Applying Swarm Intelligence to Distributed On-Chip Power Management

Divya Pathak
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA
Email: divya.pathak@drexel.edu

Ioannis Savidis
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA
Email: isavidis@coe.drexel.edu

***Abstract*—An on-chip power management technique is developed that makes use of particle swarm optimization (PSO) to improve the performance per watt of the circuit while maintaining the power integrity. On-line learning is applied to determine the optimum reference voltages of the on-chip voltage regulators set through the PSO to reduce the energy consumption of the system while preventing any timing failure due to process variation, voltage variation, temperature, and aging. The run-time adaptive voltage delivery technique is applicable to any processor architecture. Simulation results on a streaming multi-processor similar to the NVIDIA GV100 GPU in a 7 nm FinFET technology indicate an average reduction of 35%, 40%, and 5% in, respectively, the power consumption, the threshold voltage drift, and the operating temperature as compared to existing techniques that implement static voltage guardbands.**

***Index Terms*—particle swarm optimization, machine learning, power supply noise, voltage regulators, transistor aging, evolvable hardware.**

I. Introduction

Domain specific architectures (DSA) provide a means to address the increased computational demand of training deep neural networks (DNNs). Current research objectives include the optimization of DSAs and domain specific programming languages to improve the energy efficiency of the DNN. An unexplored research area with the potential to improve the cost-energy-performance of DSAs is power delivery through distributed on-chip voltage regulators (OCVRs). State of the art graphics processing units (GPUs) deployed to train DNN workloads operate with off-chip voltage regulators [1], which leads to higher latency when reacting to changes in load current. High performance CPUs, GPUs, and DSAs developed to accelerate deep learning applications require the use of advanced run time power management techniques to mitigate timing errors due to process, voltage, and temperature (PVT) variation and aging. Distributed OCVRs are, therefore, a requirement for such processor architectures to locally control the power supply voltage. However, a large and sustained current demand in GPUs subjects the OCVRs to increased aging and a higher susceptibility to process variation and noise. In addition, large dI/dt events in the GPU and other DSAs lead to power supply noise due to the finite latency of the OCVRs to react to changes in the load current. Therefore, power delivery through OCVRs is a complex research area [2], [3], with the long term implications of on-chip voltage regulation not fully analyzed.

In addition to enhancing the energy efficiency through distributed OCVRs, power integrity in advanced technology nodes, specifically FinFET based processes, must also be addressed. Challenges include circuit aging, process variation, power supply noise, and the self heating effect (SHE), which are more pronounced in sub-nanometer nodes [4]. In current literature, several techniques exist to mitigate such challenges, but the techniques address each challenge in isolation. In this paper, an evolving voltage delivery mechanism is developed that reduces 1) circuit aging, 2) power supply noise, and 3) temperature hotspots due to SHE, while accounting for process variation. The closed loop, run-time technique assigns reference voltages to the distributed OCVRs within the technology imposed guard-bands and without inducing any timing violations. A particle swarm optimizer (PSO) assigns the reference voltages to the OCVRs, which is provided data from distributed on-chip timing sensors that sense local variations in power supply voltage, temperature, age, and device parameters.

The primary contributions of this work include:

- The development of a circuit technique for adaptive voltage assignment to processing elements. The evolving voltage assignment is implemented with distributed on-chip voltage regulators of which the reference voltage is set through a particle swarm optimizer. The technique is validated on a GPU architecture.
- The run time assignment of the power supply voltage through the PSO, which compensates for the PVT variation and aging in not only the GPU functional units but also the OCVRs.
- The first work to develop a *run-time* machine learning algorithm for the power management of processing elements that is executed completely in the circuit layer.

The remainder of the paper is organized as follows. The proposed particle swarm optimization based power management methodology is described in Section II. Simulation results are provided in Section III. Conclusions are offered in Section IV.

II. Swarm intelligence for an evolvable PDN

Classical optimization methods including linear programming, non-linear programming, Newton's method, quadratic programming, and sequential unconstrained minimization assume the optimization of a continuous variable, which yields

978-1-5386-6648-7/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCD46524.2019.00079

local optimum solutions [5]. The on-chip power delivery network with OCVRs contains both discrete and continuous control variables. Applying techniques for continuous variables to discrete variables results in both an increase of the objective function and violations of inequality constraints. Evolutionary programming methods including simulated annealing, genetic algorithms (GA), tabu search, and partical swarm optimization are better suited for discrete variables and non-differential objective functions [6].

The particle swarm optimizer (PSO), however, offers a robust and simple implementation that produces superior results as compared to other evolutionary algorithms. Prior research has shown that the PSO offers different routes through the problem hyperspace as compared to GA and other optimization algorithms [7], [8]. The low overhead to store results during each iteration of the algorithm and the simplicity of implementing the circuit make the PSO algorithm an ideal choice for run-time control of the power supply voltages.

The particle swarm optimizer operates on a set of *particles*, where the position of each particle x_i in a D dimensional hyperspace represents a potential solution to the optimization problem. For a given particle p_i, the position and velocity at time t are represented as, respectively, x_i(t) = ($x_{i,1}$(t), $x_{i,d}$(t), ..., $x_{i,D}$(t)) and v_i(t) = ($v_{i,1}$(t), $v_{i,d}$(t), ..., $v_{i,D}$(t)). The current best position for particle p_i is recorded as P_{best_i}. The best position among the entire particle population is recorded as G_{best}. The velocity and position of a particle constantly change based on both the experiences of the given particle as well as the experiences of the other particles in the swarm. The velocity and position of the particles are updated as given by (1) and (2), respectively. In (1), ϕ_1 and ϕ_2 are learning factors, ρ_1 and ρ_2 are random functions in the range of [0,1], and ω is the inertia weight that is applied to constrain the influence of past velocities on the current velocity of a particle. The personal acceleration coefficient ϕ_1 provides a weight to the prior velocity of a particle when determining the current velocity. The social acceleration coefficient ϕ_2 provides a weight to the swarm when determining the new velocity of a particle. Therefore, the set values of ω, ϕ_1, and ϕ_2 establish a procedure to explore the hyperspace D.

$$v_{i,d}(t+1) = \omega \times v_{i,d}(t) + \phi_1 \times \rho_1 \times (P_{i,d}(t) - x_{i,d}(t)) + \phi_2 \times \rho_2 \times (G_{i,d}(t) - x_{i,d}(t)) \quad (1)$$

$$x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1) \quad (2)$$

A. Voltage assignment through the PSO

The distributed OCVRs operate as a swarm to locally optimize the operating voltage with the smallest guard-band needed to prevent timing violations on the local critical paths, while also compensating for aging related degradation in both the load and OCVR circuits. Distributed timing sensors provide a measure of the timing margin at the operating frequency of the circuit and transmit the data to the voltage reference control circuit of the OCVRs. The reference voltage

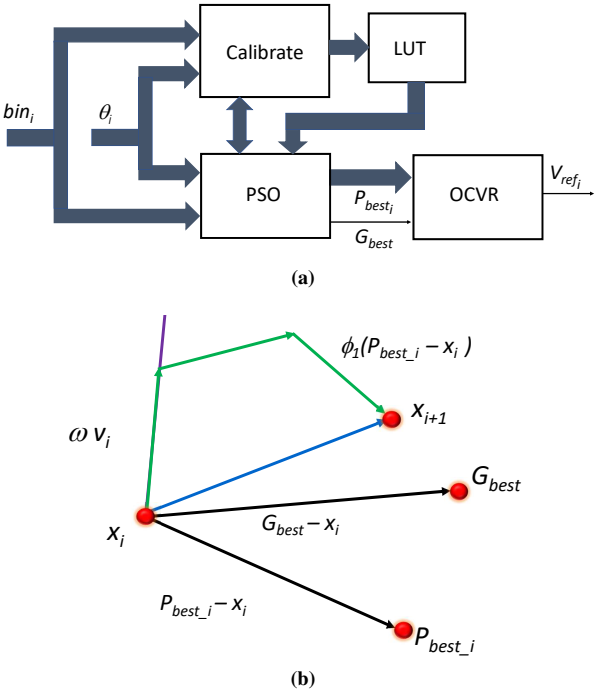


Fig. 1: Assignment of a new voltage (position) to each OCVR (particle) shown through (a) a functional block diagram of the PSO algorithm implemented in Verilog and (b) a vector diagram of particle x_i moving towards P_{best_i} or G_{best}.

of each OCVR is independently modulated based on the optimized P_{best} values provided by the on-line PSO algorithm. Alternatively, to reduce the complexity of implementing the circuit that generates the reference voltage for each OCVR, the PSO provides a global optimum voltage G_{best}, applied as a common reference to all the OCVRs.

The P_{best} of a particle (OCVR) is a function of the sensed time delay from the nearest timing sensor as given by (3). The P_{best} values are the local optimal voltages accounting for local phenomenon including *IR* drop, d*I*/d*t* noise, process and aging induced V_{th} degradation, and hotspots. The G_{best} for the swarm of particles is the maximum P_{best} value obtained across all OCVRs and is given by (4). The P_{best} and G_{best} are functions of time, providing the lowest power supply voltage to the circuit without causing any timing violations.

$$P_{best_i} = f(Temp(t), V_{noise}(t), V_{th\_aging}(t), W_{load}(t)) \quad (3)$$

$$G_{best} = max(P_{best_i}) \quad (4)$$

Existing statistical static timing analysis tools are applied to each voltage domain to determine the set of timing paths that exhibit high delay or are statistically likely to evolve as the paths with the maximum delay as the circuit ages. These set of timing paths are termed as potential critical paths (*PCPs*). A sensor to monitor timing margin violations is integrated within each *PCP*. The placement of the distributed voltage regulators and time based sensors is set based on the distribution of the

PCPs. At the first power up of the circuit, the distributed timing sensors are calibrated across the supported voltage and frequency ranges of the streaming multi-processors (SM) while executing a known workload that produces the lowest variation in the activity factor. The calibrated timing bins for each sensor at each frequency are stored in on-chip RAM as a look up table (LUT). Due to intra-die process variation, the calibrated bins vary for each timing sensor. The LUT is accessed by the on-line PSO, which compares the latest output from each timing sensor with the calibrated data in the LUT. As long as there is no violation in the timing margin (θ_i in Fig. 1(a)) of any of the PCP_i and the captured timing bins are lower than the calibrated bins, the P_{best} position is updated. In case of a violation in the timing margin of any of the *PCPs*, a recalibration of the time based sensors is performed. The inputs and outputs to the on-chip PSO block are shown in Fig. 1(a). A vector diagram depicting the assignment of an updated voltage (position) to each OCVR (particle) is shown in Fig. 1(b). As the system ages based on the executed workloads and environmental conditions, a number of re-calibrations of the sensors are performed until no further reductions to the voltage is possible that meets the timing constraints of the path.

B. Timing and aging sensor to direct the PSO

On-chip sensors are needed to inform and direct the decision of the on-chip PSO in the assignment of voltages to the swarm of OCVRs. An optimal selection and placement of the sensors is required to most effectively characterize the operating voltage, temperature, and frequency of each voltage domain. In addition, for the proposed PSO, aging sensors are integrated to further characterize the state of the circuit, with the resulting data analyzed to prevent timing violations due to aging in potential critical paths (PCPs). The construction and calibration of the timing and aging sensors is described as follows.

1) Distributed timing sensors: A timing sensor such as a latched tapped delay line [9] provides an effective and simple means to quantify the variation in the captured clock edges propagating through a chain of buffers. The variation in the captured clock edges is a function of the clock jitter, operating voltage, and temperature. Therefore, the combined effect must be characterized rather than individually quantifying the operating temperature, voltage, or load current with integrated physical sensors. Advanced circuit implementations of timing sensors are implemented in commercial microprocessors to characterize the available timing margin of critical paths (critical path monitors) [10], [11]. The proposed run-time PSO allows for the integration of commercial sensors .

A latched tapped delay line is designed in a 7 nm FinFET predictive technology model (PTM) [12] process and is used as the timing sensor of the *PCPs*. The sensors are distributed across the IC and are used to characterize and bin the location of the clock edge, with results provided to the on-line PSO. The schematic of the delay line based timing sensor is shown in Fig. 2(a). The local clock signal for the voltage domain in which the delay line is placed is applied to the buffer chain. The buffer (or bin number) at which the clock edge

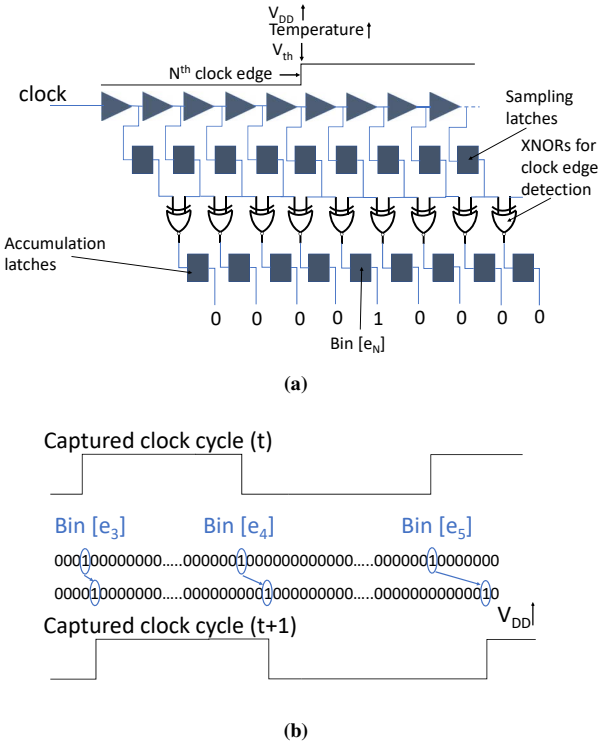


Fig. 2: A latched tap delay line [9] in a 7 nm PTM FinFET technology [12] used as a timing sensor where (a) depicts the circuit schematic of the sensor and (b) the output of the latches capturing the shift in the clock edges due to an increase in V_{DD}.

is captured provides a measure of the local physical and electrical characteristics of the circuit in the vicinity of the delay line. An illustration of capturing a clock edge is shown in Fig. 2(b), where the propagation of the edge through the delay line shifts due to an increase in V_{DD}. The clock edge propagates through additional inverters when the delay per bin (*Delayperbin*) decreases with an increase in the supply voltage or temperature. Increases in V_{th} due to process variation and aging reduce the number of bins the clock edge propagates through.

The delay of a minimum sized inverter designed in a 7 nm FinFET PTM process at a nominal voltage of 0.7 V is 2.02 ps. Optimally sized buffers are implemented and the output of each is latched into a master/slave flip-flop to reduce the size of the inverter chain and ensure that the fifth edge of the propagating clock at the highest supported operating frequency of 4 GHz is captured reliably across all PVT variation. The two consecutive latches that capture the opposite logical output of the buffers indicates the location of the propagating clock edge (rising or falling) in the buffer chain. The location of the fifth clock edge ($bin_i[e_5]$), which is transitioning high, is considered as input to the PSO algorithm since the sensitivity to V_{DD} increases the deeper the clock signal propagates into the buffer chain [9].

2) Calibration of the timing sensor: At the beginning of life of the SM, the distributed latched tapped delay lines are calibrated and the results are stored in a LUT. The calibration

TABLE I: Delay per bin of the i^{th} timing sensor stored in a look up table (LUT_i).

calib_count →

Voltage	$Delayperbin_i$	$Delayperbin_i$	...	$Delayperbin_i$
V_{min}				
⋮	⋮	⋮	⋮	⋮
P_{best_1}	D_1	D_2		
⋮	⋮	⋮	⋮	⋮
P_{best_2}		D_1		
⋮	⋮	⋮	⋮	⋮
v_{nom}				
⋮	⋮	⋮	⋮	⋮
V_{min}				

is performed at a nominal temperature of 25°C. A workload is executed on the current SM being calibrated that generates the least variation in the power supply voltage (constant activity factor). The location of the timing sensors provides a characterization of the process variation in the given voltage domain the sensor is placed in as the delay per bin amongst the timing sensors varies with differences in V_{th}. The variation in delay for a normal distribution (σ/μ of 0.9%) of V_{th} is shown in Fig. 3(a) for a 7 nm FinFET technology [12].

At each supported voltage level, the delay per bin (*Delayperbin*) for each timing sensor *i* is calculated as the difference between the edges of one clock cycle ($bin_i[e_5]$ - $bin_i[e_3]$), with the result stored in a LUT. The method to determine the new operating voltage from the P_{best} value is illustrated in Fig. 4. The structure of the LUT is depicted through the corresponding stored results listed in Table I. The $bin_i[e_5]$ value corresponding to the nominal voltage, as specified by the foundry for the given technology node, is stored in the LUT at the start of life for each timing sensor *i* as $calib_i$. The change in the $bin_i[e_5]$ location with operating voltage at a nominal temperature of 25°C and for a given V_{th} of 0.34 V is shown in Fig. 3(b). The variation in the $bin_i[e_5]$ position with operating temperature for a supply voltage of 0.7 V and a V_{th} of 0.34 V is shown in Fig. 3(c), where the clock signal propagates deeper into the buffer chain with increasing temperature due to temperature effect inversion [13]. The sensitivity of the latched tapped delay line to process, aging, voltage, and temperature, therefore, proves ideal as an on-chip sensor to provide data that directs the execution of the PSO algorithm.

During the operational lifetime of the IC, subsequent calibrations are performed if a violation in the timing margin is detected by the aging sensor. The calibration carried out during the lifetime of the IC is provided by the *CALIBRATE* procedure of the PSO as described in Algorithm 1. The *Delayperbin* value D_1 corresponding to the current best voltage assignment P_{best_1} of an OCVR is compared with the updated calibration data obtained at each supported operating voltage. The updated voltage P_{best_2} corresponding to the delay D_1 is assigned as the new reference voltage to the OCVR.

3) Aging sensor: A technique to predict timing failure is developed in [14], where the transition of the output signal of a critical path is monitored to detect any transition that occurs within the set timing interval of the guardband. A signal transition detected in the guardband interval implies

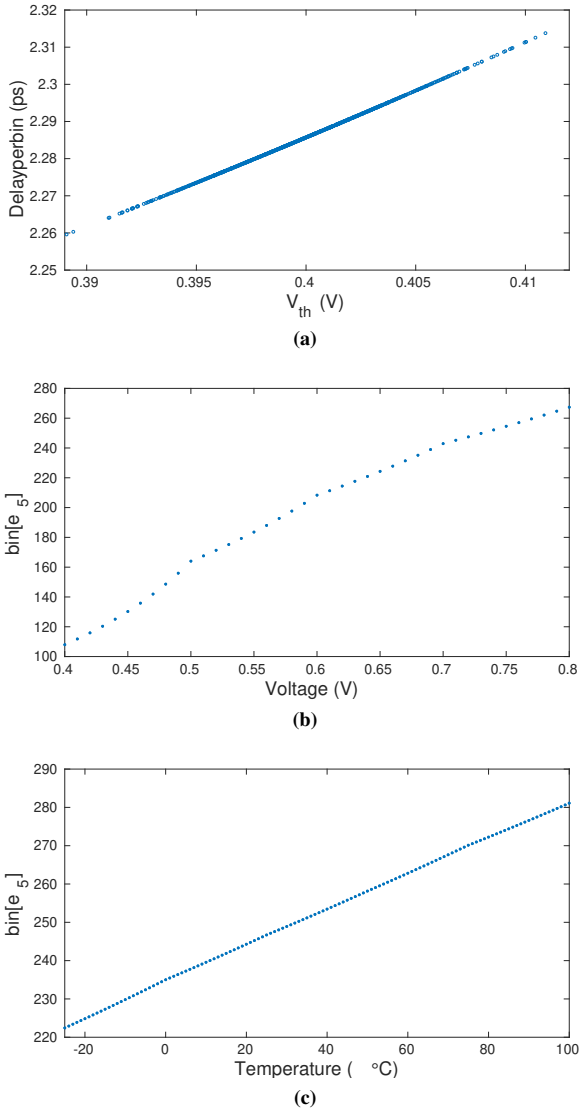


Fig. 3: Response of the timing sensor at start of life as determined by the captured bin number of the fifth clock edge ($bin_i[e_5]$) with variation in (a) intra-die V_{th} with a coefficient of variation σ/μ of 0.9 %, (b) V_{DD} (at T=25°C and V_{th}= 0.34 V), and (c) operating temperature (at V_{DD} = 0.7 V and V_{th}= 0.34 V).

that for the given logical input, the critical path has slowed due to circuit aging and is close to generating a timing fault. A monitoring circuit is embedded into the output latch of a critical path. The block diagram of the monitoring circuit, consisting of the delay element and the stability checker, is shown in Fig. 5(a). An output latch stores the result of the stability checker. The schematic of the delay element and the stability checker, is shown in Fig. 5(b). The delay element introduces a lag in the complement of the clock signal (*Clock'*). The delayed *Clock'* signal is provided as input to the stability checker, which checks for any change in the output of the critical path during the guardband interval as shown in Fig. 5(c). The global *Monitor* signal activates the

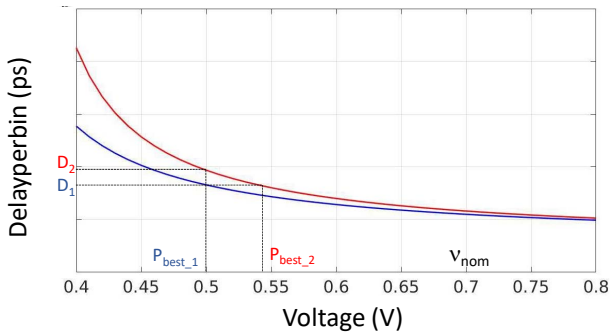


Fig. 4: An illustration of the assignment of a new voltage each time the calibration procedure is invoked due to a flag raised by an aging sensor. The assigned new voltage results in the same delay per bin in the timing sensor as observed with the previous calibration. The calibration results for each timing sensor *i*, are stored in a look up table (LUT_i) as shown in Table I.

delay element and, therefore, the aging sensor. The sensor detects transistor aging without applying any error correction or recovery techniques. For the proposed on-chip PSO based voltage assignment methodology, an aging sensor is placed in each of the potential critical paths (*PCPs*) determined during the statistical static timing analysis of a voltage domain. The output θ_i from the stability checker, for each of the PCP_i, is provided as an input to the PSO as shown in Fig. 1(a).

III. Simulated Results

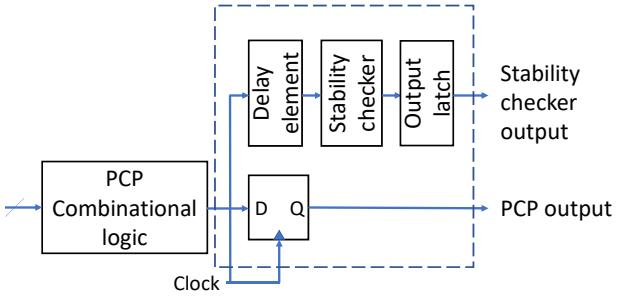
The PSO algorithm, as given by Algorithm 1, is implemented in Verilog and is used to characterize and compensate for variations in the profile of the load current, power supply noise, and threshold voltage of an SM. An SM of an NVIDIA GV100 GPU [1] is emulated with a constructed floor plan as shown in Fig. 6(a). Each SM in the GV100 is partitioned into four processing blocks, with each block containing 16 FP32 cores, 8 FP64 cores, 16 INT32 cores, a 64 KB register file, an L0 instruction cache, and two tensor cores. Three on-chip linear dropout (LDO) voltage regulators are integrated in each partition, which are roughly positioned within the FP64/INT32 cores, the register file, and the tensor cores. All four partitions of the SM are considered to operate within a single voltage domain and with a total of 12 integrated LDOs. The SM and the LDOs are designed in a 7 nm FinFET process [12]. The power traces for the functional blocks of the SM and the SM floor plan are provided as inputs to Voltspot 2.0 [15], which is a cycle accurate simulator. The voltage map across the power grid at discrete locations of the SM is obtained from Voltspot 2.0. The variation in V_{th} of the SM is characterized through VARIUS [16], assuming a 0.9% σ/μ ratio and a spatial correlation parameter ϕ of 0.2. The voltage and the V_{th} maps are provided as inputs to the developed Verilog PSO model. The per clock cycle G_{best} value obtained by the PSO is used to generate updated power traces with the same activity factor as the original power trace provided as input to Voltspot. The updated power trace is then applied as an input to the Hotspot [17] simulator, which characterizes the temperature across the SM. The simulation framework is depicted in Fig. 6(b).

Algorithm 1 Evolving power supply voltage assignment through particle swarm optimization.

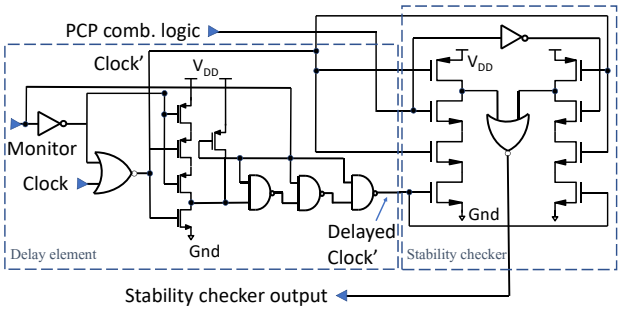
Inputs:
Set of aging sensors in each potential critical path (*PCP*): Θ
Timing sensors output: *bin(n×m)* for *n* number of timing sensors, each with a precision of *m* bits
System clock: *clk*
Outputs:
Local best voltage assignment: $P_{best}(n\times p)$ for *n* number of distributed voltage regulators and *p* bit VID code
Global best voltage assignment: G_{best}, *p* bit VID code
▷ Timing sensor calibration at beginning of life and at timing margin violation
procedure CALIBRATE(*bin*, *clk*, P_{best})
  for each bin_i, $i \in$ [1:*n*] **do**
    for each ν_j, $j \in [1{:}2^p]$ **do** ▷ for each VID code
      ▷ Compute the propagation delay per bin in the timing sensor
      $Delayperbin_j \leftarrow clk \times \left(bin_i[e_5] - bin_i[e_3]\right)^{-1}$
      if calib_count == 0 **then**
        if $\nu_j == v_{nom}$ **then**
          $calib_i \leftarrow bin_i[e_5]$
          $x_i \leftarrow \nu_j$
          calib_count ++
        end if
      else
        ▷ Search the LUT for the row with voltage equal to the P_{best_i}
        if $\nu_j == LUT_i(k,1)(P_{best_i})$ **then**
          $D_1 = LUT_i(k, calib\_count)(Delayperbin_k)$
        end if
        if $Delayperbin_j == D_1$ **then**
          $calib_i \leftarrow bin_i[e_5]$

          $x_i \leftarrow \nu_j$
          calib_count ++
        end if
      end if
      ▷ Push the *Delayperbin* computed per voltage in the LUT stack
      $LUT_i(j, calib\_count) \leftarrow Delayperbin_j$
    end for
    $P_{best_i} \leftarrow x_i$
    $v_i \leftarrow x_i$
  end for
  $G_{best} \leftarrow \max(P_{best_i})$
return *x*, *v*, *calib*, P_{best}, G_{best}
end procedure
procedure PSO(Θ, *bin*, *clk*, *x*, *v*)
  ▷ V_{max}, V_{min} : Technology imposed voltage limit for the transistor.
  ▷ Particle velocity limits
  v_{max} = γ*(V_{max}-V_{min});
  v_{min} = -v_{max};
  ▷ $\gamma \in$ [0 1]
  ▷ Constriction Coefficients:$\kappa \leftarrow$ 1, $\phi_1 \in$ [0 10] ,$\phi_2 \in$ [0 10],
  ▷ $\phi \leftarrow \phi_1 + \phi_2$;
  ▷ $\chi \leftarrow \frac{2 \times \kappa}{|\ 2 - \phi - \sqrt{\phi^2 - 4 \times \phi}\ |}$
  $\omega \leftarrow \chi$; ▷ Inertial Coefficient
  ▷ $w_{damp} \in$ [0 1]; ▷ Damping ratio of inertia coefficient
  c1 $\leftarrow \chi \times \phi_1$; ▷ Personal Acceleration Coefficient
  c2 $\leftarrow \chi \times \phi_2$; ▷ Social Acceleration Coefficient
  At every clk edge
  if $\theta_i \in \Theta$ == TRUE **then** ▷ timing violation on a *PCP*, recaliberate
    (*x*, *v*, *calib*, P_{best}, G_{best}) $\leftarrow$ calibrate(*bin*)
  end if
  for each x_i, i $\in$ [1:N] **do**
    ▷ Compute particle velocity
    $v_i \leftarrow \omega * v_i$+c1×rnd×($P_{best_i}$ - x_i) + c2×rnd×(G_{best} - x_i);

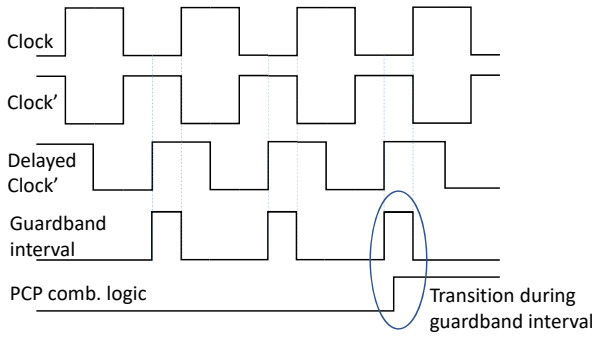
    ▷ Apply velocity limits
    $v_i \leftarrow \min(v_i, v_{max})$;
    $v_i \leftarrow \max(v_i, v_{min})$;
    ▷ Update particle position
    $x_i \leftarrow x_i + v_i$;
    if ($bin_i[e_5]$ - $calib_i$ > m'h0) && ($\theta_i \in \Theta$ == FALSE) **then**
      $P_{best_i} \leftarrow x_i$
    end if
    $\omega \leftarrow \omega \times w_{damp}$
  end for
  $G_{best} \leftarrow \max(\ P_{best_i})$
return P_{best}, G_{best}
end procedure



(a)



(b)

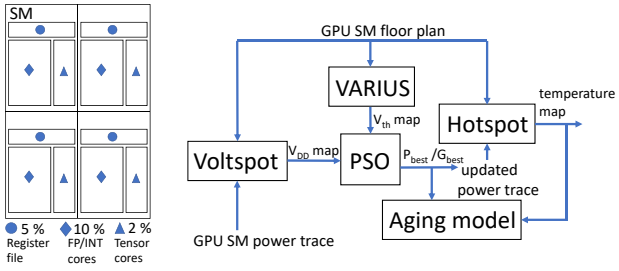


(c)

Fig. 5: Characteristics of the aging sensor [14] built into each potential critical path where (a) depicts a block diagram of the primary circuit components, (b) a circuit schematic of the sensor, and (c) a timing diagram depicting the detection of an aging violation.

A. Reduction in power and noise

Prior research analyzing the power profile of an SM determined that the caches are subject to the least variation in power consumption [18]. The FP and INT cores along with the register file (RF) are subject to large variations in power consumption per cycle. The variation in the power



(a) (b)

Fig. 6: Simulation framework to validate the PSO based voltage assignment algorithm where (a) depicts the floor plan of an SM used as an input to Voltspot and Hotspot and (b) the data flow graph depicting the link between the various tools used to characterize the execution of PSO algorithm.

consumption of the NVIDIA tensor core are not publicly disclosed. However, if assuming a constant execution of matrix multiplication and addition operations by the tensor cores, the power consumption is assumed to be less variable as compared to the FP and INT cores.

Multi-cycle power traces are generated for the FP/INT cores, the RF, and the tensor cores such that the induced voltage noise is, respectively, 10%, 5%, and 2% of the 0.7 V nominal V_{DD}. The OCVR supplies current to not only the RF but also the L0 cache, warp scheduler, and the dispatch unit. Therefore, the combined variation in the activity of the partition is set to 5%. The activity pattern (temporal) and the placement of the circuit blocks (spatial) have a combined effect on the timing and magnitude of the power supply noise at any given location on the PDN [19]. The parameters of the on-line PSO are characterized to determine the effect each has on the exploration of the voltage search space with respect to the varying activity factors of the functional blocks of the SM. The personal acceleration coefficient ϕ_1 weighs the PSO decision more towards local noise events in the vicinity of the timing sensors, which are placed in close proximity to the OCVRs in the SPICE simulation. The opposite occurs for the social acceleration coefficient ϕ_2, where the PSO decision is more heavily influenced by global noise events. The optimal voltage assignment is obtained when the personal and social acceleration coefficients are equal, which results in no timing violations. The voltage assignment becomes more conservative (less variance between P_{best} values) as the ϕ_2 coefficient is increased for a given ϕ_1. The results when setting both the personal ϕ_1 and social coefficient ϕ_2 equal to each other are shown in Fig. 7(a), whereas the results for the assignment of P_{best} when the PSO relies completely on the swarm (or social behavior) are shown in Fig. 7(b). The reduction in the variance of P_{best} when applying the voltage determined by the swarm optimization algorithm implies that blocks experiencing greater noise (overshoot due to L*dI*/*dt*) are assigned a lower voltage due to the influence of blocks with low to zero variation in activity. The reverse is true for the voltage assignment of blocks with low variation in activity, which provides less opportunity of reducing the voltage margin

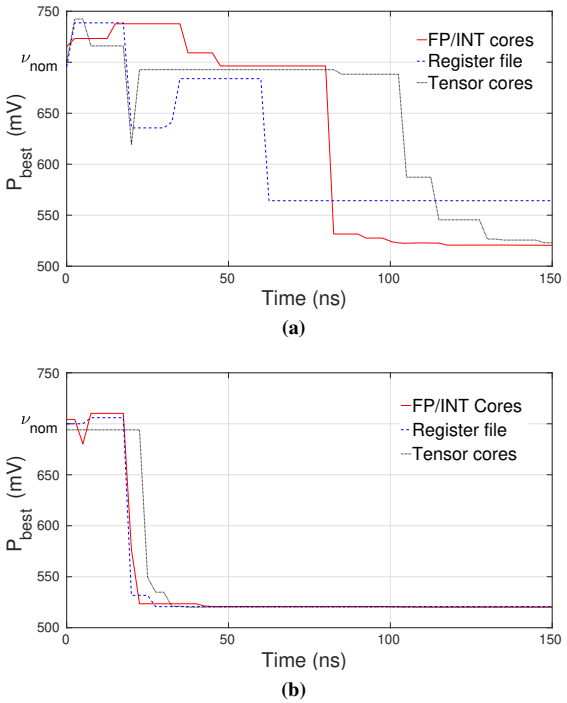


Fig. 7: Characterization of the effect of the personal (ϕ_1) and social acceleration (ϕ_2) coefficients on the decision of the PSO for varying circuit activity of the functional blocks. The evolving reference voltage assignment to a subportion of a SM that includes three OCVRs is shown. The change in P_{best} is plotted for the OCVRs supplying the FP/INT cores, register file, and tensor cores with decreasing levels of circuit activity for (a) ϕ_1 = ϕ_2 = 2 and (b) ϕ_1 = 0, ϕ_2 = 2.

of such functional blocks.

In addition, if the operating system level workload scheduler provides data to the PSO from the architectural level activity counters, the inertial coefficient ω is tuned based on the workload activity of the various functional blocks. A low value is assigned to ω when more than one functional block exhibits high variation in activity factor, which ensures that the past voltage assignment of the PSO does not dominate the current assignment and the PSO searches for a solution that satisfies the constraint on the magnitude of the transient power supply noise of the circuit. Even with a noise 10% greater than a ν_{nom} of 0.7 V on the FP and INT cores, the PSO converges to a G_{best} value significantly lower than ν_{nom} without any timing violations in the critical path(s). The average percentage reduction in the combined dynamic and static power consumption when applying the adaptive global voltage assigned to a domain as compared to an assignment of ν_{nom} is 35%.

B. Reduction in transistor aging

An off-chip voltage regulator supplying current to the entire SM (baseline) is compared with the proposed PSO based run-time voltage assignment algorithm applied to 12 distributed OCVRs as shown in Fig. 6(a) (placement of the OCVRs within the SM are shown as circles, diamonds, and triangles). The selected aging model is validated through simulation on a FinFET-based ring oscillator [20]. The process and aging induced shift in the threshold voltage $\Delta V_{th}(t)$ is mathematically expressed as given by (5). The $\Delta V_{th}(t)$ has a normal distribution with a mean $\langle \Delta V_{thA}(t) \rangle$, which is the average shift in the threshold voltage attributed to bias temperature instability and expressed by the power law given by (6). The technology parameters A and κ, and the fitting parameters, α, β, and γ, in (6) are taken from prior work [21]. Additional parameters include the temperature θ in Kelvin, the total stress time in seconds t, the duty factor of the stress signal df, and the electric field across the gate oxide E_{OX}. The variance in the threshold voltage $\sigma^2_{\Delta V_{th}}(t)$ due to process variation and aging is given by (7), where $\sigma^2_{\Delta V_{th0}}$ is the variance due to process variation at the begining of life of the SM. The SM is assumed to consist of an equal number of PMOS and NMOS transistors. The process induced variation in V_{th} is identical for the baseline and the PSO based SM.

$$\Delta V_{th}(t) = \mathcal{N}(\langle \Delta V_{thA}(t) \rangle, \sigma_{\Delta V_{th}}(t)) \tag{5}$$

$$\langle \Delta V_{thA}(t) \rangle \cong A e^{-\kappa/\theta} t^{\alpha} E_{OX}^{\gamma} df^{\beta} \tag{6}$$

$$\sigma^2_{\Delta V_{th}}(t) = \left(1 + \frac{\langle \Delta V_{thA}(t) \rangle}{0.1\ V}\right) \sigma^2_{\Delta V_{th0}} \tag{7}$$

The aging effect on the PMOS header of an OCVR implemented as a linear dropout regulator (LDO) is also accounted for when determining the effect on V_{th} due to the aging of the SM. As the output voltage from the LDOs is modulated by the PSO, both the electric field across the gate oxide (E_{OX}) of the load circuits, including the distributed timing sensors, and the operating temperature change. The updated temperatures across the SM are obtained from Hotspot. The rate of aging for the baseline SM and the SM implementing PSO voltage assignment is calculated using the aging model given by (6) for the same stress time t and duty factor df. The variation in the V_{th} of the SM designed in a 7 nm FinFET process [12] at the start of life, end of life (EOL) for the baseline SM, and EOL for the SM with adaptive voltage assignment through the PSO are shown in Fig. 8, where the EOL for both is 10 years. For the analysis, the G_{best} is applied to the 12 OCVRs of the SM. Despite accounting for high power supply noise (10% on the FP/INT cores), the cumulative effect with time of the adaptive power supply voltage results in a significant reduction in the rate of transistor aging, with a mean reduction of 40% in $\langle \Delta V_{thA}(t) \rangle$. The large improvement is due to the reduction in the applied electric field across the oxide E_{OX} as compared to the base line. The reduction in temperature due to a lower applied voltage marginally improves the aging characteristics of the circuit. The reduced degradation in V_{th} due to aging is shown for both high-V_{th} and low-V_{th} 7 nm FinFET devices in Fig. 8(a) and 8(b), respectively.

C. Reduction in operating temperature

The thermal simulator HotSpot 6.0 [17] is used to characterize the effect on the temperature profile of the SM due to the evolving voltage assignment of the circuit by the

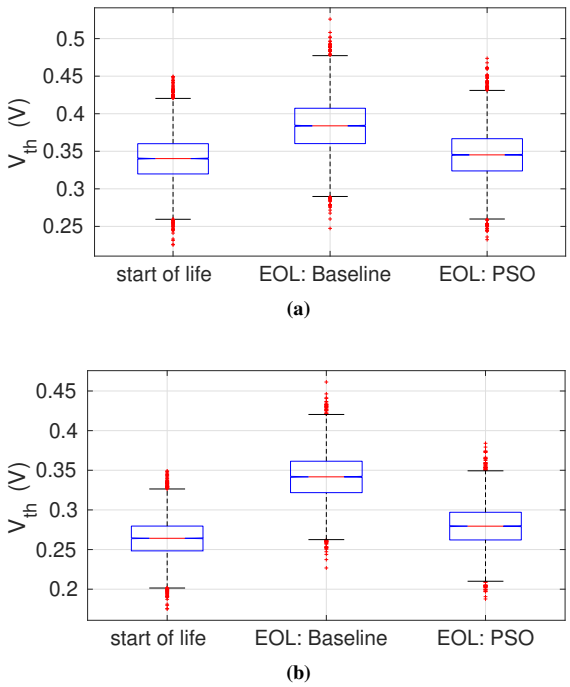


Fig. 8: Reduction in aging induced degradation of V_{th} through adaptive voltage assignment of 12 OCVRs integrated in an SM modeled on the NVIDIA Volta GV100 [1] for (a) a high-V_{th} 7 nm FinFET process and (b) a low-V_{th} 7 nm FinFET process. The end of life (EOL) of the baseline and PSO based SM is ten years.

PSO. Similar to Voltspot, the inputs to Hotspot are the SM architectural floor plan and the power trace. The floor plan of one GPU processing cluster (GPC) with 14 SMs [1] is considered for the temperature analysis. A series of differential equations are iteratively solved by Hotspot to compute the temperatures of the functional blocks. The average temperature of each functional unit is reported as an output. The transient temperatures for the RF, INT/FP cores, and tensor cores are determined through Hotspot by providing an updated power trace file computed using the per clock cycle G_{best} obtained by the PSO for 0.5 million clock cycles. Equal values of the personal (ϕ_1) and social (ϕ_2) acceleration coefficients are chosen for the analysis. The minimum, maximum, and mean temperature changes for each block in the floor plan as compared to the baseline are listed in Table II. An average reduction of 5°C is observed across the RF, INT/FP cores, and tensor cores when a global best voltage assignment is applied to the circuit that is at least 20% less than the nominal voltage recommended for the given technology node. The self heating of the FinFETs is not captured through architectural level simulators including Hotspot. As the self heating of the transistor and the local interconnects is directly proportional to the applied gate voltage, a reduction in self heating is expected when implementing PSO based voltage assignment.

D. Hardware overhead

The overhead in area and compute cycles of implementing and executing the on-line PSO for adaptive voltage assignment

TABLE II: Reduction in per block temperature determined through Hotspot for the SM with voltage assignment set by the PSO.

Functional unit	Reduction in assigned voltage	Reduction in temperature		
		Min	Max	Mean
RF	19.1%	2.9%	5.7%	4.7%
FP/INT cores	21.4%	3.4%	7.2%	6.1%
Tensor cores	22.1%	2.3%	5.1%	4.2%

TABLE III: Circuit and computational overhead to implement the run-time PSO.

Parameter	Value
On-chip LUT for 12 timing sensors	512 B
Execution time of the *PSO*	20 clock cycles per OCVR
Aging sensor per *PCP* [14]	200 transistors
Timing sensor [9]	400 transistors

is summarized in Table III. The circuit and computational overheads are determined for a voltage domain with 16 distributed OCVRs supporting six distinct reference voltage levels, 16 delay line based timing sensors, and 12 *PCPs*. The computational time to process a new P_{best} and G_{best} value for each OCVR is twenty clock cycles. The required size of the LUT increases with the operating age of the IC. In addition, the memory allocated to the LUT is reused after every two years of storing the calibration data. The fastest degradation in the threshold voltage of the OCVR and load circuit occurs in the first two years of the operating life of the IC. Beyond the first two years, the operating temperature of the circuit becomes a more critical parameter than the total stress time of the load circuit. Through the adaptive voltage assignment of the PSO, the reductions in the voltage from the design time nominal value occur less frequently near the end of life of the IC. Therefore, the calibration data from the start of life of the IC do not need to be retained in the LUT.

IV. Conclusion

A methodology that provides an evolving on-chip voltage assignment to distributed on-chip voltage regulators is proposed. The novelty of the method is in the modulation of the local supply voltages to minimize the required voltage guard-band of a circuit while assuring no timing violations occur. The distributed voltage regulators serving a voltage domain work as a swarm and use the local data collected by the timing sensors to compensate for the effects of process variation, transistor aging in both the load circuit and the voltage regulators, and variations in the temperature across the die. The variation in the delay of the timing sensors is dependent on process, aging, temperature, and power supply noise. The sensor, therefore, provides an optimal monitor of the reliability and performance of the circuit. Through simulation of a GPU streaming multi-processor in a 7 nm FinFET technology, an average reduction of 35% and 5% in, respectively, the power consumption and operating temperature of the voltage domains was observed. In addition, the end of life of the circuit increases due to an average reduction of 40% in the aging induced variation in V_{th}.

REFERENCES

[1] NVIDIA, "NVIDIA Tesla V100 GPU architecture," https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf, pp. 1–58, August 2017.

[2] D. Pathak, H. Homayoun, and I. Savidis, "SMART GRID ON CHIP: Work load balanced on-chip power delivery," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 25, No. 9, pp. 2538–2551, September 2017.

[3] M. Tavana, D. Pathak, M. Hajkazemi, I. Savidis, and H. Homayoun, "Realizing complexity-effective on-chip power delivery for many-core platforms by exploiting optimized mapping," *Proceedings of the IEEE International Conference on Computer Design*, pp. 581–588, October 2015.

[4] D. Pathak, *SMART Grid On Chip: Infusing intelligence to on-chip energy management.*, Ph.D. Thesis, Drexel University, 2018.

[5] S.C. Chapra and R.P. Canale, *Numerical methods for engineers*, Boston: McGraw-Hill Higher Education, 2010.

[6] A. P. Da Silva and P. Abrão, "Applications of evolutionary computation in electric power systems," *Proceedings of the Congress on Evolutionary Computation*, Vol. 2, pp. 1057–1062, May 2002.

[7] D. W. Boeringer and D. H. Werner, "Particle swarm optimization versus genetic algorithms for phased array synthesis," *IEEE Transactions on Antennas and Propagation*, Vol. 52, No. 3, pp. 771–779, April 2004.

[8] L. Zhang, Y. Chen, R. Sun, S. Jing, and B. Yang, "A task scheduling algorithm based on PSO for grid computing," *International Journal of Computational Intelligence Research*, Vol. 4, No. 1, pp. 37–43, June 2008.

[9] R. Franch, P. Restle, N. James, W. Huott, J. Friedrich, R. Dixon, S. Weitzel, K. Van Goor, and G. Salem, "On-chip timing uncertainty measurements on IBM microprocessors," *Proceedings of the IEEE International Test Conference*, pp. 1–7, October 2007.

[10] M. S. Floyd, A. Drake, N. S. Schwartz, R. W. Berry, C. R. Lefurgy, M. Ware, K. Rajamani, V. Zyuban, R. Willaman, and R. M. Zgabay, "Runtime power reduction capability of the IBM POWER7+ chip," *IBM Journal of Research and Development*, Vol. 57, No. 6, pp. 1–17, November 2013.

[11] J. Park and J. A. Abraham, "A fast, accurate and simple critical path monitor for improving energy-delay product in DVS systems," *Proceedings of the International Symposium on Low-power Electronics and Design*, pp. 391–396, August 2011.

[12] Nanoscale Integration & Modeling Group at Arizona State University, "Predictive Technology Model (PTM)," http://ptm.asu.edu/.

[13] X. Huang, W. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y. Choi, K. Asano, *et al.*, "Sub-50 nm p-channel FinFET," *IEEE Transactions on Electron Devices*, Vol. 48, No. 5, pp. 880–886, May 2001.

[14] M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, "Circuit failure prediction and its application to transistor aging," *Proceedings of the IEEE VLSI Test Symposium*, pp. 277–286, May 2007.

[15] R. Zhang, K. Mazumdar, B. H. Meyer, K. Wang, K. Skadron, and M. R. Stan, "Voltspot 2.0," http://lava.cs.virginia.edu/VoltSpot/.

[16] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 21, No. 1, pp. 3–13, February 2008.

[17] R. Zhang, K. Skadron, and M. R. Stan, "Hotspot 6.0," http://lava.cs.virginia.edu/HotSpot/.

[18] J. Leng, Y. Zu, M. Rhu, M. Gupta, and V. J. Reddi, "GPUVolt: Modeling and characterizing voltage noise in GPU architectures," *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 141–146, August 2014.

[19] P. I. Chuang, C. Vezyrtzis, D. Pathak, R. Rizzolo, T. Webel, T. Strach, O. Torreiter, P. Lobo, A. Buyuktosunoglu, R. Bertran, M. Floyd, M. Ware, G. Salem, S. Carey, and P. Restle, "Power supply noise in a 22 nm z13 microprocessor," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 438–439, February 2017.

[20] P. Weckx, B. Kaczer, M. Toledano-Luque, P. Raghavan, J. Franco, P. J. Roussel, G. Groeseneken, and F. Catthoor, "Implications of BTI-induced time-dependent statistics on yield estimation of digital circuits," *IEEE Transactions on Electron Devices*, Vol. 61, No. 3, pp. 666–673, January 2014.

[21] E. Cai, D. Stamoulis, and D. Marculescu, "Exploring aging deceleration in FinFET-based multi-core systems," *Proceedings of the International Conference on Computer-Aided Design*, pp. 111:1–111:8, November 2016.