

# On Nanoelectronic Architectural Challenges and Solutions

Valeriu Beiu<sup>1</sup>, Ulrich Rückert<sup>2</sup>, Sandip Roy<sup>1</sup>, and Jabulani Nyathi<sup>1</sup>

Centres for Neural-Inspired Nano Architectures

<sup>1</sup>School of EE&CS, Washington State University, Pullman, WA 99164-2752, USA

<sup>2</sup>Heneinz Nixdorf Institute, University of Paderborn, 33102 Paderborn, Germany

**Abstract** — This paper discusses the many challenges in the design of future nano architectures that result from the use of nanoelectronic devices. The relations among these challenges are studied, and an unfortunately subjective relative ranking is proposed. Possible solutions are suggested.

**Index Terms** — Computer architectures, nanotechnology, semiconductor devices.

## I. INTRODUCTION

In the nanoera, which we have entered, the growing complexity of integrated circuits (ICs) turns difficult problems (e.g., power, reliability) into great challenges [1]. Also, communication problems lead to network-on-chip and force (partly) asynchronous solutions. Out of the many nanoelectronics challenges, we shall focus here on those for which an architectural approach could make a difference. A review of the nanoelectronic devices being investigated [1], [2] can be seen in Fig. 1. As can be seen, most of them are only at the level of single devices.

In this paper we: (i) identify many challenges amenable to architectural approaches; (ii) briefly detail some of these challenges; (iii) rank a selected set of challenges based on both first-order and second-order effects; and (iv) enumerate some of the possible architectural solutions.

## II. NANOELECTRONIC ARCHITECTURAL CHALLENGES

We believe that the following set of challenges is particularly amenable to architectural solutions [1]–[4]:

- power-heat P/H – have to be reduced (this also includes power delivery/distribution, heat removal, and dealing with hot spots);
- reliability REL – has to be increased through redundancy in space, or time, or both, but the redundancy factors should be small;
- testing TST – associated costs have to be reduced;
- connectivity CONN – has to be reduced both as overall length, and as number of connections;
- communication COMM – method has to be optimized;
- hybrid integration HYB – in the near term must be achieved, including mixed design and interfacing;
- logic and (en)coding L/C – must be optimized to reduce computations and/or communications (e.g., non-Boolean, error correction, spikes).

Other challenges which might be considered are algorithm improvement ALG (e.g., stochastic/probabilistic), and reduction of the design complexity DCOM (e.g., by reuse).

Our goal is to rank these challenges with respect to their importance, when considering not only their own intrinsic importance, but also how their solution would affect the other challenges. We use an Excel spreadsheet (see Fig. 2) to correlate all these challenges. The first column contains the challenges described before. Our first step was to assign an “importance” (percentage) to each of the challenges. These are shown in the second column. Our team has rated P/H most highly at 25%. The reminders of the challenges were assigned lower percentages, with all of them totaling 100%. Our next step was to fill the correlation matrix, which is shown in the remainder of the spreadsheet. It contains coefficients that reflect second-order effects, namely how a solution improving a challenge will affect the others. Let us take CONN as an example (i.e., the fourth column in the correlation matrix):

- Reducing CONN (number of connections and/or the overall length of the connections) should reduce P/H. This shows a positive influence. We have decided that the influence factor is ‘medium’.
- CONN influences REL in a complex manner. The overall number of wires (connections) of a reliable design should be about the same as for an unreliable one, so apparently enhancing CONN should not make

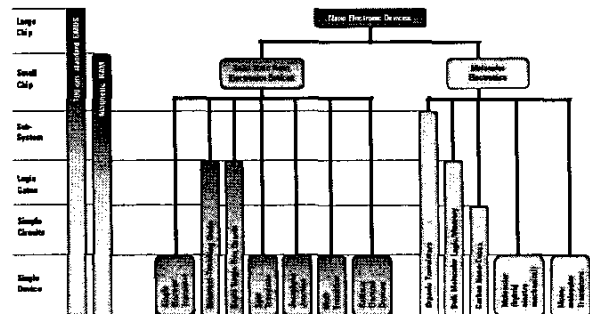


Fig. 1. The roadmap for nanotechnology [2] (see also [1]) presents many nano devices currently being investigated as an alternative to standard CMOS. As can be seen, for most of them only single devices have been realized and tested.

		%	25%	15%	12%	12%	12%	12%	12%		
			P/H	REL	TST	CONN	COMM	HYB	L/C	ALG	DCOM
Power / Heat	25%	P/H		-10%	-5%	15%	15%	10%	10%	10%	
Reliability	15%	REL	15%		5%	-5%		-10%	10%	5%	
Testing	12%	TST		25%		-10%	5%	-5%		5%	5%
Connectivity	12%	CONN	5%	-5%	-20%		5%		5%	5%	5%
Communication	12%	CoMM	5%	5%		-15%		-10%	5%	10%	5%
Hybrid	12%	HYB	5%								
Logic / Coding	12%	L/C	5%		5%	-5%				5%	
Algorithms		ALG				-5%	15%		-5%		5%
Design complexity		DCOM		-10%	-10%	-5%	10%	-10%	-10%	5%	
Overall			30%	15%	9%	9%	13%	9%	14%		

Fig. 2. Ranking nanoelectronic architectural challenges (i.e., challenges where architecture can make a difference).

REL any easier or harder. On a closer look, the total length of the wires might be increased in a redundant design (needed for enhanced REL), even if the number of connections is not. This explains the 'small' negative influence, since CONN would make redundant design slightly more difficult. [Remark: The problem is even more complex, since one should consider both faulty devices and faulty wires].

- CONN might affect TST in an adverse way, because fewer connections might make testing more difficult. Hence, a 'small-to-medium' negative influence.
- CONN clearly influences COMM, because reducing or shortening connections would make it more difficult to implement an optimal communication. Therefore, a negative 'medium' has been assigned.
- CONN does not seem to influence HYB, since the difficulty of hybrid design is not affected by the connectivity of its parts.
- CONN has a 'small' negative effect on L/C, as it might allow only for a sub-optimum encoding.

After repeating this process for all the columns in the correlation matrix, we have quantified small, small-to-medium, medium, medium-to-large, and large influences in a percentage scale as 5%, 10%, 15%, 20%, and 25%, respectively. Obviously this is subjective, but should be good enough to give us an idea of the overall importance of each challenge. This overall importance takes into account the importance initially assigned to the challenge (second column), together with the weighted sum of the second-order effects over all the other entries in each column. The results can be seen in the last row in Fig. 2. They show that: P/H becomes even more important (25→30%); REL maintains its importance (15→15%); L/C (12→14%) and COMM (12→13%) increase slightly; TST, CONN, and HYB are decreasing (12→9%).

### III. POSSIBLE ARCHITECTURAL SOLUTIONS

In this section we list many possible architectural solutions for the challenges described in Section II. We

also discuss three challenges in some more detail, and outline possible solutions. We have divided the solutions into three categories, with respect to their expected time of implementation:

- *Near term solutions* should include massively parallel, modular (cells, blocks); regular (grid processing, cellular arrays); locally connected (near-neighbor connections); higher functionality (multiple valued and threshold logic [5]); reconfigurable.
- *Medium term solutions* might include asynchronous (GALS = globally asynchronous locally synchronous); fault-tolerant (noise immune, rad-hard by design, redundant, self-testing, self-correcting); defect-tolerant (reconfigurable); adaptive (self-adaptive, self-organizing, evolvable); bio-inspired (complex functions, self-organizing, self-healing); nanophotonic (optical communication, e.g. GOLE = globally optical locally electrical); nanofluidic (e.g., for cooling); 3D interconnects; stochastic/probabilistic (algorithms, encoding, communication).
- *Long term solutions* envisaged are molecular and quantum computing, quantum-dot cellular automata, adiabatic/reversible computing, and bio-compatible.

Figure 3 presents a synthetic view of the four most important challenges identified in Section II (horizontal thrusts), and of some of the plausible architectures, in increasing order of their connectivity complexity (from left to right). The drawing also argues that both novel CAD tools, and models will be highly needed.

Our highest-ranking challenge is P/H. We believe that this should be addressed using a bottom-up approach, i.e. starting from the device level, and going all the way up to the system level. Sources of power dissipation include leakage currents, switching activity, spurious switching, operations that result in crowbarred outputs, and clock networks. Although scaling  $V_{DD}$  helps in meeting a low power budget, it requires  $V_{th}$  scaling to compensate for performance degradation. This causes an increase in leakage current. Addressing leakage power involves

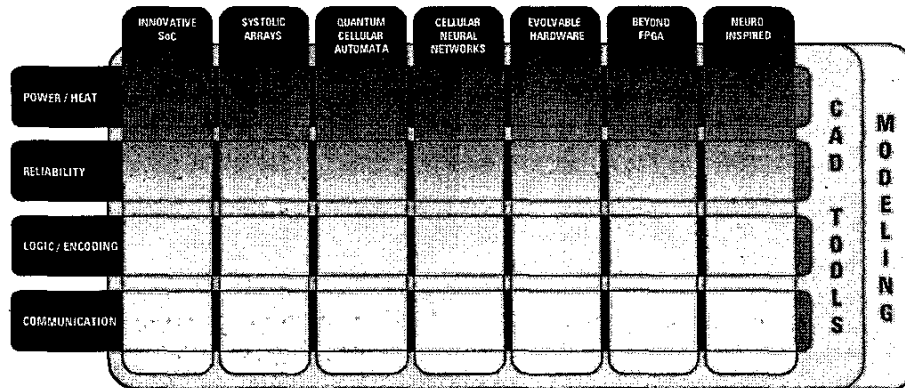


Fig. 3. Possible architectural solutions. Logic/encoding (also reconfigurability and asynchronicity) will be considered at all levels.

techniques like: *dynamic  $V_{th}$* , *stacked CMOS* (inserting leakage control transistors), or *multiple threshold voltage CMOS*. New circuit techniques, which are robust and scalable, should be used to achieve required design goals. As an example here, circuits operating at sub- $V_{th}$  with adaptive body bias (for dynamic  $V_{th}$  adjustments) have been shown to be a viable solution for ultra low-power systems that do not require high speeds. *These have only recently been investigated for medium to high-speed applications* [6]–[8]. Reconfigurability will certainly help in power aware schemes, allowing portions of the system that are not performing computations to go into sleep mode, or not to receive clocks (clock gating). The operations could even be made data-dependent, hence theoretically eliminating any unwanted switching activity (when there is no data to be acted upon). These approaches are not limited to computations, but can be applied to communication also [9]. For a large chip, like the Itanium 2, one third of the power budget is consumed in communication, while the demanding clock distribution consumes another third. Lowering clock frequencies to reduce dynamic power, while enhancing throughput, could be a viable option. A systematic use of asynchronous design styles could completely eliminate the clock, but might lead to increased communication. The architectural and micro-architectural level approach for lowering power would definitely have a significant impact. An advanced CAD tool should allow analyzing the many architectural tradeoffs when using various nano devices.

The expected higher probabilities of failures of both nano devices and interconnects, as well as higher sensitivity to noise and variations, will make it that fault- and defect-tolerance will have to be considered from the very early design phases. A well-known solution here is redundancy (with or without reconfigurability) either in space, or in time, or both. Trying to find the optimum

balance is a clear example of an architectural decision. Obviously, such a balance could be quite different for circuits made of different nano devices. First the theory has to be developed, followed by innovative CAD tools, which incorporate the models for different nano devices. Hence, it should be possible to identify a few very promising architectures. Currently, the theory, the CAD tools, and the models (to be used in these CAD tools) are just starting to be developed [6]. We expect that their integration should come about in the near future.

As the artist's rendition shows (Fig. 4), another challenge is communication, which could be (partly) solved by using optical interconnects. It is expected that GALS would be an easier solution for the near future, since it will not require optical wave-guides. Still, one could envisage GOLE, or a combined solution. The communication challenge will also be met at the protocol level. In developing on-chip communication protocols, we must keep in mind the following two special features of chip-level communication:

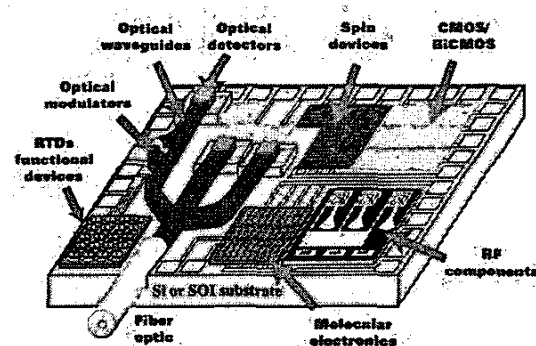


Fig. 4. In the near future it is expected that combinations of several devices will be used in a hybrid system-on-chip way (artistic rendition adapted from [2]).

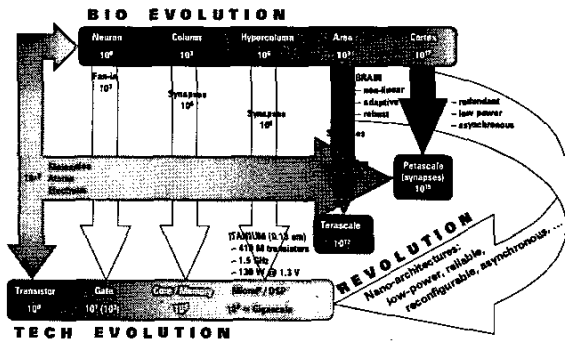


Fig. 5. Comparing the Bio and the Tech evolutions.

- *Simplicity* in the communication paradigm is of utmost importance, because of the difficulty of implementing complex communication protocols at nano scales, and because of the low-power requirement. Simplicity in this case refers to both the need for sparse but reliable communication, and the need for easy-to-implement protocols.
- The relatively *regular (grid) structure* of the communicating units is a special feature of such a system, that should be exploited to develop efficient protocols (but also for novel distributed circuit design which could take advantage of standing waves for achieving high speeds and low power).

These two features clearly motivate protocols without acknowledgement and with simple route selection, which are robust to faults and defects in the architecture.

#### IV. CONCLUSIONS

The building blocks for ICs and for the Brain are the same at nanoscale level: electrons, atoms, and molecules, but their evolutions have been radically different (see Fig. 5). The fact that *reliability, low-power, reconfigurability, as well as asynchronicity* are brought up so many times in recent conferences and articles, makes it compelling that the Brain should be an inspiration (at many different levels), suggesting that *future nano architectures could be neural inspired* (see Fig. 6, and [6]–[10]).

#### V. ACKNOWLEDGMENT

VB wishes to extend special thanks to: M. Rajashekharaiyah M. Sulieman, S. Tatapudi, and P. Upadhyaya (all from Washington State University) for fruitful discussions.

#### REFERENCES

[1] *The International Technology Roadmap for Semiconductors (ITRS)*, 2003. Available: <http://public.itrs.net/>

Fig. 6. A joint effort on nano architectures [10].

[2] R. Compañó, L. Molenkamp, and D.J. Paul (Eds.), *Technology Roadmap for Nanoelectronics*, European Commission, IST Programme, Future and Emerging Technologies, 2000. Available: <http://www.cordis.lu/espri/src/melna-rm.htm>

[3] V. Beiu, "The Next Generation of Neural Network Chips: Shaping the Hardware Solutions for the Next Millennium," *Special Session, Intl. Symp. Engineering of Intelligent Sys. EIS'98*, Tenerife, Spain, Feb. 1998.

[4] "Silicon Nanoelectronics and Beyond," *SCR-NNI Workshop*, Portland, Oct. 2003. <http://www.src.org/member/event/e002238/>

[5] V. Beiu, J. M. Quintana, and M. J. Avedillo, "Threshold Gates: Past, Present, and Future," *Special Session, Intl. Workshop Artif. & Natural Neural Networks IWANN'03*, Mahon, Spain, Jun. 2003.

[6] V. Beiu, and U. Rückert, "Neural Inspired Architectures for Nanoelectronics," *Special Workshop, Neural Inform. Proc. Sys. NIPS\*03*, Whistler, Canada, Dec. 2003. Available: [http://www.eecs.wsu.edu/~vbeiu/workshop\\_nips03/](http://www.eecs.wsu.edu/~vbeiu/workshop_nips03/)

[7] U. Rückert, and V. Beiu, "Brain Inspired Emerging Architectures: Design and Technological Challenges," *Special Session, Intl. Joint Conf. Neural Networks IJCNN'04*, Budapest, Hungary, Jul. 2004.

[8] V. Beiu, and U. Rückert (Eds.), *Emerging Brain-Inspired Nano-Architectures*, 2004, in progress.

[9] V. Beiu, M. Sulieman, and S. Roy, "Majority multiplexing: Practical applications to nanoscale computations and communication," *Proc. Application-specific Sys., Arch. & Processors ASAP'04*, Sep. 2004, to appear.

[10] Centers for Neural-Inspired Nano Architectures See <http://www.cnina.org/>