A Spatial Multi-Bit Sub-1-V Time-Domain Matrix Multiplier Interface for Approximate Computing in 65-nm CMOS

Srinivasan Gopal, *Student Member, IEEE*, Pawan Agarwal, Joe Baylon, *Student Member, IEEE*, Luke Renaud, *Student Member, IEEE*, Sheikh Nijam Ali, *Student Member, IEEE*, Partha Pratim Pande, *Senior Member, IEEE*, and Deukhyoun Heo, *Senior Member, IEEE*

***Abstract*— Large-scale parallel implementation of matrix multiply and accumulate (MAC) core poses significant energy and area constraints in analog voltage domain under reduced supply voltage. A spatial multi-bit sub-1-V time-domain matrix multiplier interface is presented using multi-bit back-gate-driven delay elements as a scalable alternative for various approximate computing applications. A single-chip solution is demonstrated for two application modes: a high-throughput digitally driven mode for acceleration and a low-energy analog front-end mode for sensing. In accelerate mode, the system achieves an aggregate throughput of 21.6 GMAC/s with 9 TOPS/W energy efficiency. In sense mode, the system exhibits an energy efficiency of 55.3 TOPS/W for classification purpose. The proposed architecture utilizes 16-parallel 6-bit input vectors to perform matrix MAC computations using time-domain signal processing with 3-bit resistive weights at a sub-1-V supply of 0.7 V. An integrated speculative time-to-digital converter (is employed for 6-bit time-domain quantization with an on-chip mismatch calibration scheme. The prototype is fabricated in 65-nm CMOS technology and occupies an active area of 0.04 mm^2. The system performs image recognition of handwritten digits using a machine learning scheme and demonstrates an average classification accuracy of 84.3% on the MNIST dataset. The resultant energy per MAC computation in the proposed spatial architecture is about 15× lower than a digital CMOS combinational logic-based parallel-tree MAC.**

***Index Terms*— Time-domain signal processing, VTC, DTC, TDC, machine learning, approximate computing, neuromorphic computing, IoT, MAC, spatial, accelerator.**

I. Introduction

MATRIX multiply-and-accumulation (MAC) core arithmetic units are pervasive in scientific computing, machine learning, and real-time signal processing.

Manuscript received January 7, 2018; revised May 15, 2018; accepted June 1, 2018. Date of publication July 3, 2018; date of current version September 11, 2018. This work was supported in part by the U.S. National Science Foundation under Grants CCS-1514269, CNS-1564014, and CNS-1705026, and in part by the NSF Center for Design of Analog–Digital Integrated Circuits. This paper was recommended by Guest Editor A. Marongiu. *(Corresponding author: Deukhyoun Heo.)*

S. Gopal, J. Baylon, L. Renaud, S. N. Ali, P. P. Pande, and D. Heo are with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99163 USA (e-mail: dheo@wsu.edu).

P. Agarwal was with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99163 USA. He is now with Maxlinear Inc., Carlsbad, CA 92008 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JETCAS.2018.2852624

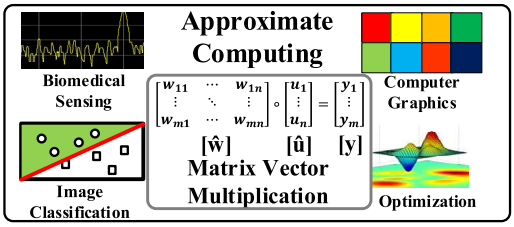


Fig. 1. Approximate computing applications with matrix multiply-accumulate (MAC) operation as core unit.

Fig. 1 shows potential applications of approximate computing for sensing and high-performance computing that can function under reduced precision for approximate computing [1]–[10], such as computer vision, speech recognition, biomedical sensing, neuromorphic computing, acceleration etc. [3], [6], [10]. These applications can function at low precision motivating the use of MAC arithmetic operations with analog signal processing as an energy-efficient alternative to existing approaches [9]. By implementing MAC arithmetic operations closer to the extreme edge of sensor networks, far from the cloud access, faster results can be realized in system designs by reducing communication bandwidth requirements to central processing nodes. MAC embedded front-end computation units thus reduce latency of cloud access providing energy-efficient transmission and processing [1]–[4].

The parameters to consider in the design of MAC core are computation speed (throughput), energy-efficiency, area, dynamic range and resolution. Depending on the application, some parameters are given higher relative importance to others. For instance, image classification is performed in [6] for inference using real images from the CIFAR-10 database using machine learning with 6-bit input data. A self-calibrating accelerator for received signal separation from noise is developed in [10] for GPS acquisition using 2-bit input data. A neuromorphic computing core is developed in [16] using 1-bit input data for handwritten digit recognition.

In analog signal processing, reduced supply voltage offers limited signal-to-noise ratio (SNR) with technology scaling. Maintaining dynamic range under a reduced supply voltage necessitates mismatch and thermal noise reduction

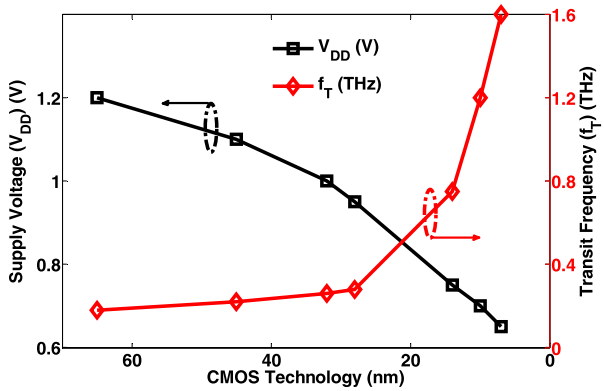


Fig. 2. ITRS Technology trend across CMOS process generations.

using large-sized passive or active devices. Consequently, reduced supply voltage leads to increased chip area and larger power consumption across multiple parallel channels in the analog-domain [12], [15]. Analog-domain implementation of a multiple parallel-input spatial MAC architecture requires power-hungry and physically large digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), to interface with the digital environment.

MAC implementation in the analog-domain therefore offers diminishing performance returns with technology scaling trends. However, an increase in transition frequency (f_T) has improved transistor speed resulting in enhanced resolution in the time-domain. The increase in resolution in time is illustrated in Fig. 2, which plots the supply voltage on the left Y-axis (in black), and the f_T on the right Y-axis (in red) with technology scaling. The increase in f_T presents an opportunity for time-domain signal processing to operate under a reduced supply voltage for increased energy efficiency leveraging technology scaling trends [18]. Limited electronic design automation tools for analog-domain make digital implementation of time-domain signal processing circuits more favorable in terms of design scalability of multiple parallel inputs. Thus, time-domain signal processing using digital implementation is an alternate scalable technique substituting its analog counterparts in nanoscale CMOS processes across different applications such as digital phase-locked loops [19] and delay-line based ADCs [18].

This work demonstrates an approximate computing circuit with two operating modes using 16-parallel, 6-bit back-gate-driven time-domain matrix multiplier as a scalable interface which operates under a 0.7 V supply voltage in 65 nm CMOS technology. The approximate computing modes can broadly be classified into two application modes based on the nature of input data vectors (u_i).

1) accelerate-mode: We consider u_i as the multi-bit digital vectors where the proposed system co-processes matrix computations in time-domain by interfacing with digital environment applicable for hardware accelerators [10].
2) sense-mode: In this mode, u_i are inherent analog input vectors where the proposed system functions directly as an analog interface, performing machine-learning classification in the time-domain for direct inference from sensors [27].

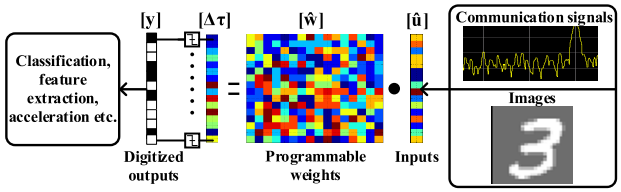


Fig. 3. Time-domain matrix MAC for various signal processing applications.

A brief description of related work is presented in Section II. In section III, the proposed architecture of spatial multi-bit time-domain matrix multiplier interface is presented. Section III presents the architecture details of proposed speculative TDC. An on-chip rising-edge mismatch calibration mechanism is presented in Section IV. Measurement results for accelerate-mode and sense-mode with an application demonstration are presented in Section V, and Section VI concludes this paper.

II. Related Work

In the recent past, embedded analog-domain MAC computation front-end systems have increasingly gained wide attention [9]–[11], [15], [27]. Analog-domain MAC implementations can be classified as active and passive approaches. Active approaches are based upon transconductance cells that use weighted current-mode summing [8]–[10]. As the multiplication dramatically increases dynamic range requirements, active approaches have a severe energy cost to meet them limited by the voltage/current headroom against the noise level [27]. On the other hand, passive approaches are built using switched capacitor-based MAC units, implemented by charge scaling and addition [5]–[7]. However, with high levels of parallelism for large scale design implementation, charge-redistribution is subject to non-idealities such as gain error and signal level degradation. Charge redistribution limits dynamic range in spatial architecture of switched capacitor-MAC restricting large scale implementation [12]. These non-idealities are aggravated with technology scaling. In the analog-domain, it is quite challenging to implement a power- and area-efficient MAC for a multi-bit parallel-input spatial architecture.

Alternately, a time-domain implementation offers robust computation with readily interfaced digital circuits [15]–[17]. As illustrated in Fig. 3, the MAC operations are performed in time-domain on the elements of multi-bit input vector (u) with elements of programmable weight vector (w), and the resultant output data ($\Delta\tau$) undergoes time-based quantization process to obtain a digital output (y) for various signal processing applications. However, conventional time-domain techniques are digitally driven by standard combinational logic gates limiting their application space to digital interfaces, and unsuitable for analog sensing. Conventional single-bit time-domain implementations using resistive and capacitive techniques have design challenges and associated trade-offs [15]–[17]. Resistive time-domain technique is employed in [15] using 5-transistor stacked logic limiting high speed operation at sub-1V due to parasitic capacitors present in stacked signal path of devices. Capacitive time-domain technique using capacitor-bank is applied in [17]. The area and energy

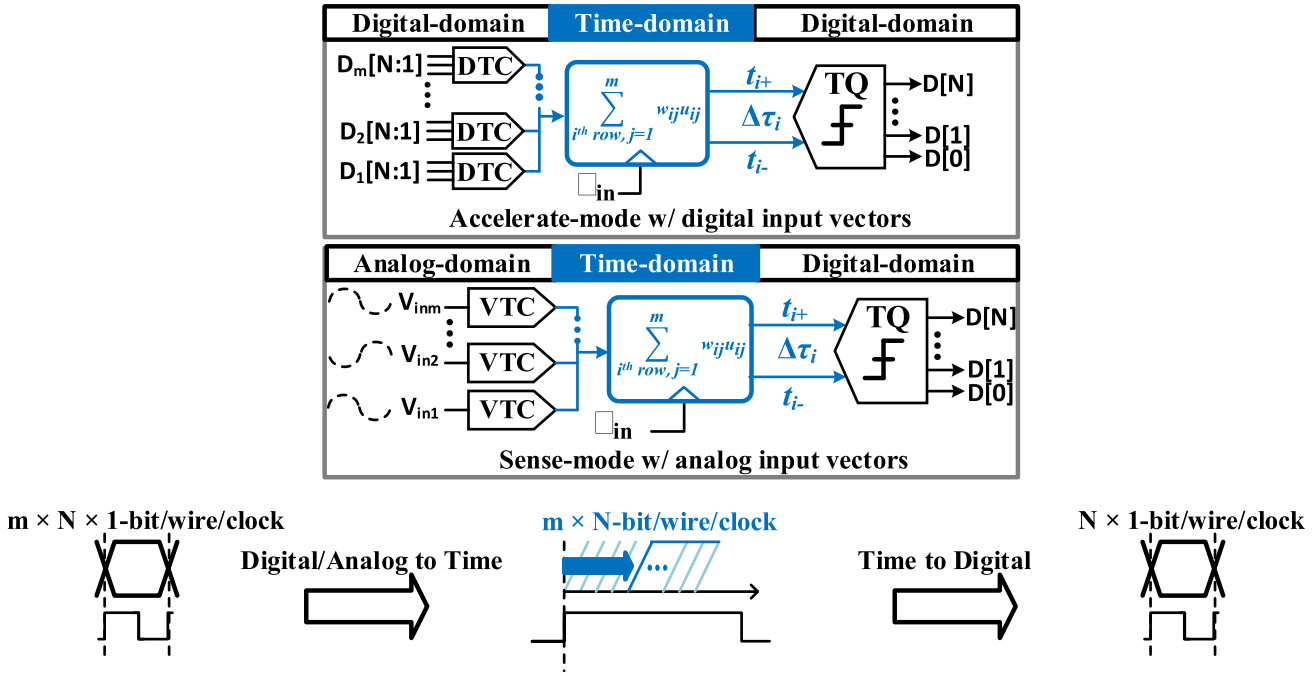


Fig. 4. Spatial multi-bit time-domain MAC for accelerate- and sense- computing modes with information encoded as time-difference.

consumption increase exponentially for N-bit weight programmability in time-domain MAC implementation, limiting speed, energy- and area- efficiency as shown in (1) and (2), where C_u is a unit element capacitor and V corresponds to the voltage swing.

$$Area = 2^N \times C_u \tag{1}$$

$$Energy = 2^N \times C_u \times V^2 \tag{2}$$

This work aims to expand the time-domain computing approach to multiple signal processing environments by addressing the limitations of conventional approaches. Multi-bit time-domain computing with back-gate technique is introduced to span wide range of applications from high-performance computing to analog front-end sensing, and achieve multi-mode functionality. The core computing delay element in our chip implementation is scalable in terms of the number of inputs. Although our prototype implements a unit-row matrix computation demonstrating our basic principle, multiple parallel-unit implementation can increase system throughput.

III. Proposed Spatial Multi-Bit Time-Domain MAC Interface Architecture

Fig. 4 illustrates the fundamental idea of spatial multi-bit time-domain signal processing of the proposed spatial multi-bit time-domain MAC interface architecture in which the input data (analog/digital) is converted to time difference. The matrix MAC arithmetic operation using appropriate weights is then performed in time-domain. Finally, the processed output undergoes time-based quantization process. Thus time-domain signal processing uses delay difference ($\Delta\tau_i$) to encode the accumulated signal information.

The spatial MAC architecture offers high amounts of data parallelism using m parallel N-bit digital (voltage) -to-time converters (DTC/VTC) to convert the digital (analog) signal into time-domain in accelerate-mode and sense-mode, respectively (see Fig. 4). The matrix MAC operations are performed on the elements of input vector (u_{ij}) with elements of programmable weight vector (w_{ij}), to obtain the resultant accumulated time-domain output ($\Delta\tau_i$) for an i^{th}-row MAC operation as shown in (3):

$$\Delta\tau_i = \sum_{j=1}^{m} w_{ij}u_{ij} \tag{3}$$

The size of the weight vector matrix is $n \times m$, where the number of MAC operations is m. This operation is repeated n times for n rows using programmable matrix (w_{ij}) as shown in (4):

$$\begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{bmatrix} \times \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} w_{1j}u_{1j} \\ \vdots \\ \sum_{j=1}^{m} w_{nj}u_{nj} \end{bmatrix} = \begin{bmatrix} \Delta\tau_1 \\ \vdots \\ \Delta\tau_n \end{bmatrix} \tag{4}$$

The resultant time-domain data ($\Delta\tau_i$) undergoes time-based quantization process using a time-to-digital converter (TDC) for obtaining digitized output.

A. Proposed Spatial Multi-Bit DTC: Back-Gate Driven Time-Domain Matrix Multiplier Interface

The proposed spatial multi-bit DTC architecture is based on a back-gate-driven time-domain matrix multiplier

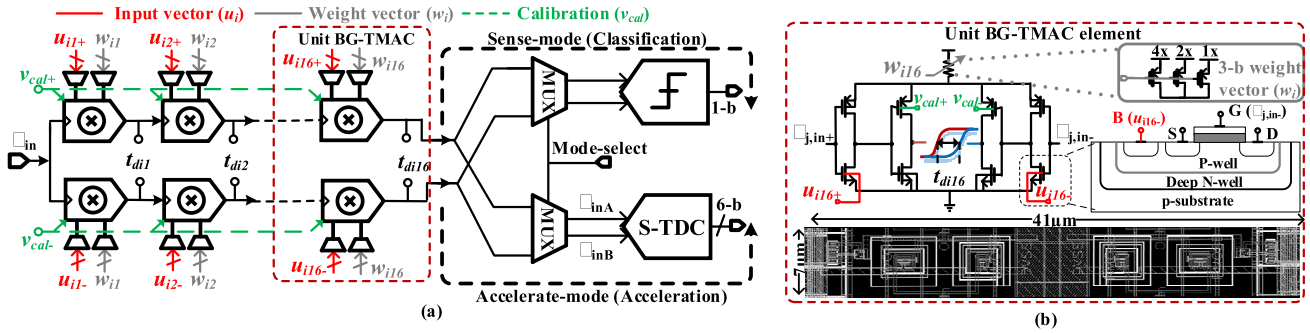


Fig. 5. (a) Proposed spatial multi-bit time-domain back-gate matrix multiplier interface (BG-TMAC) supporting sense-mode and accelerate-mode. (b) BG-TMAC unit element circuit and layout.

interface (BG-TMAC) supporting multi-mode approximate computing applications, as shown in Fig. 5(a). An input vector size of 16 parallel pseudo-differential signals (u_{ij}) are given to bulk terminals of NMOSs placed in deep N-well layers for noise immunity from digital substrate switching noise. The clock signal (Φ_{in}) is fed into the gate of MOSFETs. The delay calibration differential signals (v_{cal}) are given through PMOS bulk terminals. The elements of weights vector (w_{ij}) are applied through resistive PMOS switches to 3-bit programmability. The delay difference at each parallel input vector is cascaded enabling accumulation process. The BG-TMAC architecture thus offers systematic accumulation in time-domain in an asynchronous manner. The unit element circuit and the layout of BG-TMAC circuit are shown in Fig. 5 (b). The accumulated output phase difference carries the output MAC value in the time-domain. The time-domain output quantization mode is selected by two MUX blocks controlled by a mode-select control signal. In sense-mode computing, the mode-select allows the time-domain output to be quantized by a flip-flop for binary classification [16], [27]. While in accelerate-mode computing, the mode-select allows the time-domain output to be quantized by a 6-bit speculative TDC.

B. Linear Back-Gate-Driven Voltage-to-Time Conversion

Back-gate driven technique for generating linear delay characteristics is used in ultra-low power oscillators [29] and linear fine delay element in Vernier TDC applications [30]. This sub-section derives the back-gate voltage-to-time transfer function (VTTF) characteristic used for the design of each delay element. The VTTF characteristics is given by the weight vector derivation (w_{ij}) of unit delay element. For a unit delay element with index (i, j), the input is given to the bulk terminal and the weight factor is modified using 3-bit programmable supply regulation, that results in an output delay difference. The weight vector derivation and its simulation give an insight on the delay behavior with the change in input bulk voltage.

Fig. 6(a) shows the transition diagram of proposed delay element. The propagation delay (τ_{pd}) of a unit delay cell is given by the total output capacitance (C_L). Assuming fast rise\fall signal transitions (an order of magnitude faster compared to the clock period), a unit delay cell's τ_{pdij} is given by (5) [31]:

$$\tau_{pdij} = \frac{C_L}{g_{m,maxij} + g_{mb,maxij}} = \frac{C_L}{g_{mT,maxij}} \tag{5}$$

Here, the total load capacitance (C_L) constitutes both external loading capacitance and internal parasitic capacitance. While $g_{m,maxij}$ and $g_{mb,maxij}$ represent the maximum gate- and bulk-transconductance of the NMOS (PMOS) while operating in strong-inversion region for rise (fall) transition of the input signal. The $g_{mT,maxij}$ and V_{THij} for an NMOS (PMOS) of a unit delay element with index (i, j) using alpha (α)-power law, is shown in (6) and (7) [31]:

$$g_{mT,maxij} = \frac{g_{m0}}{2^{\alpha-1}}\left(1 + \frac{\gamma}{2\sqrt{|2\Phi + V_{SBij}|}}\right)\left[1 - \frac{(\alpha-1)V_{THij}}{V_{DD}}\right] \tag{6}$$

$$V_{THij} = (V_{TH0} - \gamma\sqrt{|2\Phi|}) + \gamma\sqrt{|2\Phi| + V_{SBij}} \tag{7}$$

g_{m0} is a proportionality constant, γ and Φ are technology dependent parameters, and V_{TH0} is V_{TH} at $V_{SBij} = 0$.

For a small-signal source-to-bulk voltage excitation (Δv_{SBij}), around the common mode V_{BCMopt}, ($V_{SB} = V_{BCMopt} + \Delta v_{SBij}$), the derivative of (5) with respect to Δv_{SBij} gives the expression in (8):

$$\frac{\Delta\tau_{pdij}}{\Delta v_{SBij}} = \frac{-C_L}{g_{mT,maxij}^2} \times \frac{\Delta g_{mT,maxij}}{\Delta v_{SBij}} \tag{8}$$

While $\Delta g_{mT,maxij}$ represents the change in the maximum transconductance, which corresponds to (9) where $\Delta g_{m,maxij} = 0$, corresponding to the only change in bulk-to-source transconductance due to small signal bulk-to-source voltage excitation (Δv_{SBij}).

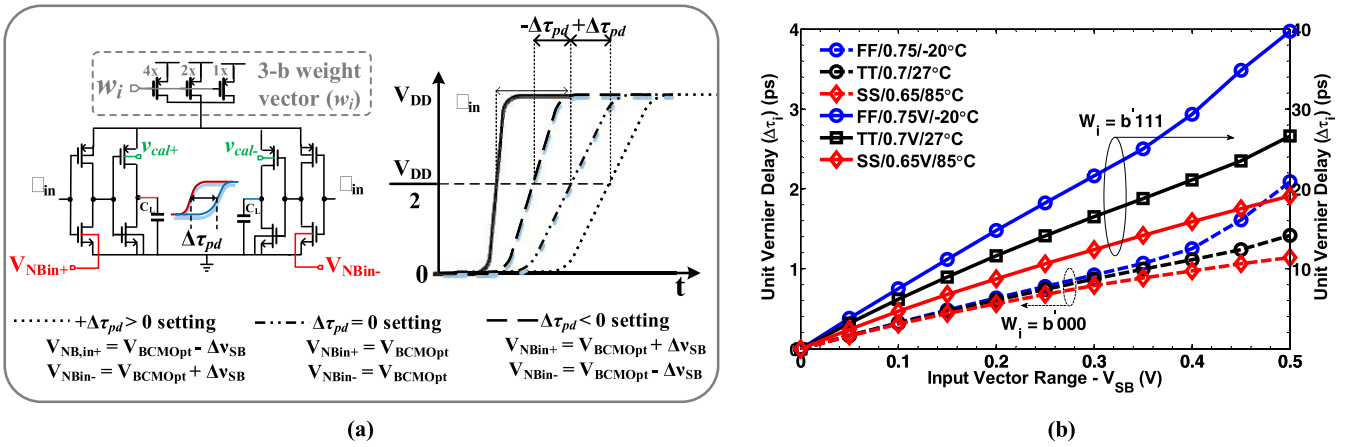
$$\Delta g_{mT,maxij} = \Delta g_{m,maxij} + \Delta g_{mb,maxij} \tag{9}$$

The value of $\Delta g_{mb,maxij}$ can be evaluated as in (10):

$$g_{mb,maxij} = \frac{\gamma\, g_{m,maxij}}{\sqrt{(|2\Phi| + V_{SBij})}} \tag{10}$$

Obtaining the derivative of (10) due to the small signal excitation (Δv_{SBij}) results in (11):

$$\Delta g_{mb,maxij} = \frac{-\gamma\, g_{m,maxij}\, \Delta v_{SBij}}{2(|2\Phi| + V_{BCMopt})^{\frac{3}{2}}} \tag{11}$$



(a) (b)

Fig. 6. (a) Bi-directional delay difference using pseudo-differential bulk-driven inputs and its transition diagram. (b) Simulated piecewise linear back-gate-driven voltage-to-time characteristics for minimum and maximum weights.

From (8), (9) and (11), the delay difference due to bulk driven voltage excitation is derived in (12):

$$\Delta\tau_{pdij} = \frac{-C_L}{g^2_{mT,maxij}} \times \frac{-\gamma\, g_{m,maxij}\Delta v_{SBij}}{2(|2\Phi| + V_{BCMopt})^{\frac{3}{2}}} \tag{12}$$

As per (12), a piecewise linear delay function can be obtained by bulk-driven voltage excitation (Δv_{SBij}). Thus the generated unit weight vector (w_{ij}) is given by the amount due to the bulk-to-source input signal ($u_{ij} = \Delta v_{SBij}$) is computed as in (13):

$$w_{ij} = \frac{-C_L}{g^2_{mT,maxij}} \times \frac{-\gamma\, g_{m,maxij}}{2(|2\Phi| + V_{BCMopt})^{\frac{3}{2}}} \tag{13}$$

Fig. 6(b) shows the simulated unit element voltage-to-time transfer function (VTTF) for minimum ($w_{ij} = b'000$) and maximum ($w_{ij} = b'111$) weight settings respectively, across process-voltage-temperate (PVT) variations.

An important requirement of MAC in time domain is the implementation of signed arithmetic circuits. The bulk-driven MOSFET modulates the sign and magnitude of delay difference, based on applied body bias differential voltage, as illustrated in Fig. 6(a) [29]. The NMOS devices are body biased at a common mode voltage ($V_{BCM,Opt}$) to maximize the linear range in delay. Then an incremental excitation ($+/-\Delta v_{SB}$) can be applied to $V_{BCM,Opt}$ in order to achieve bi-directional delay difference corresponding to the signed magnitude.

C. Sensitivity Analysis of Delay Function Using Bulk-Driven Voltage

Applying small-signal analysis for the NMOS bulk potential around the optimum common mode V_{BCMopt}, ($V_{SB} = V_{BCMopt} + \Delta v_{SB}$), the slope of the change in delay is shown in (14) on the derivative of (12) and (14) indicates a uniform step size of $\Delta\tau_{pdij}$ can be obtained for a given common mode bias at V_{BCMopt}.

$$\frac{\partial\Delta\tau_{pdij}}{\partial\Delta v_{SBij}} = \frac{-C_L}{g^2_{mT,maxij}} \times \frac{-3\gamma\, g_{m,maxij}}{4(|2\Phi| + V_{BCMopt})^{\frac{5}{2}}} \tag{14}$$

To obtain the sensitivity of bulk-driven voltage control, second derivative of change in delay on (14) establishes zero small-signal sensitivity ($\partial^2\Delta\tau_{pdij}/\partial\Delta v^2_{SBij} = 0$). Intuitively, the transconductance increases with a decrease in MOSFET threshold voltage (V_{THij}) and *vice versa*. Thereby, a higher $g_{mT,maxij}$ results in a faster charging current at the output capacitor, hence achieving a smaller $\Delta\tau_{pdij}$ and *vice versa*.

IV. Proposed Speculative TDC

In a time-domain MAC architecture, the accumulated delay-difference is quantized using a TDC. TDCs find ubiquitous applications such as in digital PLLs [20], high resolution PET imaging [33], delay-line analog-to-digital converters (ADC) [18], etc., as time-based quantization can be superior to voltage-based quantization in nanoscale CMOS technologies [20].

Conventional Vernier TDC requires 2^N-1 time-comparators (arbiters or flip-flops) for resolving N-bit resolution. Consequently, the conventional architecture results in a design with large power consumption and low area-efficiency. To overcome this issue, binary-search and SAR based TDCs are explored in [15] and [34]. However, there is an associated wait time (T_{wait}) added at each stage in order to match the delay of a flip-flop at each stage of the conversion. For an N-bit TDC, this imposes a severe constraint on the overall speed of TDC as T_{wait} accumulates along the delay line generating a total accumulated latency of $N \times T_{wait}$ for an N-bit binary-search based TDC as illustrated in Fig. 7 (a). Based on simulations, a 30% to 45% throughput reduction in system is expected for a 6-bit resolution due to the accumulated latency.

A speculative TDC is proposed based on a speed-enhanced speculative binary-search algorithm with N comparators while eliminating the delay due to the flip-flop (T_{wait}). As shown in Fig. 7(b), for an i^{th} stage time-to-digital conversion, the speculative binary search algorithm speculates by pre-calculating both the positive delay-difference ($+2^i\Delta\tau$) and negative-delay difference ($-2^i\Delta\tau$) in parallel to comparator's output decision. The output of the comparator selects the appropriate delay (positive or negative) based on the output

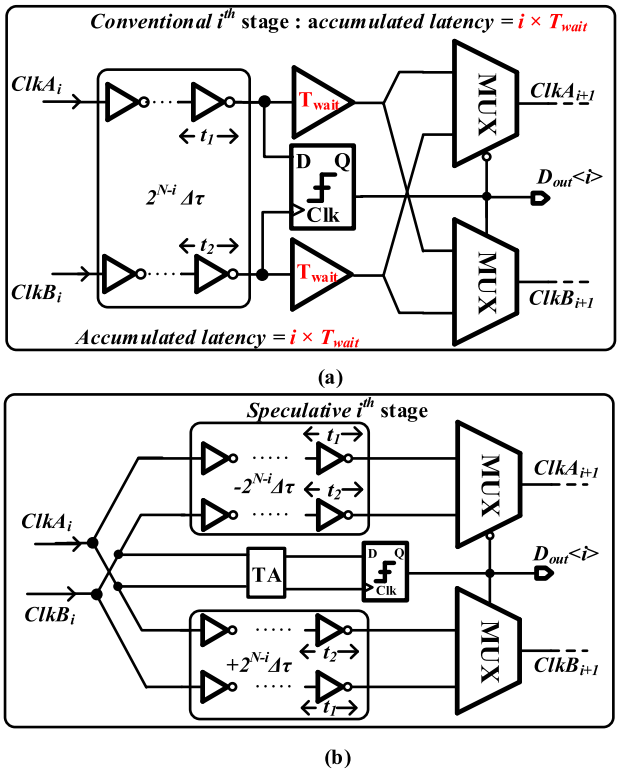


Fig. 7. (a) Conventional binary search based TDC architecture at i^{th}-stage. (b) Proposed speculative TDC architecture.

decision that reduces the delay difference by an amount of $2^i \Delta\tau$ for the subsequent stage quantization. Parallel comparison with delay propagation facilitates a faster conversion process compared to conventional serial comparison.

Conventional TDC systems employ sense-amplifier based flip-flops for achieving fine resolution in making 1-bit decisions without any meta-stability [22]. This becomes more severe while operating in sub-1V voltage regime. To resolve the meta-stability issue, a time-amplifier (TA) is placed before sense-amplifier to increase the input-referred resolution. Time-amplifiers are conventionally used in a multi-stage TDC [35]. While, the proposed speculative TDC uses time amplification as a method to expedite 1-bit decisions of sense-amplifier flip flops.

As the clock-signal has a finite propagation delay through the delay chain, delay between each comparator decision has to be synchronized in order to correctly sample the digital output bits. A clock re-timing circuit is used to synchronize the digital outputs correctly. A second sampling stage is designed to re-sample the outputs in a synchronous manner using a retimed clock on another set of flip-flops. This technique maintains setup and hold time requirements for obtaining the correct samples [20].

The speculative TDC architecture employs the same delay elements as the BG-TMAC circuit described in Section II, to ensure PVT tracking. This is because the relative edge between *ClkA* and *ClkB* signals remain the same at the comparator input employing 1-bit decisions [15]. The simulated linearity performance of speculative TDC across PVT corners are plotted in Fig. 8. Linearity performance is evaluated from differential non-linearity (DNL) and integral non-linearity (INL) estimations. DNL represents the measure of differential error in delay-difference, while INL represents the cumulative error in delay-difference. As per Fig. 8, the worst-case integral non-linearity (INL) is about 0.7 LSB error in the slow corner at 75°C under a supply of 0.65 V.

V. On-Chip Rising-Edge Mismatch Calibration Scheme for Spatial Time-Domain MAC Architecture

Device sizing in delay elements can severely impact non-linearity performance of delay line based VTC, DTC and TDC systems [22]. For instance, it is possible to use minimum sized devices to facilitate low power and compact design, which can also result in large inter-device mismatch in a cascaded delay line. As a result, the linearity of delay-line is distorted in the cascaded delay line for multi-bit parallel-input spatial architecture. This section details the mismatch analysis of spatial time-domain architecture based on the length of delay line. Further, an on-chip calibration mechanism to correct for mismatch is presented.

A. Non-Linearity of Parallel-Input BG-TMAC Delay Elements

Global delay variations due to PVT fluctuations affects all the BG-TMAC delay elements equivalently, resulting in global offset and gain errors. Simple digital calibration can cancel these effects. However, local variations including random dopant profiles and V_{TH} differences causes mismatch in the unique delay behavior of each delay element. As the length of parallel-input based delay line increases, the accumulated mismatch error in delay difference increases. The mismatch error is the total effect of variance in local delay variations. For an m-parallel-input delay line, assuming σ_j is the standard deviation at j^{th} input point, the total accumulated variance σ_T^2 is given by (15):

$$\sigma_T^2 = \sum_{j=1}^{m} \sigma_j^2 \tag{15}$$

For uncorrelated local delay variation in delay elements, $\sigma_j = \sigma_D, \forall j \in m$. Thus, (15) reduces to (16):

$$\sigma_T^2 = m \times \sigma_D^2 \tag{16}$$

Thus, the total delay line's standard deviation (σ_T) grows in proportion to $\sqrt{m}$ for an m-parallel-input delay line. This effect can be seen in the Monte-Carlo simulations plotted in Fig. 9 performed on the delay lines using 65-nm CMOS devices. The error accumulation 2-D map shows the extent of variation as the length of the delay line increases. As per the histogram plot in Fig. 10, a 32-parallel-input delay line has a standard deviation of about 5 times compared to a 2-parallel-input delay line. The Monte-Carlo simulation provides an insight on the accumulated mismatch error in delay differences by scaling the number of parallel inputs, as shown in Fig. 10. Thus, calibration is required to correct mismatch errors in delay differences that causes non-linearity. In our design,

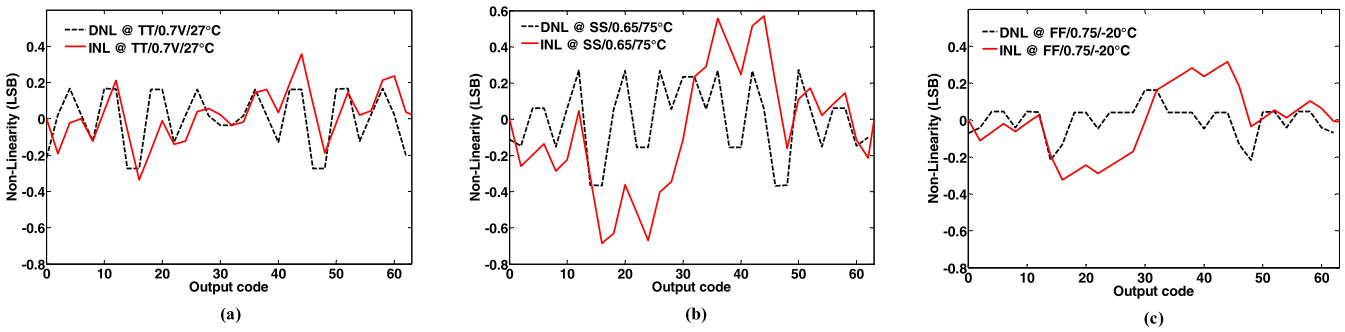


Fig. 8. Simulated non-linearity performance of speculative TDC across process, voltage and temperature (PVT) variations (PVT corner information: typical – TT/0.7V/27°C, slow- SS/0.65V/75°C, fast – FF/0.75/-20°C.

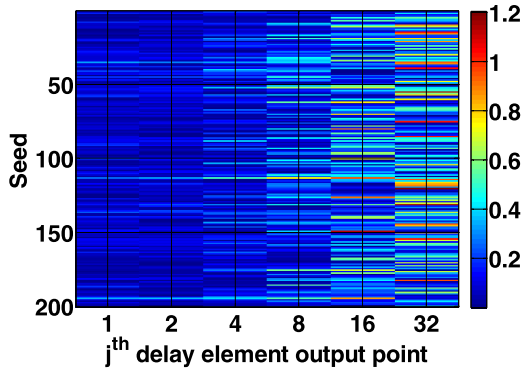


Fig. 9. Mismatch analysis of growing delay line length using Monte Carlo simulations with different seed values.

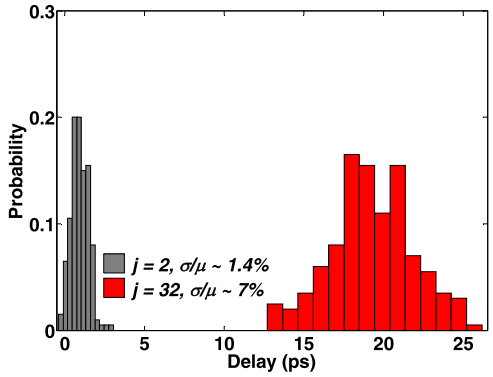


Fig. 10. Histograms at delay element position 2 and 32.

we implemented a 16-parallel-input architecture in consideration of chip area and limited availability of calibration probe pads in measurements.

B. On-Chip Rising Edge Mismatch Calibration

In this sub-section, calibration coefficient is derived based on the amount of mismatch error in delay difference. On-chip calibration is performed for correcting non-linearity due to mismatch errors in delay differences within the chip. In the proposed calibration scheme, PMOS bulk inputs provide the required voltage for on-chip mismatch calibration by using PMOS bulk voltage-to-time transfer function at the rising edge of the signal transition. The amount of required voltage is adaptive depending on the delay correction for a given test chip. This is given by (17) similar to the derivation in (12).

$$\Delta\tau_{pd,calibration} = \frac{+C_L}{g^2_{mpmosT,max}} \times \frac{\gamma\, g_{mpmos,max}\Delta v_{cal}}{2(|2\Phi| + V_{BCMPopt})^{\frac{3}{2}}} \tag{17}$$

In (17), the PMOS bulk-transconductance using bulk input voltage (Δv_{cal}) is leveraged for delay correction by an amount of $\Delta\tau_{pd,calibration}$ and the common mode voltage is set at an optimum value ($V_{BCMPopt}$).The BG-TMAC architecture has multiple parallel inputs with a cascaded delay line of n delay elements.

For an n-parallel input BG-TMAC architecture, with cascaded n delay elements, the actual delay-difference of j^{th} delay element along a row is given by (18):

$$D_j = D + \Delta D_j \tag{18}$$

In (18), D is the nominal delay-difference of the delay element, and ΔD_j is the random error due to mismatch effect that has zero-mean random error with standard deviation σ_D. As D is a measure of unit least significant bit (LSB), the differential non-linearity (*DNL*) and the integral non-linearity (*INL*) values are obtained as shown in (19) and (20) for k delay elements as j runs from 1 to k.

$$\text{DNL[k]} = \frac{\Delta D_j}{D} \tag{19}$$

$$\text{INL[k]} = \sum_{\text{j=1}}^{\text{k}} \frac{\Delta D_j}{D} \tag{20}$$

DNL[k] represents the measure of differential error in delay-difference, while INL[k] represents the cumulative error in delay-difference, with standard deviations $\sigma_{DNL[k]} = \sigma_D/D$ and $\sigma_{INL[k]} = \sqrt{k}\sigma_D/D$, respectively.

The total delay including the delay variation is given by (21):

$$\Delta\tau_{in} = nD + \sum_{j=1}^{n} \Delta D_j \tag{21}$$

Mismatch calibration is performed such that the total delay of the chain given by (21) is adjusted such that $\sum_{j=1}^{n} \Delta D_j$ is

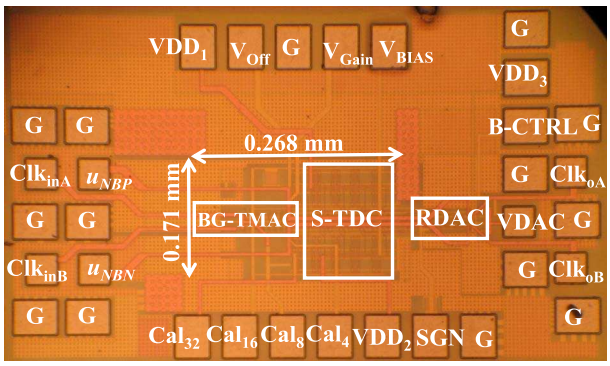


Fig. 11. Chip micrograph of spatial multi-bit BG-TMAC interface.

minimized. ΔD_n is adjusted to a reference nD. Thus, the actual delay of j^{th} delay element becomes the expression in (22):

$$D_j = \frac{nD \cdot (D + \Delta D_j)}{nD + \sum_{j=1}^{n} \Delta D_j} \tag{22}$$

The normalized calibration coefficient (Cal[k]) is evaluated based upon the residual delay correction word for kth delay element given by DNL[k]. Cal[k] can be approximated to (23) from (22) which is normalized to nominal delay-difference D.

$$\text{Cal}\,[k] = \frac{\Delta D_k}{D} - \frac{1}{k}\frac{\sum_{j=1}^{k} \Delta D_j}{D} \tag{23}$$

The entire delay line with multiple inputs is calibrated using the calibration coefficients generated from (23). It is noted that the amount of calibration correction increases as index-k runs from 1 through n along the length of delay line.

VI. System Measurement Results and Application Demonstration

A 16-parallel-input prototype of spatial multi-bit BG-TMAC interface is fabricated in a 65 nm CMOS GP process with chip micrograph shown in Fig. 11. The interface occupies an area of 0.045 mm^2. The clock signals are given using a Tektronix DTG 5274 that feeds in pseudo-differential BG-TMAC circuit. The offset correction is initially performed at the output of the interface based on the measured delay difference and resultant output code by feeding the same clock signal at the inputs with no phase difference. Experimental demonstration of two modes of approximate computing are described in this section. 6-bit quantization is performed on time-domain output in accelerate-mode which is typically used for co-processing an enormous amount of parallel MAC computations in hardware accelerator applications. 1-bit quantization is performed in sense-mode which is mainly used in machine-learning classifiers for performing direct inference on the data obtained from analog sensors.

A. Accelerate-Mode Computing Measurements

In the accelerate-mode, co-processing and acceleration is performed in time-domain. These operations are conventionally performed in digital domain. Accelerate-mode computing processes digital input vectors clocked at a high speed of 1.35 GHz. The measurement set up is shown in Fig. 12(a). A signal generator using HP 8644A provides the ramp input digital word with a minimum voltage resolution of ~11 mV. To measure the system's average linearity performance, all sixteen multipliers are configured with the same weights and inputs. The total accumulated output delay difference is measured using Tektronix DSA 3800 oscilloscope. Fig. 12(b) plots the measured average output delay difference versus digital word across programmable 3-bit weight vectors (w_i). The output delay range at a weight vector setting of $w_i = \text{b}'000$ is about 30 ps. As the weight vector increases to a value of $w_i = \text{b}'111$, the output delay range is increased to 250 ps. Linearity degrades as the value of weight vector increases. Figs. 12(c) and (d) display the measured un-calibrated and calibrated non-linearity data of BG-TMAC circuit, respectively. The displayed measurement data in Fig. 12(c) and Fig. 12(d) are at the worst-case binary weight vector settings of $w_i = b'111$. The un-calibrated maximum INL at BG-TMAC output is about 1.5 LSB, and on-chip mismatch calibration reduces the maximum INL to 0.52 LSB.

The 6-bit digital output of speculative TDC is captured on an analog output pad (VDAC) using an on-chip 6-bit resistive digital-to-analog converter (RDAC). Figs. 12(e) and (f) plot the measured un-calibrated and calibrated non-linearity data of output code, respectively. Non-linearity is evaluated by comparing the step width of resistive DAC output voltage to a nominal step value, which is evaluated from averaging across output digital codes. The linearity of the complete spatial interface architecture is limited by speculative TDC. The measured un-calibrated INL of the output code is about 2.8 LSB as per Fig. 12(e). Post calibration, it can be seen in Fig. 12(f) that the INL of the output code reduces to 1.35 LSB. The entire system has an effective number of bits (ENOB) of 4.63-bit at 1.35 GHz sampling rate.

The energy per unit MAC computation is about 42 fJ/MAC. The speculative TDC presents a fixed energy overhead of 69 fJ/MAC operation. Overall, the chip exhibits an energy efficiency of 9 TOPS/W including 6-bit quantization by speculative TDC at 0.7 V. The power due to resistive DAC and the output buffers are excluded. The 16-parallel input spatial architecture achieves an aggregate throughput of 21.6 GMACs/s in this mode.

B. Sense-Mode Computing Measurements

Sense-mode computing processes data which is obtained directly from analog sensing inputs. Fig. 13(a) illustrates frequency domain measurement set-up to capture output sinusoidal tones [23]. The square wave clock signal ($\Phi(t)$) has a time domain response given in (24).

$$\phi\,(t) = \frac{4w_i}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}sin\,((2n-1)\,2\pi f_c t) \tag{24}$$

The clock amplitude is controlled by weights (w_i). Matrix multiply operation with back-gate-driven input sinusoid ($u(t) = Asin(2\pi f_{NB} t)$) yields $y(t)$, given in (25).

$$y\,(t) = \frac{4Aw_i}{\pi}\sum_{n=1}^{\infty}\frac{1}{2n-1}sin(4n\pi - 2\pi f_c t)\,sin\,(2\pi f_{NB} t) \tag{25}$$

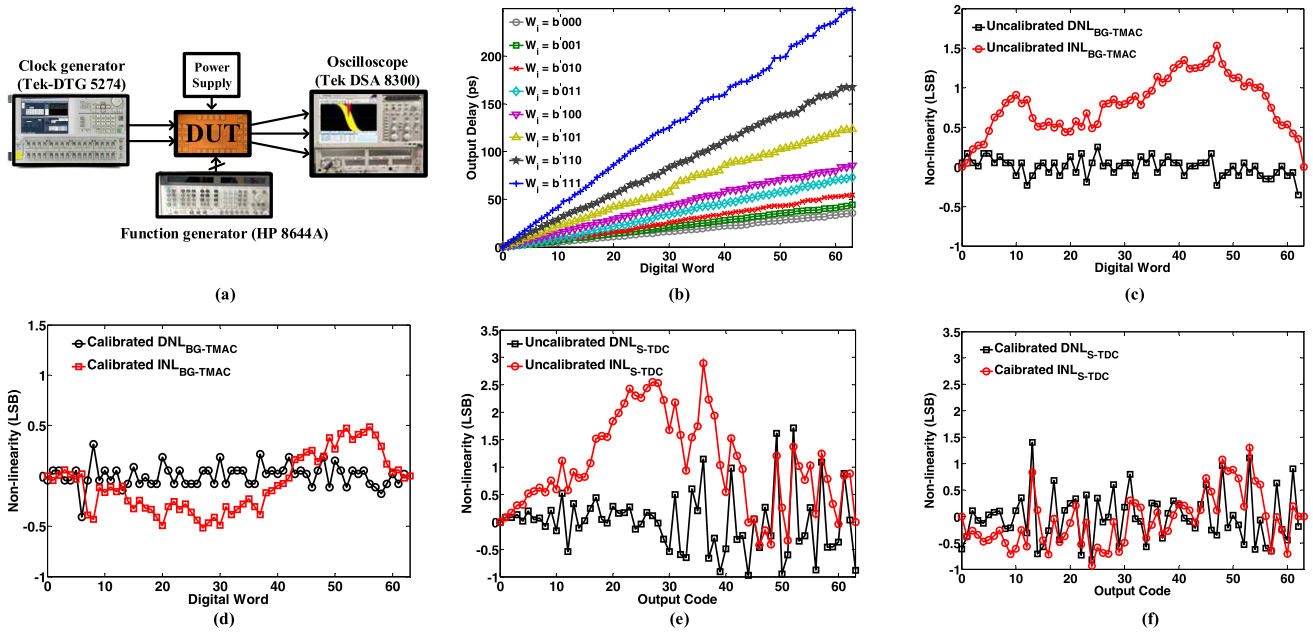


Fig. 12. (a) Accelerate-mode measurement set up. (b) Measured BG-TMAC output delay vs digital word across 3-b reconfigurable resistive weight settings. Linearity performance at BG-TMAC output for $w_i = \text{b}'111$ weight settings for (c) un-calibrated and, (d) calibrated non-linearity measurements. Output code linearity performance at $w_i = \text{b}'111$ weight settings with (e) un-calibrated and, (f) calibrated non-linearity measurements.

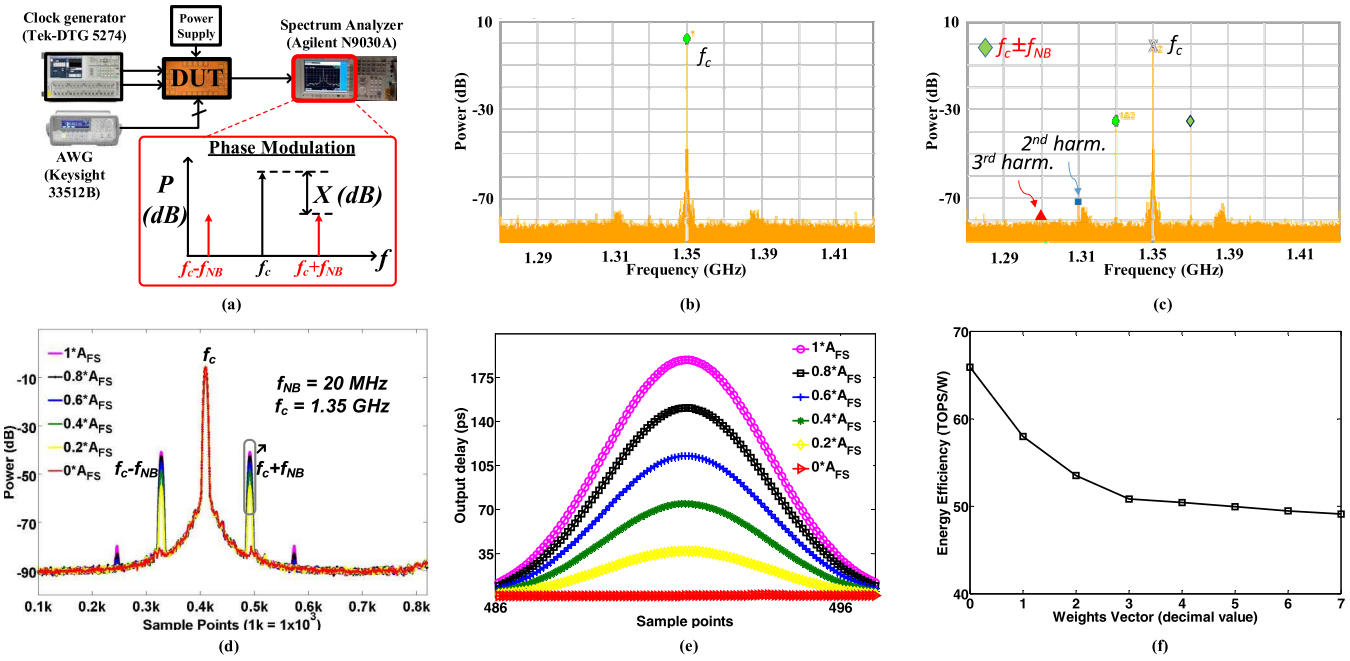


Fig. 13. (a) Sense-mode measurement set up (b) PXA Signal Analyzer Snapshot un-modulated carrier at 1.35 GHz, (c) Calibrated NMOS modulation tone at 20 MHz offsets around 1.35 GHz carrier with reduced 2nd and 3rd harmonic distortions, (c) measured phase modulation and, (e) corresponding ouput delay variation with sinusoidal excitation of variable amplitudes, (f) measured energy efficiency across weight settings.

The expression of output $(y(t))$ in (25) can be re-written as (26) using a standard trigonometric identity.

$$y(t) = \frac{2Aw_i}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \cos(f_{tone+}) - \cos(f_{tone-}) \quad (26)$$

As seen in (26), $y(t)$ is a modulated square wave with two tones at $f_{tone+} = (2n-1)f_c + f_{NB}$ and $f_{tone-} = (2n-1)f_c - f_{NB}$, respectively. Here, f_{NB} is the back-gate input frequency. The amount of modulation is obtained from the spectrum analyzer as illustrated in Fig. 13(a). The spur level difference in log-scale (X dB) between $y(t)$ and clock $(\Phi(t))$ determines the ratio of delay modulation (t_{dij}) with respect to clock edge with time period $(1/f_c)$ shown in (27) [24]:

$$X(dB) = 20 \times log_{10}(t_{dij} \times f_c) \quad (27)$$

The un-modulated carrier of 1.35 GHz (f_c) is captured in Fig. 13(b). Fig. 13(c) displays the fundamental tone of 20 MHz

TABLE I
PERFORMANCE SUMMARY AND COMPARISON WITH RECENT WORKS

References	This Work (BG-TMAC)		JSSC'17 [16]	TCAS-I'17 [27]	JSSC'17 [9]		ASSCC'16 [11]	ISSCC'16 [10]
Technology	65 nm CMOS		65 nm CMOS	130 nm CMOS	40 nm CMOS		28 nm FD-SOI	65 nm CMOS
Domain	Time (back-gate-driven)		Time	Analog current	Analog charge		Analog charge	Analog current
Application mode	Accelerate	Sense	Accelerate	Sense	Accelerate	Sense	Accelerate	Accelerate
Number of parallel channels	16		128	48	1[a]		16	4096
Sampling rate	**1.35 GHz**	0.75 GHz	-	1.3 MHz	39 MHz	15 MHz	2.4 MHz	170 MHz
Speed (MACs/s)[b]	**21.6 G**	12 G	-	63 M	2.5 G	1 G	24 M	-
Supply (V)	**0.7**		1.0	1.2	1.1	1.0	1.0	1.2
Energy/MAC (pJ)	0.042[c]	**0.018[c]**	0.020[d]	0.051	0.13[e]	0.11[e]	0.07	0.027
Energy efficiency (TOPS/W)	9[f]	**55.3[c]**	48.2[d]	19.6	7.77[e]	8.77[e]	9.61[g]	36.8[h]
Resolution[i]	6b/3b/6b	Analog/3b/1b	1b/1b/1b	Analog/5b/1b	6b/3b/6b	Analog/3b/6b	8b/8b/8b	2b/1b/3b
Active area (mm^2)	0.04		3.61[d]	0.0206	0.012[e]		0.084	0.325

[a]Serial matrix-vector product
[b]Cumulative multiply-accumulate rate across parallel inputs
[c]averaged across weights excluding output buffer, [d]1-bit quantization and excludes external I/O and includes SRAM memory, [e]includes a 6b ADC and excludes output buffer
[f]includes 6-bit TDC and excludes output buffer, [g]includes 8-b ADC, [h]includes 3-b ADC
[i]Resolution Notation: Input/Weights/Output

tone (f_{NB}) of single ended output with the carrier (f_c) at 1.35 GHz measured by Agilent PXA Signal Analyzer. The harmonics of the fundamental tone are suppressed under -75dB. Fig 13(d) shows the measured output phase modulation in logarithm scale from minimum input to full scale amplitude (A_{FS}) and its corresponding output delay variation in linear scale is plotted in Fig. 13(e). Fig. 13(f) plots the measured average energy efficiency (units-TOPS/W) of about 55.3 TOPS/W across different weight settings.

C. System Demonstration: Handwriting Recognition

The BG-TMAC implements the matrix computations in a neural network for handwritten digit recognition. In this work, single layer neural network simulations using our BG-TMAC prototype are performed. We present a hardware chip solution, where the focus of this work is to demonstrate our idea of multi-bit time domain computing circuit approach. We have validated our chip over a single layered neural network. The simulated classification accuracy is 84.3%. This circuit could be extended to a multi-layered computational system. For instance, as the number of layers is increased to three, the classification accuracy of handwritten recognition can be increased to a value greater than 95% [16]. Perceptron learning algorithm is used from [36], and the training of digits and post-processing such as weight-update and gain-error correction are performed externally [27].

The choice of design parameters is made by maximizing energy efficiency while maintaining a required system quantization accuracy. The 6-bit quantization is implemented for approximate computing systems that can perform statistical inference at a precision of at least 4-bit accuracy [9]. The clock rate of 1.35 GHz is set by the maximum rating of available data timing generator (DTG 5274). The clock frequency can be tuned to a lower value for increasing energy efficiency and ENOB at reduced throughput. As the supply voltage scales below 0.65 V, the ENOB reduces below 4.5-bit compromising linearity performance with reduced energy consumption. In order to maintain ENOB of at least 4.5-bit, the supply voltage of 0.7V is chosen. In our design, sixteen parallel inputs are implemented due to the chip area restriction. The number of inputs can be increased to enhance the overall throughput.

Performance comparison with recent works (Table I) demonstrates that BG-TMAC shows the best energy efficiency with parallel high-throughput computing at sub-1V operation in standard CMOS technology. In sense-mode, the

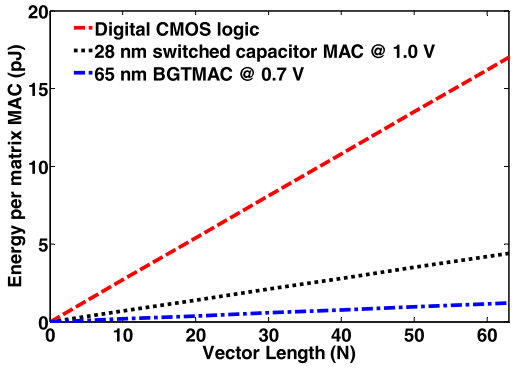


Fig. 14. Energy consumption with parallel vector length N.

16-parallel-input BG-TMAC interface consumes an average energy of 18 fJ/MAC computation, exhibiting an energy efficiency of 55.3 TOPS/W. In accelerate-mode operation by sampling at a high clock rate of 1.35 GHz, the 16-parallel-input BG-TMAC interface computes at the highest measured rate of 21.6 GMAC/s throughput while consuming an average energy consumption of 42 fJ/MAC computation and exhibits an energy efficiency of 9 TOPS/W including 6-bit TDC. By sampling at a reduced clock rate of 800 MHz in accelerate-mode, the average energy consumption reduces to 19.2 fJ/MAC that is about 4× lower compared to the switched capacitor MAC (SC-MAC) implementation [11]. From the energy estimates in [11] and [37], the resultant energy consumption per MAC computation for a parallel vector length-N using sub-1V BG-TMAC is about 15× lower compared to digital CMOS logic implementation designed by N multipliers and $N-1$ adders consuming 270 fJ/MAC, as shown in Fig. 14.

VII. Conclusion

A spatial multi-bit time-domain matrix multiplier interface is presented as a scalable alternative for multi-mode approximate computing applications across various signal processing interfaces. The architecture employs parallel back-gate-driven input vectors to perform matrix multiply-accumulate (MAC) computations in time-domain and operates under sub-1V supply voltage. The proposed system demonstrates parallel high-throughput and ultra-low computing energy for two application modes supporting accelerator and sensing front-end applications, respectively. In a digitally-driven accelerate-mode, the 16-parallel channel system demonstrates a high throughput of 21.6 GMAC/s and consumes 42 fJ/MAC of energy per computation. The proposed system functions directly as an analog interface, for classification in the time-domain for direct inference from sensors. In the analog front-end sense-mode, the interface exhibits a high energy efficiency of 55.3 TOPS/W which corresponds to a computation energy of 18 fJ/MAC at sub-1V supply voltage. An on-chip mismatch calibration mechanism for non-linearity correction is presented for spatial time-domain MAC architectures. The proposed sub-1V N-parallel-input architecture offers a 15× lower energy consumption per computation compared to an N-parallel-input digital static CMOS combinational logic implementation. The proposed spatial multi-bit time-domain architecture is an attractive and scalable solution for massively parallel approximate computing applications.

References

[1] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.
[2] F. Conti and L. Benini, "A ultra-low-energy convolution engine for fast brain-inspired vision in multicore clusters," in *Proc. Design, Automat. Test Eur. Conf. Exhib.*, 2015, pp. 683–688.
[3] V. Sze, Y. H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, 2018, pp. 1–8.
[4] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Proc. Deep Learn. Unsupervised Feature Learn. Workshop, NIPS*, 2011, pp. 1–8.
[5] D. Bankman and B. Murmann, "Passive charge redistribution digital-to-analogue multiplier," *Electron. Lett.*, vol. 51, no. 5, pp. 386–388, 2015.
[6] E. H. Lee and S. S. Wong, "A 2.5 GHz 7.7 TOPS/W switched-capacitor matrix multiplier with co-designed local memory in 40 nm," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 418–419.
[7] J. Zhang, Z. Wang, and N. Verma, "A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.
[8] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.
[9] E. H. Lee and S. S. Wong, "Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 261–271, Jan. 2017.
[10] S. Skrzyniarz *et al.*, "A 36.8 2 b-TOPS/W self-calibrating GPS accelerator implemented using analog calculation in 65 nm LP CMOS," in *IEEE ISSCC Dig. Tech. Papers*, Jan./Feb. 2016, pp. 420–422.
[11] D. Bankman and B. Murmann, "An 8-bit, 16 input, 3.2 pJ/op switched-capacitor dot product circuit in 28-nm FDSOI CMOS," in *Proc. IEEE Asian Conf. Solid-State Circuits (ASSCC)*, Nov. 2016, pp. 21–24.
[12] S. Joshi, C. Kim, S. Ha, and G. Cauwenberghs, "From algorithms to devices: Enabling machine learning through ultra-low-power VLSI mixed-signal array processing," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2017, pp. 1–8.
[13] *ITRS Technology Trend*. Accessed: Jan. 2016. [Online]. Available: http://www.itrs.net
[14] R. Mohan, S. Zaliasl, G. G. E. Gielen, C. Van Hoof, R. F. Yazicioglu, and N. Van Helleputte, "A 0.6-V, 0.015-mm^2, time-based ecg readout for ambulatory applications in 40-nm CMOS," in *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 298–308, Jan. 2017.
[15] D. Miyashita *et al.*, "An LDPC decoder with time-domain analog and digital mixed-signal processing," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 73–83, Jan. 2014.
[16] D. Miyashita, S. Kousai, T. Suzuki, and J. Deguchi, "A neuromorphic chip optimized for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing," *IEEE J. Solid-State Circuits*, vol. 52, no. 10, pp. 2679–2689, Oct. 2017.
[17] M. Liu, L. R. Everson, and C. H. Kim, "A scalable time-based integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition capabilities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr./May 2017, pp. 1–4.
[18] G. Li, Y. M. Tousi, A. Hassibi, and E. Afshari, "Delay-Line-Based Analog-to-Digital Converters," *IEEE Trans. Circuits Syst., II, Exp. Briefs*, vol. 56, no. 6, pp. 464–468, Jun. 2009.
[19] A. Ravi *et al.*, "A 2.4-GHz 20–40-MHz channel WLAN digital outphasing transmitter utilizing a delay-based wideband phase modulator in 32-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, pp. 3184–3196, Dec. 2012.
[20] R. B. Staszewski *et al.*, "All-digital TX frequency synthesizer and discrete-time receiver for Bluetooth radio in 130-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 12, pp. 2278–2291, Dec. 2004.
[21] P. Dudek, S. Szczepanski, and J. V. Hatfield, "A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line," *IEEE J. Solid-State Circuits*, vol. 35, no. 2, pp. 240–247, Feb. 2000.
[22] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, "A local passive time interpolation concept for variation-tolerant high-resolution time-to-digital conversion," *IEEE J. Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, Jul. 2008.

[23] J. Z. Ru, C. Palattella, P. Geraedts, E. Klumperink, and B. Nauta, "A high-linearity digital-to-time converter technique: Constant-slope charging," *IEEE J. Solid-State Circuits*, vol. 50, no. 6, pp. 1412–1423, Jun. 2015.
[24] C. Palattella, E. A. M. Klumperink, J. Z. Ru, and B. Nauta, "A sensitive method to measure the integral nonlinearity of a digital-to-time converter based on phase modulation," *IEEE Trans. Circuits Syst., II, Exp. Briefs*, vol. 62, no. 8, pp. 741–745, Aug. 2015.
[25] S. Sievert *et al.*, "A 2 GHz 244 fs-resolution 1.2 ps-peak-INL edge interpolator-based digital-to-time converter in 28 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 52–54.
[26] M. Maymandi-Nejad and M. Sachdev, "A monotonic digitally controlled delay element," *IEEE J. Solid-State Circuits*, vol. 40, no. 11, pp. 2212–2219, Nov. 2005.
[27] Z. Wang and N. Verma, "A low-energy machine-learning classifier based on clocked comparators for direct inference on analog sensors," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 11, pp. 2954–2965, Nov. 2017.
[28] Z. Wang, J. Zhang, and N. Verma, "Realizing low-energy classification systems by implementing matrix multiplication directly within an ADC," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 6, pp. 825–837, Dec. 2015.
[29] Y.-L. Lo, W.-B. Yang, T.-S. Chao, and K.-H. Cheng, "Designing an ultralow-voltage phase-locked loop using a bulk-driven technique," *IEEE Trans. Circuits Syst., II, Exp. Briefs*, vol. 56, no. 5, pp. 339–343, May 2009.
[30] I. Hayashi, T. Matsubara, S. Kumaki, A. H. Johari, H. Ishikuro, and T. Kuroda, "A phase-to-digital converter for wide tuning range and PVT tolerant ADPLL operating down to 0.3 V," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2010, pp. 1–4.
[31] T. Maeda and T. Tokairin, "Analytical expression of quantization noise in time-to-digital converter based on the Fourier series analysis," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1538–1548, Jul. 2010.
[32] K.-H. Cheng, Y.-C. Tsai, Y.-L. Lo, and J.-S. Huang, "A 0.5-V 0.4–2.24-GHz inductorless phase-locked loop in a system-on-chip," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 5, pp. 849–859, May 2011.
[33] A. S. Yousif and J. W. Haslett, "A fine resolution TDC architecture for next generation PET imaging," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 5, pp. 1574–1582, Oct. 2007.
[34] H. Chung, H. Ishikuro, and T. Kuroda, "A 10-bit 80-MS/s decision-select successive approximation TDC in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 5, pp. 1232–1241, May 2012.
[35] S.-K. Lee, Y.-H. Seo, H.-J. Park, and J.-Y. Sim, "A 1 GHz ADPLL with a 1.25 ps minimum-resolution sub-exponent TDC in 0.18 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2874–2881, Dec. 2010.
[36] *Stanford UFLDL Tutorial. Exercise: Supervised Neural Networks*, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, May 2011.
[37] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

Srinivasan Gopal (S'13) received the B.Tech. and M.Tech. degrees in electrical engineering from IIT Madras, Chennai, India, in 2008. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA. From 2008 to 2012, he held analog mixed-signal design engineering positions with Intersil and also with PMC-Sierra, and involved in precision data converter design and optical links.

During 2015 and 2016, he was an Analog Serial I/O Design Intern with Intel, San Jose, CA, USA, where he developed high-speed wireline transceiver designs for field-programmable gate arrays. His current research interests include wireless 3-D interconnect-based high-speed links and time-domain signal processing for neuromorphic computing.

Mr. Gopal was a recipient of the Intersil Summit Award for developing a comprehensive behavioral model with calibration for SAR analog-to-digital converter product line.

Pawan Agarwal (S'11) received the B.Tech. and M.Tech. degrees in electrical engineering from IIT Madras, Chennai, India, in 2009, and the Ph.D. degree from the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA, in 2017.

Since 2014, he has been a Research Intern with Maxlinear Inc., Carlsbad, CA, USA, where he involved in wireless systems for millimeter-wave backhaul applications.

Dr. Agarwal was a recipient of the IEEE MTT-S Graduate Fellowship Award and the MTT-S International Microwave Symposium Best Student Paper Competition Award.

Joe Baylon (S'12) received the B.Eng. degree from The Cooper Union, New York City, NY, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with Washington State University. His research focuses on low-power, wideband mm-wave wireless transceiver design for wireless network-on-chip. He is a Student Member of the IEEE Microwave Theory and Techniques Society.

Luke Renaud (S'12) received the B.S. degree (*summa cumme laude*) in electrical engineering from Washington State University, Pullman, WA, USA, in 2013, where he is currently pursuing the Ph.D. degree in RF microelectronics.

In 2015 and 2016, he was an Intern with EM Microelectronic-US, Colorado Springs, CO, USA, where he conducted research on low-power sensing methods for human interface devices. His current research works include the development of new envelope tracking systems and RF blocks for use in millimeter-wave systems.

Sheikh Nijam Ali (S'12) received the B.Sc. degree from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2009, and the M.Sc. degree from The University of British Columbia, Kelowna, BC, Canada, in 2012. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA.

In 2015, he joined the Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, as an Advanced RF Circuit Design Intern, where he was involved in wideband Doherty PA design. During 2018, he was an RFIC Design Intern with Skyworks Solutions Inc., San Jose, CA, USA, focusing on multi-band PA multi-chip-module design for next-generation cellular applications. His current research interests include spectral and energy-efficient RF and millimeter-wave integrated circuits and systems for 5G mobile communications. He received the 2018 Best Graduate Researcher Award from the School of Electrical Engineering and Computer Science, Washington State University.

Partha Pratim Pande (SM'11) received the M.S. degree in computer science from the National University of Singapore, Singapore, and the Ph.D. degree in electrical and computer engineering from The University of British Columbia, Vancouver, BC, Canada.

He is currently a Professor and the Holder of the Boeing Centennial Chair in computer engineering with the School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA. His current research interests include novel interconnect architectures for multicore chips, on-chip wireless communication networks, and hardware accelerators for biocomputing.

Dr. Pande was a recipient of the NSF CAREER Award in 2009. He received the Anjan Bose Outstanding Researcher Award from the College of Engineering, Washington State University, in 2013. He is the Editor-in-Chief (EIC) of the IEEE TRANSACTIONS ON MULTI-SCALE COMPUTING SYSTEMS and an Associate EIC of the *IEEE Design & Test*. He currently serves on the editorial boards of the *ACM Journal of Emerging Technologies in Computing Systems* and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, and on the Program Committees of many reputed international conferences.

Deukhyoun Heo (S'97–M'00–SM'13) received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2000. In 2000, he joined National Semiconductor Corporation, Santa Clara, CA, USA, as a Senior Design Engineer, involved in the development of silicon RFICs for cellular applications.

In 2003, he joined the Electrical Engineering and Computer Science Department, Washington State University, WA, USA, where he is currently the Frank Brands Analog Distinguished Professor of electrical engineering. He has authored or co-authored approximately 120 publications, including 55 peer-reviewed journal papers and 66 international conference papers. His current research interests include RF/microwave/opto transceiver design based on CMOS, SiGe BiCMOS, and GaAs technologies for wireless and wireline data communications, and intelligent power management system for sustainable energy sources, adaptive beam formers for phased-array communications, low-power high data-rate wireless links for biomedical applications, and multilayer module development for system-in-package solutions.

Dr. Heo is a Technical Program Committee Member of the IEEE Microwave Theory and Techniques Society (IEEE MTT-S), the IEEE MTT-S International Microwave Symposium (IMS), and the International Symposium of Circuit and Systems. He was a recipient of the 2000 Best Student Paper Award presented at the IEEE MTT-S IMS and the 2009 National Science Foundation CAREER Award. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II: EXPRESS BRIEFS during 2007–2009 and the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES.