol. 18, no. 3, p. 035204, 2007.

[9] Y. Su and W. Rao, "Defect-Tolerant logic mapping on nanoscale crossbar architectures and yield analysis," in *Proceedings of the 2009 24th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*. IEEE Computer Society, 2009, pp. 322–330.

[10] Y. Dotan, N. Levison, R. Avidan, and D. Lilja, "History index of correct computation for fault-tolerant nano-computing," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 7, pp. 943 –952, july 2009.

[11] P. Narayanan, K. W. Park, C. O. Chui, and C. Moritz, "Manufacturing pathway and associated challenges for nanoscale computational systems," in *Nanotechnology, 2009. IEEE-NANO 2009. 9th IEEE Conference on*, 26-30 2009, pp. 119 –122.

[12] P. Narayanan, M. Leuchtenburg, T. Wang, and C. A. Moritz, "CMOS control enabled Single-Type FET NASIC," in *Proceedings of the 2008 IEEE Computer Society Annual Symposium on VLSI*. IEEE Computer Society, 2008, pp. 191–196.
[13] Y. Huang, X. Duan, Y. Cui, L. J. Lauhon, K. Kim, and C. M. Lieber, "Logic gates and computation from assembled nanowire building blocks," *Science*, vol. 294, no. 5545, pp. 1313–1317, Nov. 2001.
[14] P. Narayanan, M. Leuchtenburg, J. Kina, P. Joshi, P. Panchapakeshan, C. O. Chui, and C. A. Moritz, "Variability in nanoscale architectures: Bottom-up integrated analysis and mitigation," 2010, submitted for publication.
[15] "Hspice user's manual," 2007, synopsys, Inc.