Incorporating Heterogeneous Redundancy in a Nanoprocessor for Improved Yield and Performance

Priyamvada Vijayakumar, Pritish Narayanan, Israel Koren, C. Mani Krishna,and Csaba Andras Moritz
Department of Electrical and Computer Engineering
University of Massachusetts
Amherst, MA, USA
{vijayakumar, pnarayan, koren, krishna, andras}@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.**

Keywords: Heterogeneous, Homogeneous, NASICs, nanowires, Effective Yield, Performance.

I. INTRODUCTION

Semiconductor nanowires [1][2], carbon nanotubes [3] and molecular devices [4][5] 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 [6][7] and reconfiguration [8][9] have been explored for emerging nano-computational fabrics to achieve fault tolerance [10][11]. Built-in fault tolerance is a promising direction since it does 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 consumed 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 new heterogeneous redundancy scheme is generic and can be implemented on any design in nano-computational fabric. However, the evaluations here were carried out for a processor design based on the NASIC fabric [6][7].

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 II 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 III and Section IV present the experiments conducted and results obtained. Section V concludes the paper.

II. HETEROGENEITY IN NANO FABRICS

A. *Fabric and Design Overview*

The principle of heterogeneity can be applied to any nano-computational fabric. In this paper, the heterogeneous schemes have been extensively explored on WISP-0 processor implemented on NASICs fabric.

NASICs [6][7][12][13] is a computational fabric based on a 2D grid of semiconductor nanowires with external dynamic control for data streaming and cascading. WISP-0 is a stream processor with a five-stage pipelined streaming architecture using five nanotiles: Program Counter - PC, Read Only Memory - ROM, Decoder - DEC, Register File - RF and Arithmetic Logic Unit - ALU [6][7]. Adjacent nanotiles communicate using nanowires, with each nanotile being driven by surrounding microwires.

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

Before applying heterogeneous redundancy, WISP-0 wasfurther balanced with respect to timing and delay. The nominal time delay of the various pipelined stages 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 nine stages has been then re-evaluated. As can be seen in Fig.1, the frequency of operation of the stages has been made more balanced. The frequencies of operation plotted in Fig.1 are for the nine stages with no redundancy.

TABLE I. DELAY COMPARISON OF WISP-0 TILES

Tiles	*Timing Delay (ps)*
Inc	47.419
Rom	55.2182
Dec	13.5242
Ide8	12.184
Ide14	13.216
Mux41	56.7714
Mux21	52.256
Alu	220.49

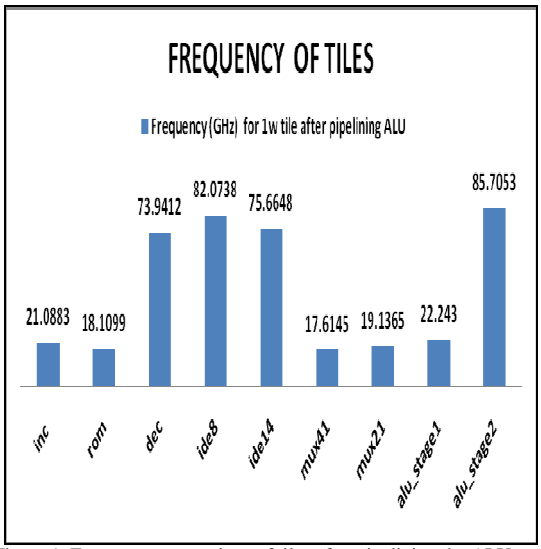

Figure 1. Frequency comparison of tiles after pipelining the ALU

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 conventional CMOS. Typically, a 5%-10% device level defect rate is expected [14], which in conjunction with the high densities of nanoscale fabrics translates into 10^8-10^9 defects per cm^2. Comprehensive fault-tolerance strategies are therefore necessary to achieve acceptable yield.

It should be noted that the new heterogeneous scheme of redundancy 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.

Different techniques have been proposed to incorporate fault tolerance in NASIC fabrics. For example, Biased Voting Scheme and FastTrack have been explored in [15]. 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) reveals the opportunity of applying heterogeneous redundancy by introducing higher levels of redundancy into the faster tiles. Applying 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.2. 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 triplicated blocks. 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.

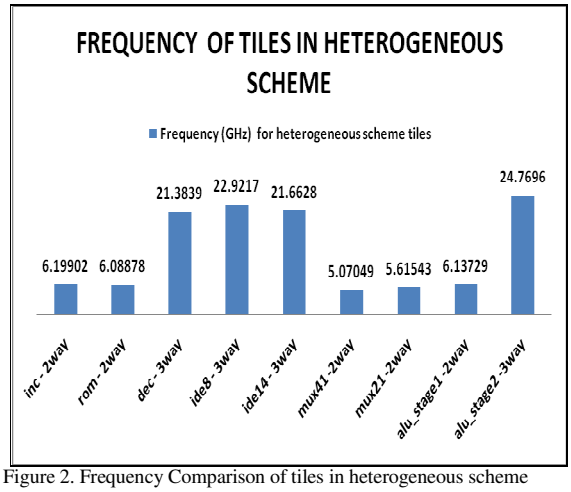

Figure 2. Frequency Comparison of tiles in heterogeneous scheme

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 [12]. In this pathway, stuck-on transistors are 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. 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 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 test for their impact on the logical functioning of the circuit.

Timing faults can also be detected by the simulator. Delay characterization of NASIC circuits was done using HSPICE [16] 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% [14].

The normalized performance represents the 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. This product is a metric that 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 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 2*w* and 3w, respectively.

2) *FastTrack redundancy*

The FastTrack scheme is based on the following observation: i) some inputs (in some of the blocks) 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. 3) [15]. For example, (3w,2w)$FTV_0^{2/5}$ 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)HFTV$_0^{2/5}$ 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.4) . 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.

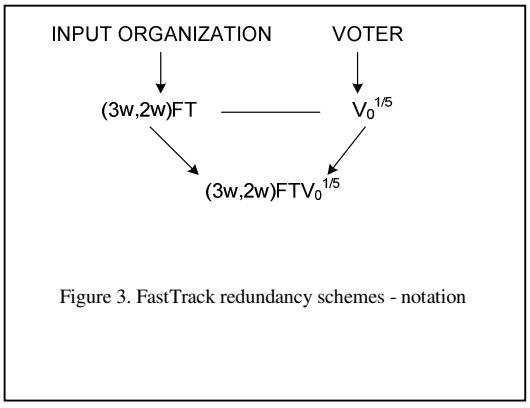

Figure 3. FastTrack redundancy schemes - notation

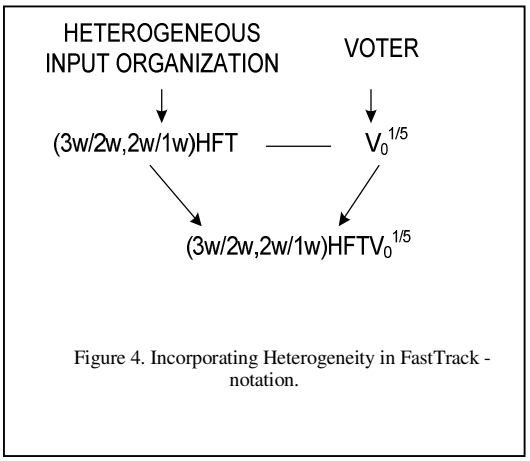

Figure 4. Incorporating Heterogeneity in FastTrack - notation.

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. 5 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 is less efficient in Region II as it cannot provide the redundancy required to combat the large number of faults in this region. The homogeneous 3w scheme is less efficient 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.

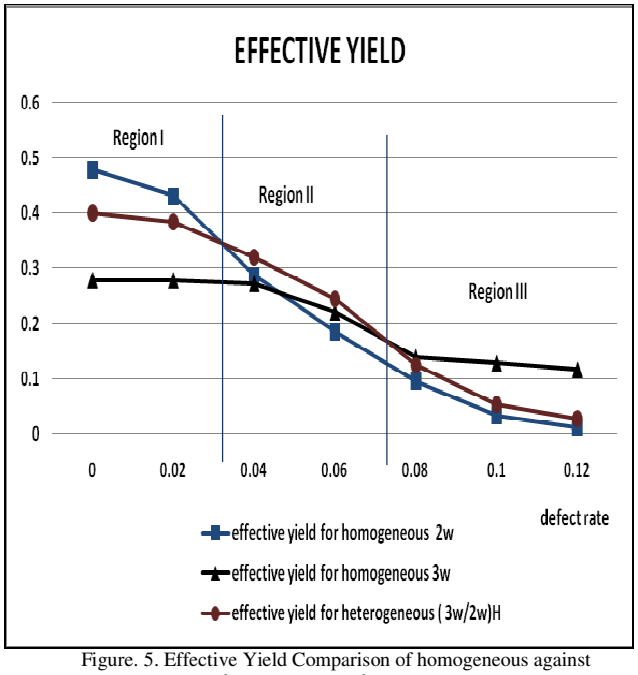

Figure. 5. Effective Yield Comparison of homogeneous against heterogeneous schemes

Fig. 6 shows the 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 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. 7 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. 8 that the performance of (3w,2w)$FTV_0^{1/5}$ is the same as that of the 2w homogeneous scheme ascertaining the performance benefit

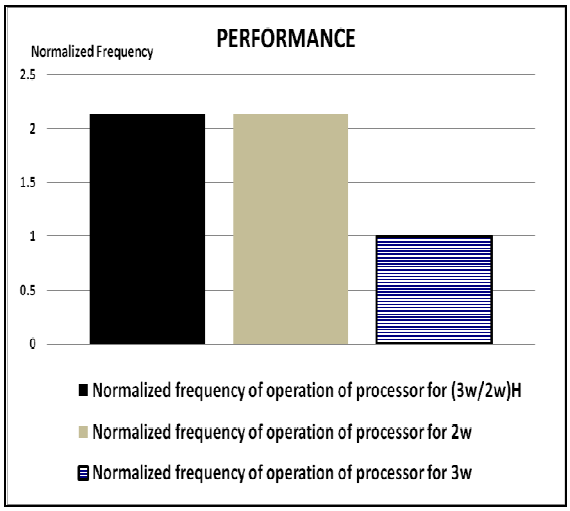

Figure. 6. Performance of homogeneous v/s heterogeneous schemes

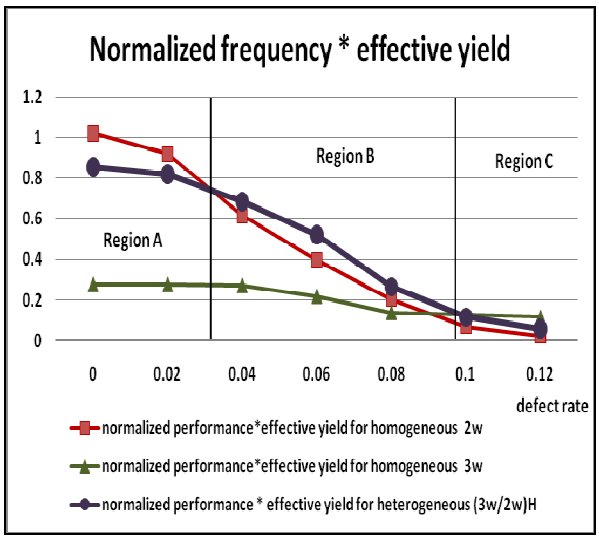

Figure. 7. PEY plot comparing homogeneous and heterogeneous schemes

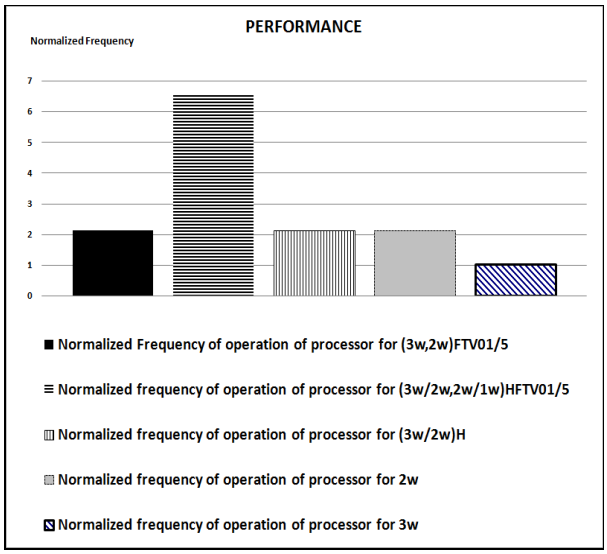

Figure 8. Performance of various redundancy schemes

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. 8 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. 9 that the incorporation of heterogeneity into FastTrack suffers from low effective yield.

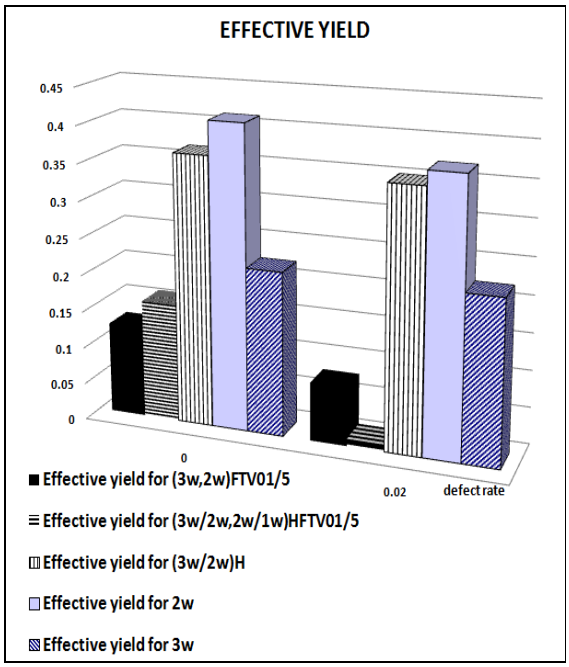

Figure 9.Comparison of effective yield for various schemes

V. CONCLUSIONS

In this paper, we have studied the application of heterogeneous redundancy to a nanoarchitecture. While heterogeneous schemes were extensively explored 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 schemes 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 schemes 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 schemes. 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 on a given nano-computational fabric design implementation.

REFERENCES

[1] W. Lu and C. M. Lieber, "Semiconductor Nanowires," *J. Phys. D: Appl. Physics,* vol. 39, pp. R387-R406, October 2006.

[2] Y. Cui, X. Duan, J. Hu, and C. M. Lieber, "Doping and Electrical Transport in Silicon Nanowires," *Journal of Physical Chemistry B*, vol. 104, pp. 5213-5216, May 2000.

[3] Z. Chen *et. al.*, "An Integrated Logic Circuit Assembled on a Single Carbon Nanotube," *Science*, vol. 311, p. 1735, Mar. 2006.

[4] C.P. Collier, E.W. Wong, M. Belohradský, 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, pp. 391-394, Jul. 1999.

[5] K.K. Likharev and D. B. Strukov, "CMOL: Devices, circuits, and architectures,"Introducing molecular electronics," *Lecture Notes Phys.*, vol. 680, pp. 447–477, 2005.

[6] T. Wang, P. Narayanan, and C. A. Moritz, "Heterogeneous 2-level Logic and its Density and Fault Tolerance Implications in Nanoscale Fabrics," *IEEE Trans. on Nanotechnology*, vol. 8, no. 1, pp. 22-30, January 2009.

[7] C. A. 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, pp. 2422-2437, 2007.

[8] D. B. Strukov and K. K. Likharev, "Reconfigurable Hybrid CMOS Devices for Image Processing," *IEEE Transactions on Nanotechnology*, vol. l6, pp. 696-710, November 2007.

[9] G. S. Snider and R. S. Williams, "Nano/CMOS architectures using a field-programmable nanowire interconnect," *Nanotechnology*, vol. 18, pp. 1-11, 2007.

[10] Y. Su and W. Rao, "Defect-tolerant Logic Mapping on Nanoscale Crossbar Architectures and Yield Analysis," *24th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, 2009.

[11] Y. Dotan, N. Levison, R. Avidan and D. J. Lilja,"History Index of Correct Computation for Fault-Tolerant Nano-Computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 7,July 2009.

[12] P. Narayanan, K. W. Park, C. O. Chui, and C. A. Moritz, "Manufacturing Pathway and Associated Challenges for Nanoscale Computational Systems," *IEEE Conference on Nanotechnology (NANO'09)*, Genoa, Italy, 2009.

[13] P. Narayanan, M. Leuchtenburg, T. Wang, and C. A. Moritz, "CMOS Control Enabled Single-Type FET NASIC," *Symposium on VLSI, 2008. ISVLSI '08. IEEE Computer Society Annual*, pp. 191-196, 2008.

[14] Y. Huang, X. Duan, Y. Cui, L. J. Lauhon, K. Y. Kim, and C. M. Lieber, "Logic Gates and Computation from Assembled Nanowire Building Blocks," *Science*, vol. 294, no. 5545, pp. 1313-1317, November 2001.

[15] 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," submitted for publication.

[16] *HSPICE User's Manual*, Meta-Software, Inc., Campbell, CA, 1992.