Wave Pipelining for Majority-based Beyond-CMOS Technologies

O. Zografos∗+, A. De Meester†, E. Testa‡, M. Soeken†,
P.-E. Gaillardon§, G. De Micheli‡, L. Amarú¶, P. Raghavan∗, F. Catthoor∗†, R. Lauwereins∗†

∗imec, Kapeldreef 75, B-3001 Leuven, Belgium
†KU Leuven, ESAT, B-3001 Leuven, Belgium
‡Integrated Systems Laboratory, EPFL, Switzerland
§Laboratory for Nanointegrated Systems, University of Utah, UT, USA
¶Synopsys Inc., Sunnyvale CA, USA

Abstract—The performance of some emerging nanotechnologies benefits from wave pipelining. The design of such circuits requires new models and algorithms. Thus we show how Majority-Inverter Graphs (MIG) can be used for this purpose and we extend the related optimization algorithms. The resulting designs have increased throughput, something that has traditionally been a weak point for the majority of non-charge-based technologies. We benchmark the algorithm on MIG netlists with three different technologies, Spin Wave Devices (SWD), Quantum-dot Cellular Automata (QCA), and NanoMagnetic Logic (NML). We find that the wave pipelined version of the netlists have an improvement in throughput over power of 23×, 13×, and 5× for SWD, QCA, and NML, respectively. In terms of throughput over area ratio, the improvement is 5×, 8×, and 3×, respectively.

I. INTRODUCTION

The continuous downsizing of the complementary metal oxide semiconductor (CMOS) technology, dictated by Moore’s Law [1], has enabled the semiconductor industry to higher density electronic circuits at reduced costs. However, this downsizing will reach its limits in the following decade [2], which has given rise to a need for logic components that can operate at high frequencies, be extremely compact and also consume ultra-low power [3]. The exploration and study of novel logic components has been a main research focus in the past decade [4], in pursuance of extending the semiconductor industry roadmap beyond the CMOS technology [4].

Beyond-CMOS device concepts include a wide variety of elements such as charged-based components like carbon nanotubes [5], graphene [6], and Quantum-dot Cellular Automata (QCA) [7]. Additionally, the research community has also focused on exploring non-charge-based solutions such as spin-based components like Spin Wave Devices (SWD) [8], All-Spin Logic [9], and nanomagnets [10], [11]. Interestingly, several beyond-CMOS concepts claim to be, in principle, non-volatile which would eliminate the need for a constant supply voltage and reducing the standby power consumption. Three characteristic examples of such concepts are SWD, QCA, and NanoMagnetic Logic (NML). However, the non-volatility property comes at a cost; in order to cascade elementary devices, the complete circuits need to be clocked [8], [12], [13]. Additionally, the aforementioned technologies use as a primitive gate a majority voter, which means that they are excellent candidates to be optimized utilizing majority-based synthesis, and more specifically Majority-Inverter Graphs (MIG) [14–16].

The combination of the non-volatility property and clocking requirement of the aforementioned technologies, creates a prime opportunity for a wave pipeline application [17]. The premise of wave pipelining is based on the fact that the rate at which logic can propagate through the circuit depends not on the longest path delay but on the difference between the longest and the shortest path delays [18]. One of the main sources of delay variations that limit the application of wave pipelining in CMOS technology is gate-delay data-dependence, where gate delay is not independent from the input pattern [18]. This however is not present in some beyond-CMOS technologies (including SWD, QCA, and NML) and the fact that utilizing MIG synthesis results in a circuit comprising one logic primitive (majority gate) increases the applicability of wave pipelining. Enabling wave pipelining for such beyond-CMOS technologies would require to equalize the delays of all paths in the circuit and therefore increase its size.

The main contributions of this paper are the utilization of both logic and memory capabilities of emerging technologies and the exploration of the wave pipelining impact and benefits on all crucial metrics of MIG circuits assuming beyond-CMOS implementations with SWD, QCA, and NML technologies. We show that applying wave-pipelining to beyond-CMOS technologies boosts significantly their performance metrics in terms of throughput per unit power and per unit area, compared to non-wave-pipelined counterparts. More specifically, we show improvements in the aforementioned metrics of 25× and 23× for SWD, 8× and 13× for QCA, and 3× and 5× for NML.

The remainder of this paper is organized as follows. In Section II, we introduce the central concepts used in this work. Section III presents the buffer insertion algorithm. Fan-out restriction that is required for beyond-CMOS technologies described in Section IV. Section V presents benchmarking results for our algorithmic extension, followed by conclusions in Section VI.

II. BACKGROUND & MOTIVATION

This section introduces MIG synthesis along with the specific assumptions for the selected technologies. Additionally, the
general wave pipelining concept and how this can be targeted for beyond-CMOS technologies is described.

A. Majority-Inverter Graph

Majority-Inverter Graphs [14]–[16] are logic representation forms based on the majority and complement function as the only logic primitives. A MIG is a data structure for Boolean function representation and optimization. It is defined as a homogeneous logic network consisting of 3-input majority nodes and regular/complemented edges. MIGs can efficiently represent Boolean functions thanks to the expressive power of the majority operator, which contains both AND and OR operation and is one of the basis for basic operation of binary arithmetic [19]. As a consequence of the AND/OR inclusion by MAJ, traditional AND/OR/INV Graphs (AOIGs) are a special case of MIGs and MIGs can be easily derived from AOIGs [20]. Fig. 1 shows an example of MIG representations derived from its optimal AOIG. Intuitively, MIGs are at least as compact as AOIGs. We refer the interested reader to [16] for an in-depth discussion on MIG optimization recipes.

B. Technology Implementation

Hereafter, we introduce the operating principles of each technology, and their basic implementations of an inverter and a majority gate shown in Fig. 2.

1) Spin Wave Devices: Are logic components that utilize spin waves (propagating oscillations of magnetization in ferromagnetic materials) as the carrier of information and were introduced in [8]. SWDs have been put forward as a competitive option to CMOS in [3], [21], [22]. The operating principle of these circuits relies on a synthetic multiferroic stack used to generate and detect spin waves, called Magneto-Electric (ME) cell [8], [23]. The generated spin waves propagate in ferromagnetic wires, called spin wave buses. The computation principle is based on the interference of propagating spin waves, where the information is encoded in the phase of the waves. In Fig. 2a-INV we present the inverter component which is a simple waveguide, with a magnetically pinned layer, that inverts the phase of the propagating spin wave. The majority gate (Fig. 2a-MAJ) is the simple merging of three spin waveguides.

2) Quantum-dot Cellular Automata: Are logic components based on the interaction of quantum-dot cells. Each quantum-dot cell is composed by four quantum dots arranged in a square fashion and coupled by tunnel barriers. Two free electrons contained in each cell are able to tunnel through the barriers and occupy different quantum dots. Due to Coulomb repulsion, the two electrons are always positioned in opposite corners of the cell and thus amounting to two possible stable polarization states. These two states are used to represent logic ‘0’ and ‘1’. However for these polarization states to be energetically stable, the operating temperature of this concept is limited to ~1 K [12]. Fig. 2b shows the layout of a QCA inverter and majority gate.

3) NanoMagnetic Logic: Also known as Magnetic Quantum Cellular Automata, was first introduced by Cowburn et al. [10] and Csaba et al. [11]. In NML, the information is encoded in the perpendicular magnetization (along + or - ) of ferromagnetic dots. The computation is mediated through dipolar coupling between nanomagnets, which forces neighboring nanomagnets to orient their magnetic moment anti-parallel to each other. Fig. 2c shows the layout of a NML inverter and majority gate. NML is the most mature and feasible of the considered technologies, with experimentally proven small in size circuits [24], and since its operation is based on individual nanomagnets, its non-volatility property is the largest of the technologies considered in this work. However, this means that the energy and delay required for operation are the highest among SWD, QCA, and NML.

C. Wave Pipelining and beyond-CMOS Application

Ordinary pipelined systems [25], [26] can process more than one instructions on a set of data simultaneously and are divided in several stages, isolated by registers. Each of these stages is nominally performing its part for a separate instruction than the rest of the stages. The data flow through each stage is determined by the global clock signal which allows processing of a new set of data only once the previous set has propagated to the next stage. In contrast, wave pipelining is the use of multiple coherent set of data (or ‘waves’) between registers of one stage [27]. Fig. 3 presents a simplified view of data waves propagating through a system increasing the maximum rate at which this system can be pipelined [17].

Wave pipelining has been explored thoroughly for CMOS technology application, from synthesis techniques [29] to
VLSI chips [27], [30]. However, a number of open problems still exist for broad use of wave pipelining [18]. One of the most important constraints which wave pipelining imposes is that all the propagation paths from the combinational circuit’s inputs to outputs have approximately the same delay, then each data wave propagates uniformly to the outputs without interfering with adjacent waves.

The aforementioned constraint does not seem too limiting under the assumption of a non-volatile cell-based technology that requires a clock for signal regeneration and propagation. Especially, since such technologies (SWD, QCA, NML) can exploit MIG synthesis that basically introduces only one primitive gate, delay variations are reduced to a minimum. Fig. 4 depicts an example of how a three-phase regeneration clock can implement wave pipelining in a cell-based emerging technology.

![Fig. 4. Schematic of three-phase clock application to an all-buffer chain. Data waves propagate from cell to cell.](image)

In this section, we describe the buffer insertion algorithm and present its impact on a MIG netlist. The buffer insertion algorithm will balance each path of the netlist such that every path from input to output has equal length. We assume that the input of the algorithm is an already optimized MIG netlist. The algorithmic is kept technology-agnostic by assuming generic components for majority and negation operations. However, we have included in the implementation the possibility to adjust component weights so that the final result can be tailored to different technologies.

The main operation of the algorithm is to traverse the MIG and compare component depths and insert balancing buffer elements between two components. For a good understanding of the algorithm operation, we need to introduce the following definitions:

- Distance ($D$) between two different components, is the set of lengths of any path going from the source to the destination. This means that from a distance set $D$ we can get minimum and maximum distance.
- Base distance ($BD$) of a component, is the set of lengths of any path going from any netlist input to the component. The maximum of this set is the depth of the component.
- Exclusive base distance ($xBD$) of a component. The same as $BD$ but excluding the component itself. The maximum of the $xBD$ is one level lower than the depth of the component.

Objectives of the buffer insertion algorithm are: (a) all paths from one component to another must be equal length; (b) maximum base distance of all netlist outputs must be equal. Differently worded, for any two connected components the minimum distance must be equal to its maximum distance. If the first goal is also obtained, then the base distance of all outputs must be equal as this set will only contain one number. Additionally, it will do this by inserting the minimal amount of buffers. The algorithm utilizes a greedy method, and a short version of the algorithm is given in Algorithm 1. It goes through each component with a non-empty fan-out, and a short version of the algorithm is given in Algorithm 1.

The algorithm iterates over all components that have a fan-out, meaning majority gates and inputs. For a given component $comp$ the fan-out (set of connected components to $comp$) is sorted based on their maximum $xBD$. Starting from the closest component in $comp$’s fan-out, the algorithm adds $m$ buffers between $comp$ and closest component, where $m$ is based on the difference between the current component depth and closest component depth. After the algorithm iterates over the entire graph, all paths between any two connected components will be equal length. As a last step the buffer insertion algorithm, adds ‘padding’ buffers to all output components to make sure that all of them are at the same $BD$. The proof of correctness and solution optimality...
Fan-out restriction (from 2 to 5) over the original critical path length increases with the original critical path length, especially in non-critical paths and introduces a large number of additional buffer components. To study the impact of this algorithm and throughout this paper we use a large set of MIG benchmarks that was used in [16]. Fig. 5, shows the number of buffers added versus the original MIG benchmark size.

![Graph showing number of buffers added versus original MIG benchmark size](image)

Fig. 5. Number of balancing buffers added to the each netlist versus the netlist sizes.

We observe that this number follows a power trend in the form of $B(s) = 7.95s^{0.9}$, where $B(s)$ is the number of buffers inserted and $s$ is the original circuit size. On average, the number of buffers inserted ranged from 2× to 4× the original netlist size. Therefore, we conclude that, as intuitively expected, the algorithmic extension has a large impact on the number of components of a MIG netlist. However, if the wave pipelining requirements were to be taken into account during the original MIG optimization, then the size of the netlists could be reduced. Here we assume that the input netlist is already optimized for depth.

IV. FAN-OUT RESTRICTION

The fan-out restriction algorithm is created in order to limit the cascading of one component, so that it is feasible for beyond-CMOS implementation, given that several emerging technologies have no intrinsic gains. This fan-out restriction is similar to the addition of drive cells in contemporary CMOS designs but here the fan-out limitations are assumed to range from 2 to 5 (relatively small numbers compared to CMOS capabilities). An example would be that a fan-out of 3 is a reversed majority node. The primary goal of the algorithm is to limit the fan-out size. The second goal is to insert as little components as possible. An example fan-out restriction is shown in Fig. 6, where the resulting netlist has a maximum fan-out of 3. The proof of correctness and solution optimality (in terms of number of fan-out components inserted) have been derived but are not included in this paper for brevity.

![Diagram of fan-out restriction algorithm](image)

Fig. 6. Example of fan-out restriction algorithm, where the resulting fan-out is limited to three. (a) Initial condition of node with fan-out larger than three. (b) Result after fan-out restriction, with three fan-out gates and a buffer added, and two nodes delayed.

Although this algorithm is implemented separately from the buffer insertion algorithm, it takes into account the general goal of wave pipelining and tries to not ‘leave’ residual paths that jump through graph levels and thus would need re-balancing (see buffer in Fig. 6b). An aspect of this algorithm that can strongly impact the netlist, is that in effort of inserting minimum amount of fan-out components (and use them efficiently) it can introduce path delays (see delayed nodes in Fig. 6b). Fig. 7 shows the increase of critical path length after fan-out restriction (from 2 to 5) over the original critical path length of the MIG benchmarks.

![Graph showing increase of critical path length](image)

Fig. 7. Increase of critical path length after fan-out restriction.
the extreme fan-out restriction of 2. On average the critical
path length increases by 140%, 57%, 36%, 26%, for fan-out
restrictions of 2, 3, 4, and 5, respectively.

Since the fan-out restriction algorithm increases the depth
of the netlist, in order to fully enable a MIG netlist for wave
pipelining it has to be performed before the buffer insertion
algorithm. Fig. 8 shows the impact of both algorithms, ran
separately and together, in terms of normalized netlist size
averaged all the benchmarks.

Fig. 8. Impact on number of components (normalized and averaged over all
benchmarks). BUF is the buffer insertion algorithm, FOx is fan-out restriction
to x, and FOx+BUF is both algorithms one feeding to the next.

Three important statements arise from Fig. 8; (a) the number
of added buffers from the combination of the two algorithms
is larger than number of buffers inserted when the algorithms
are performed individually, (b) the number of fan-out gates
inserted is independent of the buffer insertion algorithm, (c)
the best case of impact of the proposed algorithmic extension
is a 5× increase in netlist such which is significant. The first
observation is directly related to the added delay of the fan-
out restriction algorithm. The third observation could seem
negative but it is important to keep in mind that the target is
the instruction parallelism enablement in these technologies so
this netlist size cost has to be further benchmarked.

V. BENCHMARKING RESULTS

As mentioned in Section III, we use the MIG benchmarks
from [16] to benchmark the proposed wave pipelining enable-
ment. Table I introduces the primitive area, delay, and energy
constants used for each technology, extracted from [22] for
SWD, [12] for QCA and [11], [24] for NML. Additionally,
Table I shows the relative cost for each component included
in the wave pipelined enabled netlists. For the following ban-
chmarking results, we have only considered fan-out restriction
to 3 and assumed that the fan-out gate (FOG) is equivalent to
a reversed majority gate.

We used 37 benchmarks to study the impact of wave pipeli-
ning, however for brevity Table II presents the benchmarking
results for selected benchmarks for each of the three selected
technologies. As expected the size, depth, and calculated area
of the benchmarks in all three technologies are increased.
Additionally, we observe an increase of power in the case of
NML which is also expected, since the size of the benchmarks
is increased dramatically. However, the calculated power me-
tric for SWD and QCA technologies tends to decrease for

the wave pipelined benchmarks which is counter-intuitive. In
fact, this power decrease is an artifact related to the increased
delay of operation and the technology assumptions (where we
include a power dominant sense amplifier [22] for SWD and
the large QCA inverter). To highlight wave pipelining benefits
we need to focus on the throughput it enables. We assume
a three-phase clocking scheme as the one shown in Fig. 4.
A circuit can simultaneously process \( N = d/3 \) instructions,
where \( d \) is the depth of each benchmark. The throughput of
the non-pipelined benchmarks is calculated based on its complete
delay. The last two columns of Table II shows the normalized
Throughput over Area unit (T/A) and over Power unit (T/P)
for each benchmark.

Fig. 9 summarizes the benefits of wave pipelining where
the normalized T/A and T/P gains, averaged over all the 37
benchmarks, are 5× and 23× for SWD, 8× and 13× for QCA,
and 3× and 5× for NML. It is worth noting that in this work
we ignore the overhead that the required clocking network
introduces but this is fair since it is not included for either of
the benchmarking calculations.

![Fig. 9. Normalized T/A and T/P ratios for each technology implementation.](image)

VI. CONCLUSIONS

In this work, a synthesis framework that enables wave
pipelining as an extension of MIG synthesis for emerging
technologies is presented, along with a study on its impact
on the performance of technology implementations. Through
the proposed algorithms, we construct feasible netlists for the
beyond-CMOS technology standards and efficiently exploit
their inherit non-volatility. We show that wave pipelining is
a strong candidate for performance boosting that might be

<table>
<thead>
<tr>
<th>Technology</th>
<th>Area (( \mu m^2 ))</th>
<th>Delay (( \mu s ))</th>
<th>Energy (( fJ ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWD</td>
<td>0.002304</td>
<td>0.42</td>
<td>1.44·10⁻⁸</td>
</tr>
<tr>
<td>QCA</td>
<td>0.000410</td>
<td>0.0012</td>
<td>9.80·10⁻⁷</td>
</tr>
<tr>
<td>NML</td>
<td>0.0098</td>
<td>1.0</td>
<td>5.00·10⁻⁴</td>
</tr>
</tbody>
</table>

![Fig. 8. Impact on number of components (normalized and averaged over all
benchmarks). BUF is the buffer insertion algorithm, FOx is fan-out restriction
to x, and FOx+BUF is both algorithms one feeding to the next.](image)
needed in beyond-CMOS technologies. The throughput over area unit and over power unit show significant increase when wave pipelining is applied.

Acknowledgments: This research was supported by the European Research Council (H2020-ERC-2014-ADG 669354 CyberCare) and the Swiss National Science Foundation (200021_169084 MAJesty).

REFERENCES


TABLE II

<table>
<thead>
<tr>
<th>SWD</th>
<th>Depth (10^10)</th>
<th>Size</th>
<th>Area (µm²)</th>
<th>Power (µW)</th>
<th>Throughput (MOPS)</th>
<th>T/A ratio (×)</th>
<th>T/P ratio (×)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASC</td>
<td>Original</td>
<td>WP</td>
<td>Original</td>
<td>WP</td>
<td>Original</td>
<td>WP</td>
<td>Original</td>
</tr>
<tr>
<td>DES AREA</td>
<td>22</td>
<td>38</td>
<td>4187</td>
<td>13325</td>
<td>63.24</td>
<td>123.82</td>
<td>108.23</td>
</tr>
<tr>
<td>MUL32</td>
<td>36</td>
<td>58</td>
<td>9097</td>
<td>18998</td>
<td>140.95</td>
<td>201.95</td>
<td>66.14</td>
</tr>
<tr>
<td>HAMMING</td>
<td>61</td>
<td>96</td>
<td>2072</td>
<td>11523</td>
<td>32.50</td>
<td>82.43</td>
<td>0.74</td>
</tr>
<tr>
<td>MUL64</td>
<td>109</td>
<td>135</td>
<td>25773</td>
<td>139914</td>
<td>403.65</td>
<td>978.79</td>
<td>7.55</td>
</tr>
<tr>
<td>REVX</td>
<td>143</td>
<td>225</td>
<td>7517</td>
<td>34911</td>
<td>112.75</td>
<td>266.95</td>
<td>1.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QCA</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP/Original</th>
<th>WP/Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASC</td>
<td>6</td>
<td>9</td>
<td>622</td>
<td>1885</td>
<td>2.65</td>
<td>3.34</td>
<td>0.27</td>
<td>0.23</td>
<td>41666.67</td>
</tr>
<tr>
<td>DES AREA</td>
<td>22</td>
<td>38</td>
<td>4187</td>
<td>13325</td>
<td>14.88</td>
<td>20.46</td>
<td>0.41</td>
<td>0.33</td>
<td>11363.64</td>
</tr>
<tr>
<td>MUL32</td>
<td>36</td>
<td>58</td>
<td>9097</td>
<td>18998</td>
<td>39.48</td>
<td>45.04</td>
<td>0.67</td>
<td>0.48</td>
<td>6944.44</td>
</tr>
<tr>
<td>HAMMING</td>
<td>61</td>
<td>96</td>
<td>2072</td>
<td>11523</td>
<td>9.67</td>
<td>14.11</td>
<td>0.10</td>
<td>0.09</td>
<td>4098.36</td>
</tr>
<tr>
<td>MUL64</td>
<td>109</td>
<td>135</td>
<td>25773</td>
<td>139914</td>
<td>17.96</td>
<td>168.74</td>
<td>0.66</td>
<td>0.77</td>
<td>2293.58</td>
</tr>
<tr>
<td>REVX</td>
<td>143</td>
<td>225</td>
<td>7517</td>
<td>34911</td>
<td>30.62</td>
<td>44.50</td>
<td>0.13</td>
<td>0.12</td>
<td>1748.25</td>
</tr>
<tr>
<td>DIFFEQ1</td>
<td>219</td>
<td>282</td>
<td>17726</td>
<td>306937</td>
<td>88.66</td>
<td>1654.57</td>
<td>8.48</td>
<td>6.59</td>
<td>10.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NML</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP</th>
<th>Original</th>
<th>WP/Original</th>
<th>WP/Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASC</td>
<td>6</td>
<td>9</td>
<td>622</td>
<td>1885</td>
<td>16.85</td>
<td>44.60</td>
<td>6.88×10⁻³</td>
<td>1.22×10⁻²</td>
<td>8.33</td>
</tr>
<tr>
<td>DES AREA</td>
<td>22</td>
<td>38</td>
<td>4187</td>
<td>13325</td>
<td>106.21</td>
<td>316.76</td>
<td>1.18×10⁻²</td>
<td>2.04×10⁻²</td>
<td>2.27</td>
</tr>
<tr>
<td>MUL32</td>
<td>36</td>
<td>58</td>
<td>9097</td>
<td>18998</td>
<td>248.28</td>
<td>468.51</td>
<td>1.69×10⁻²</td>
<td>1.98×10⁻²</td>
<td>1.39</td>
</tr>
<tr>
<td>HAMMING</td>
<td>61</td>
<td>96</td>
<td>2072</td>
<td>11523</td>
<td>58.20</td>
<td>254.30</td>
<td>2.34×10⁻³</td>
<td>6.30×10⁻³</td>
<td>0.82</td>
</tr>
<tr>
<td>MUL64</td>
<td>109</td>
<td>135</td>
<td>25773</td>
<td>139914</td>
<td>718.38</td>
<td>3039.22</td>
<td>1.62×10⁻¹</td>
<td>5.32×10⁻²</td>
<td>0.46</td>
</tr>
<tr>
<td>REVX</td>
<td>143</td>
<td>225</td>
<td>7517</td>
<td>34911</td>
<td>200.26</td>
<td>784.77</td>
<td>3.43×10⁻³</td>
<td>8.55×10⁻³</td>
<td>0.35</td>
</tr>
<tr>
<td>DIFFEQ1</td>
<td>219</td>
<td>282</td>
<td>17726</td>
<td>306937</td>
<td>495.89</td>
<td>6220.95</td>
<td>5.55×10⁻³</td>
<td>5.41×10⁻²</td>
<td>0.23</td>
</tr>
</tbody>
</table>