A Time-Domain Computing Accelerated Image Recognition Processor with Efficient Time Encoding and Non-linear Logic Operation

Zhengyu Chen, *Student Member, IEEE,* and Jie Gu, *Member, IEEE*

***Abstract*—Time-domain computing (TC) has drawn significant attention recently due to its highly efficient computation for applications such as image processing and neural network computing. This paper presents novel time-domain circuit techniques including (1) double encoding strategy, (2) bit-scalable design which accelerates the performance compared with previous linear coding, and (3) shared time generator (TG) with variation-aware design technique which significantly improves the error tolerance of time-domain computing. A feature-extraction and vector-quantization processor accelerated by TC has been developed for real-time image recognition. A 55nm prototype chip shows 72 fps/core (@1.33 GHz) operation with up to 42% area and power saving from time-domain computing compared with conventional digital implementation.**

***Index Terms*—time-domain computing (TC), image recognition, bit-scalable design, double-encoding scheme, median filter, winner-take-all.**

I. INTRODUCTION

THE energy improvement of conventional digital circuits mainly relies on the scaling of technology and supply voltages (V_{dd}), as given in the energy equation: CV_{dd}^2 where C is the capacitance of the circuits. However, as the technology scaling is becoming harder, there is an urgent demand in finding alternative computing methods that bring efficiency beyond conventional digital approach. To improve the computing efficiency, several non-conventional digital techniques such as approximate computing and stochastic computing have been proposed providing a good tradeoff between energy/area consumption and accuracy. Such a tradeoff is made possible based on the fact that 80% of daily application has certain degree of error tolerance in the applications leading to feasibility of approximation in computing [2]. In addition, statistical computing approaches were proposed by exploiting the statistical significance of the computation and using error correction scheme to correct error from the most-significant bits due to voltage over-scaling [3, 4]. Despite of the different methodologies used in the above low power design techniques, the energy reduction is still based on conventional voltage and technology scaling leaving little room for further improvement assuming logic optimization has been well obtained from the modern design automation tools.

On the other hand, the analog computing which has been explored over decades, offers several attractive features such as high energy efficiency in certain applications such as analog multiplier, mixed-signal FIR filter [5, 6, 7, 8], etc. However, analog computing suffers from its limitation on voltage scalability due to the design of amplifier, process variation sensitivity, the static current from amplifier, and incompatibility to conventional digital design flow. As a result, analog computing has not been chosen as primary design method for general purpose computing compared with its digital counterpart.

Recently, the mixed-signal time-domain computing emerges as an interesting alternative to existing computing methods [9, 10, 11, 12]. The time-domain computing, due to its mixed-signal nature of the design, combines the advantages of both conventional digital and analog computing. On one hand, circuit-wise, the time-domain computing (TC) utilizes all digital components to encode/decode and process information in time-domain which brings the benefit of technology scalability and compatibility of the current digital design flow. On the other hand, from signal processing point of view, time-domain computing is similar to analog computing as the information can be more densely encoded in a single signal leading to benefits similar to the analog based processing, such as energy efficiency and a desirable error resiliency/scalability where most-significant-bit is least likely to show errors.

A. Previous Work

Several solid demonstrations have been provided in recent year in using time-domain computing for realizing popular signal processing jobs. For instance, a time-based low-density parity-check (LDPC) design has been demonstrated showing a 2X reduction in area compared with digital counterpart [9]. A time-domain neural network design has been shown with simple use of ring-oscillator and counters to realize integrate-and-fire neuron with significantly enhanced energy efficiency [10]. A high energy efficient time-based neuromorphic chip is proposed for deep leaning [11]. More recently, a time-domain reinforcement learning engine for robotic cart has been demonstrated showing 45% reduction in power compared with digital design [12].

However, there has been a lack of discussion on robustness and variation impact to the time-domain computing, which is the crucial consideration in this type of

design. As a result, most of existing work suffers from the following issues: (1) the existing design utilized a low efficient and variation vulnerable multiple-gate time encoding (TE) circuit limiting the advantages of the technique [9]; (2) existing design only contains low bits precision, e.g. 3 bits in [9, 10], and does not address the process variation impact to the design; (3) The strong capability of TC in various nonlinear operations, e.g. MAX, MIN operation, has not been well explored leaving limited improvement from the techniques [10, 11, 12]; (4) Variation impact which is critical to the time-domain computing design is not well considered and analyzed [10, 11, 12]. The stringent timing control and matching requirement of the technology poses challenges to the adoption of the techniques into a large-scale design. Thus, there still lacks a comprehensive and efficient design methodology of time-domain computing.

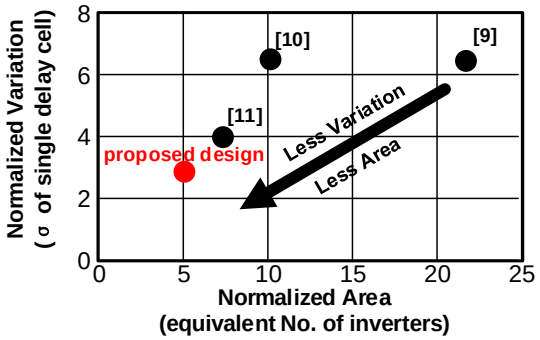


Fig. 1. Efficiency vs. variation of timing encoding circuits from prior work [9, 10, 11] and our proposed work.

As time encoding or digital-to-time conversion is the first step and the most critical job in the time-domain computing, the selection of time encoding determines the robustness of the time. Unfortunately, none of the prior work clearly address that issue. For instance, the design of LDPC in [9] and the design of time-domain neural network [10] utilize the very power and area consuming circuits for time encoding. In addition, the variation or mismatch from previous work increase dramatically with the number of bits encoded which make the designs error vulnerable. Although the more recent demonstrations on neural network based design [10, 11] exhibit higher degree of tolerance to the errors, it is still important to develop a robust time encoding technology for designs that tolerate less of the error. This work targets to address such issues with a series of circuit techniques that bring robustness and efficiency to the time-domain computing. Fig. 1 shows the efficiency and variation comparison of time encoder between prior work and our work [9, 10 ,11]. The area is defined as the area cost of equivalent number of inverters and variation is normalized to the standard deviation (σ) of single inverter. More detailed discussion will be provided in the later sections.

B. *Contributions of This Work*

As extended from previous works in [13, 25], this work develops comprehensive techniques of time-domain computing including shared time generator, double-edge operation, and bit-scalable design. Also, a comprehensive modeling of variation along with variation-driven design is provided. More specifically, the contributions of this paper are highlighted as below:

1) Efficient and robust time encoding circuit techniques are proposed including (a) shared time generator design with 4X variation reduction, (b) double-edge operation strategy which improves the energy efficiency by 2X, and (c) bit-scalable design which accelerates the design by 16X and also improves variation tolerance of time-domain computing.
2) We exploit the efficient non-linear operation of time-based computing, e.g. MAX, MIN, CMP, to realize complex functions like median filter and winner-take-all. Compared with digital counterparts, the TC design achieves 20% ~40% improvement in area, energy and performance.
3) Models for variation impacts to time-domain computing design are provided and a variation-driven design methodology is developed for an optimized trade-off between energy and robustness.
4) A complete implementation of proposed methodology on an image recognition test chip. We show that the proposed design achieves state-of-art energy efficiency and throughput compared with conventional design with similar algorithms.

The organization of this paper is given below. In Section II, a brief introduction of the principle and building blocks of TC is described. An overview of the proposed design methodology is introduced. In Section III, a variation-awareness design methodology is proposed to relieve the variation concern which is critical in time-domain computing. In section IV, complex TC non-linear operations including winner-take-all and median filter are introduced along with algorithm optimization and circuit implementation. A comprehensive TC design flow and a test chip implementation on image recognition are provided in section V. Measurement that supports the results of the proposed design flow is presented in Section VI. Section VII concludes the paper discussion.

II. Time Domain Computing

A. *Basic Circuit Building-blocks*

Time-domain computing (TC), or also referred as time-domain signal processing (TDSP) converts the task of signal processing into “time” or delay of digital cells which can be processed efficiently for numerous operation. The digital binary inputs are first encoded and processed in time domain and either reconverted back into digital domain through time

decoder or results are directly obtained at time domain without time decoder. Essentially, the digital information is represented by the delay or pulse width of the standard modules. Fig. 2 (a) shows an overall setup of the TC which consists of key building blocks. Here, T_{in} is the time-domain input signal, D_{in} is the digital input signal which used to determine the time delay of generated time-domain signal from TE, and T_d is referred as the time delay of a single delay cell, e.g. buffer.

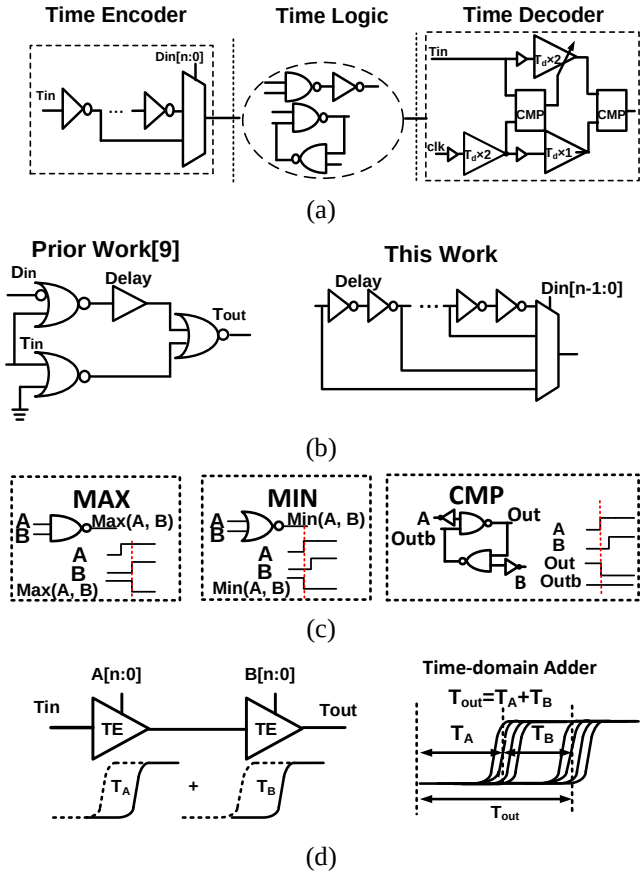


Fig. 2. (a) System overview of time-domain computing; (b) Time encoder circuit: prior work [9] (left), proposed TE in this work (right); (c) Time-domain logic circuits; (d)An example of n-bit TC adder: circuit schematic (left), waveform (right).

1) Time Encoder or Digital-to-Time Converter (DTC)

Fig. 2 (b) shows 1-bit TE from prior work (left) [9] and the proposed n-bit TE (right) in this work. Our proposed TE has a simple inverter chain to generate the delay passed through selection multiplexer to convert from digital input to time-domain signal [13]. Compared with the design in LDPC work [9], our design has benefit in area, energy and robustness. In order to decrease the variation impact to TC design, a more practical design based on this TE is proposed in Section III-C.

2) Energy and Area Efficient Time Logic

Fig. 2 (c) summarizes the schematic of basic time-domain logic cells that process the signal in time-domain once the information is encoded into time domain from TE. Compared with conventional CMOS logic design, many operations can be performed much more efficiently. For example, the MAX, MIN and Compare (CMP) operations are realized by a single or two logic gates leading to tremendous saving from conventional CMOS operation. Note that, the symmetric output load constraint is critical to the CMP circuit.

3) Time Decoder

Time decoder (TD), or time-digital-converter (TDC) has been extensively developed in all-digital phase-locked-loop (ADPLL) design with the state-of-art TDC achieving 1ps resolution [14, 15]. Right-hand side of Fig. 2 (a) shows a 2-bit base-line TD design based on binary-search. However, due to the high energy and area cost of such a TD, in this work, we develop algorithms that eliminate the use of TD leading to much improved area and energy efficiency.

The benefits of TC come from the following interesting facts: (1) similar as analog computing, multiple bits can be encoded into single transition leading more efficient information delivery. An example is the adder circuit as shown in Fig. 2 (d) where the n-bit operation only consumes transitions of a few inverters rendering 3X improvement in energy efficiency; Here, A[n:0] and B[n:0] are the digital control signal of the TEs; T_A and T_B are the corresponding time-domain signals. (2) Some logic operation can be efficiently carried, which is discussed in Section II-A-2); (3) Owing to the analog nature of operation, the TC is intrinsically immune to large magnitude of error. In other word, compared to digital design, the analog-based design has much less chance to have the error happen at the most-significant bit (MSB). Since the computation output is more affected by MSB errors in applications like facial recognition, TC is more error tolerant compared with digital counterpart when error occurs; (4) Although the information is processed in time-domain, the information carriers are still binary digital signals processed by conventional logic circuits which makes the entire design digital-friendly.

B. Double-edge Operation

In this work, we propose a double-edge operation (also referred as double encoding in this work), where logic operation is processed at both rising and falling transitions as shown in Fig. 3 (a). The energy taken from source to charge the gate capacitance is $E = V * Q = V * V * C = CV^2$. Half of the energy is dissipated during rising transition; the other half is dissipated during falling transition. In previous design [9], only single transition is used. Thus, the other half is purely a waste. In our proposed design, we utilize both transitions, which provides us with 2X energy efficiency. Area consumption is also reduced by around 30% because the buffer stage is shared for both rising and falling transitions. Fig. 3 (d) shows the energy and area saving come from the double-encoding scheme. Fig. 3 (b) and (c) shows the logic encoding concepts between conventional

complementary logic design where pull-up and pull-down realize complementary logic functions and the dual-encoding strategy for TC. Interestingly, for TC design, for rising and falling operation, the design could perform two totally different logic operations as compared with conventional design where complementary operations have to be performed. As shown in Fig 3 (c), (1) during the falling edge of time-domain input signal T_{in}, the pull-up part of the circuit is turned on, which processes $D_{up} = A + B$; (2) During the rising edge of T_{in}, the pull-down part of the circuit is turned on, which processes $D_{dn} = \overline{CD} + E$. This means that theoretically TC could realize more functionality with the same amount of pull-up and pull-down logic circuits as compared with conventional digital design. Simulation shows that the pull-up and pull-down operation can be completed decoupled without delay impact to each other as long as the input slew rate is much faster than the encoded 1-bit delay, which is guaranteed by adding inter-stage buffers.

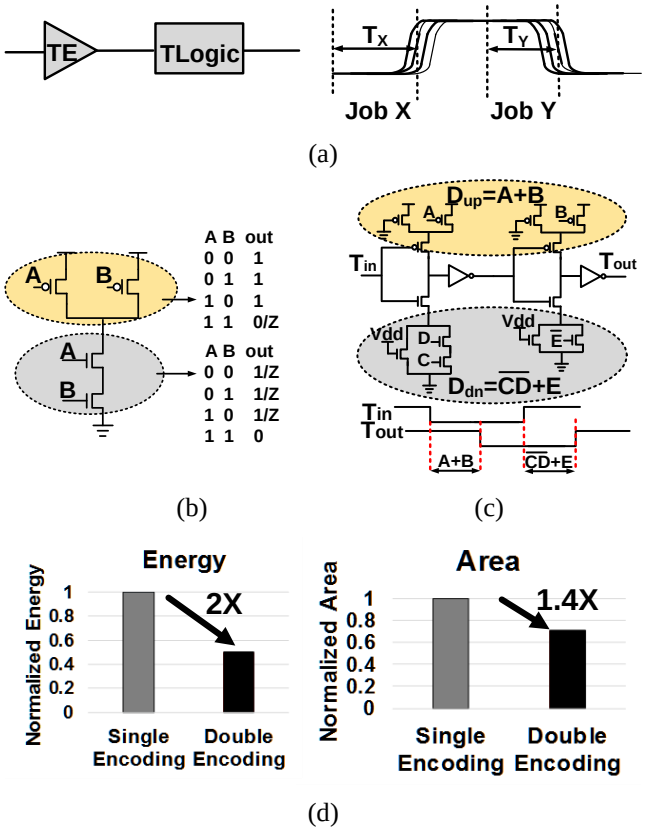


Fig. 3. Proposed double-edge operation: (a) overview, (b) conventional digital complementary logic, (c) Two 1-bit adders using dual-encoding scheme with different logic operations for rising and falling transitions, (d) energy and area comparison between single-encoding and double-encoding schemes.

C. Bit-Split Scheme

As previous work shows only limited bit precision, e.g. 3 bits [9, 10], we propose a bit-split technique that splits an input vector into smaller bit groups leading to a scalable high-resolution encoding without exponentially increasing the delay. As shown in Fig. 4 (a), 8-bit inputs A and B are split into 2 sub-groups, e.g. A[7:4] (referred as MSBs) and A[3:0] (referred as LSBs). In this work we encode 4-bit MSBs operation in the falling-edge and 4-bit LSBs in the rising-edge rending 16X reduction of delay and 2-4X reduction of energy on a 8-bit non-linear operation, e.g. MIN(MAX(A, B), (C+D)) (Fig. 4 (b)). This technique also makes TC designs scalable with the number of bits since large number of bits can be split into several small groups. To allow the split of the bits, a digital combination logic has to be added to combine the decision from each sub-group. This incurs digital equal comparison to deal with the situation that MSBs are equal. An example of the combine algorithm is shown in Fig. 4 (c): the comparison between A and B goes through the following steps: (1)check if A's 4-bit LSBs equals to B's, if no, go to step (2), else go to step (3); (2) Check weather A's 4-bit MSBs is larger than B's; (3) Check weather A's 4-bit LSBs is larger than B's. Then based on the value of these conditions to determine the mathematical relation between A and B.

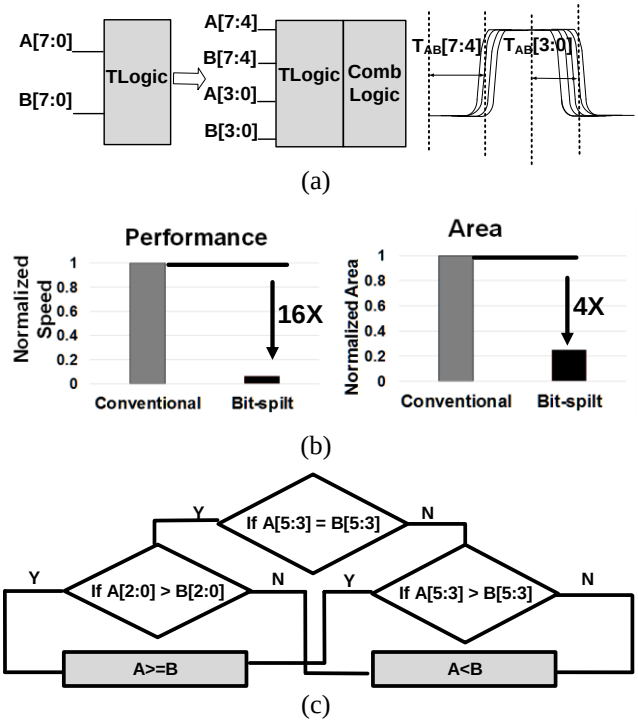


Fig. 4. Proposed bit-split technique: (a) technique overview, (b) performance and energy saving come from the bit-split scheme, (c) combine algorithm.

D. Design Challenges in Time-domain Computing

As all the information is carried and processed in time-domain, the timing control is critical to guarantee the error resiliency due to the sensitivity of delay to variations including both global variation and local mismatches. Since the least-significant-bit (LSB) resolution is pre-defined, e.g. 25ps used in this work, a mismatch of timing beyond this value will lead to single-bit error. Note that, global variation does not hold significant concern as circuit's delay equally

scales at global variation, as will be introduced in Section III-B. However, local variation or mismatch creates the largest threat to the operation. An example is the compare (CMP) operation where the two signals are compared for timing. The two signals are required to be carefully matched so that no systematic mismatch leads to a variation beyond 1 LSB. This is on top of local random mismatch which requires upsizing to reduce the mismatch, which is similar to the analog design. In summary, TC suffers from sensitivity to process variation and matching requirement despite of its advantages in digital friendly circuit design. To deal with this issue, we propose a variation-driven design method discussed in Section III.

III. Variation-Aware Design Methodology

A. Design Overview

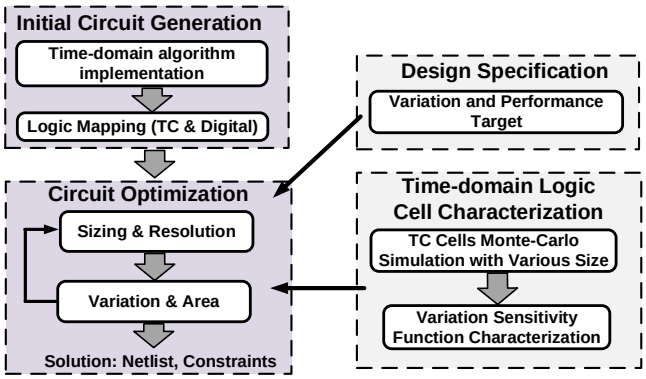


Fig. 5. Variation-awareness design flow of TC.

Fig. 5 shows the overall flow of our proposed variation-awareness design methodology. The flow includes two steps: (1) initial circuit generation which first convert the target algorithm/application into time-domain algorithm and then map the time-domain algorithm into time-domain logic cells; (2) Resizing the time-domain logic cells and adjusting the single-bit delay (resolution) to meet the design specification, e.g. variation and performance target. More specifically, (a) the capacitive loads of the time-domain circuits determine the delay of the circuits, e.g. time encoder; (b) The larger size of the transistor, the smaller variation it has but the capacitive load/energy consumption also increases. Thus, the size of time-domain logic cells must be carefully selected to satisfy the design specification. Meanwhile the time-domain logic cells are characterized based on Monte-Carlo simulation.

B. Global Variation VS. Local Variation

As time encoder (TE) holds the most stringent requirement on the timing control accuracy, we performed the global process-voltage-temperature (PVT) variation analysis to a 4-bit time encoder in a 55nm technology using the base-line TE as shown in Section II-A. Since relative delay among signals is most critical to keep the computation error-free or at low error rate, linearity matters the most compared with the absolute delay values. As shown in the left side of Fig. 6 (a), (b), and (c), the linearity is well preserved across PVT corners. The right side of Fig. 6 (a), (b), and (c) show the integral nonlinearity (INL) of the time encoder where the integrated nonlinearity is represented as a fraction of least-significant bit (LSB). The INL variation of the time encoder is well maintained within 15% of LSB across PVT corners. Hence, through a proper budgeting of variation, the global PVT variation does not introduce significant concern to the functionality of the TC design.

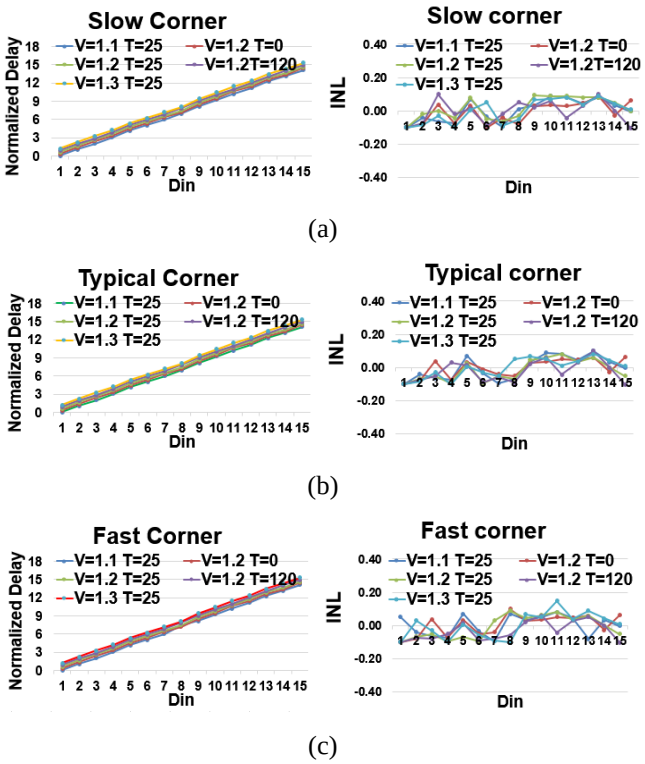


Fig. 6. Normalized delay across PVT corners and INL of (a) slow corner, (b) typical corner, (c) fast corner.

On the other hand, local variation/mismatch poses more challenges to the TC design compared with global variation due to the linearity requirement of time encoder. Monte-Carlos simulation with random threshold voltage variation is performed to evaluate the impact of local mismatch. In order to precisely model the conversion factor from threshold voltage variation to the delay of time encoder, a detailed variation-awareness design methodology will be introduced in Section III-D.

C. Shared Time Encoder Design

The time encoder poses the most stringent variation requirement on the timing control accuracy as the generated signal experiences multiple stages of variation impact during signal generation stage. For example, for a 4-bit TE, the longest delay generated by such a TE sums up the variation of 15-stage inverter chain. Assume the variation (also referred as the standard deviation) of a single stage inverter is σ_{inv}, the variation V_{ind-TE} for a n-bit individual TE (Fig. 7 (a)) in the worst scenario is shown in (1):

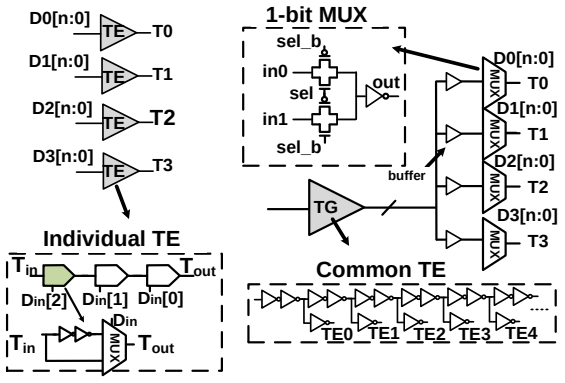
$$V_{ind-TE} = \sqrt{(2^n-1)\sigma_{inv}{}^2 + n\sigma_{inv}{}^2}$$
$$= \sigma_{inv}\sqrt{(2^n-1)+n} \quad (1)$$

The term, $(2^n-1)\sigma_{inv}{}^2$,represents the variation comes from (2^n-1)-stage inverter chain when input is set to maximum value. The other term, $n\sigma_{inv}{}^2$, comes from the n-stage multiplexer at the end of the inverter chain. As been discussed in Section III-B, relative delay among signals is most critical, we derive the mismatch between two individual TEs in (2):

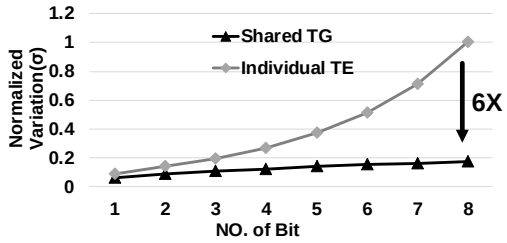
$$V_{ind-TE-CMP} = \sigma_{inv}\sqrt{2(2^n-1)+2n} \quad (2)$$

The mismatch between two individual TEs increases dramatically when the number of bit increases. In order to relieve the variation concern comes from the time encoder, we proposed the shared time generator (TG) scheme which uses a common inverter chain to generate timing signals relieving the variation impact. Fig. 7 (b) shows the circuit schematic of the shared time generator which consists of (1) a common inverter chain used to generate the all possible delay, (2) distributing multiplexers which are used to select the desired delay for each output. When we consider the mismatch of two different output from TG, the worst scenario is when the delay difference of two output is just single-bit delay away from each other. Since all the output are generated by the common inverter, the variation comes only from the distributing multiplexers eliminating the mismatch source from inverter chains rendering much improved variation resiliency. The mismatch between two outputs from TG can be represented by (3):

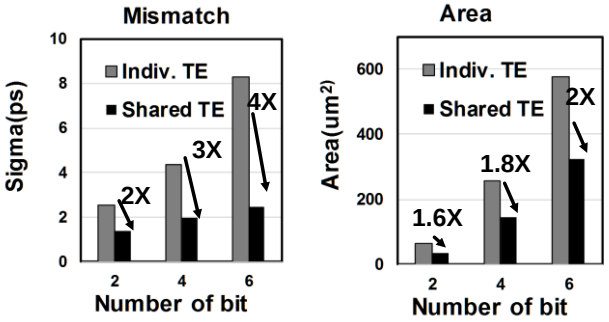
$$V_{TG-CMP} = \sigma_{inv}\sqrt{2n} \quad (3)$$



(a) (b)



(c)



(d)

Fig. 7. (a) Circuit schematic of individual TE scheme [9], (b) proposed shared time encoder scheme, (c) variation comparison between shared TG and individual TE, (d) mismatch and energy comparison between induvial TE and shared TE (8-output).

Compared with individual TE design whose mismatch in a n-bit TE is proportional to $\sqrt{(2^n-1)+n}$, the shared TG has mismatch only proportional to $\sqrt{n}$ leading to 3~4X less mismatch rendering shortened single bit delay and smaller cell size. Fig. 7 (c) shows the variation trend of shared TG and individual TE cases. As we can see, the variation of individual TE case grows dramatically compared with the shared TG case. Besides, the area consumption of TG is also smaller due to the sharing of common inverter chain. Fig. 7 (d) shows the mismatch improvement and area saving come from proposed common TG design. Note that, the more paths/operations that share the same TG, the more complexity of distributing network will be needed, e.g. tree-structure buffers and distributing MUXs. In our case, it is a clear win for using shared TG. But it is possible that the shared PG becomes too expensive if too many signals are generated. In that case, the design needs to be performed towards using individual TE.

D. Variation-Awareness Design

In this section, we introduce the variation-awareness design flow. The goal is to meet the variation budget of the critical paths while minimizing the area/power consumption of the design.

We define the 3-sigma variation of TC modules, which is a function of the size s as $\sigma(s)$. Apparently, the $\sigma(s)$ decreased as s increases. Also, the area is a function of the size s as $A(s)$. The sensitivity of variation of a module can be defined as (1):

$$F_{sen}(s) = \gamma\frac{d\sigma(s)}{dA(s)} \quad (1)$$

Where $\frac{d\sigma(s)}{dA(s)}$ term represents the variation sensitivity comes from module itself without considering the whole placement topology. The γ term represents the significance of the module, e.g. module in a convergent path.

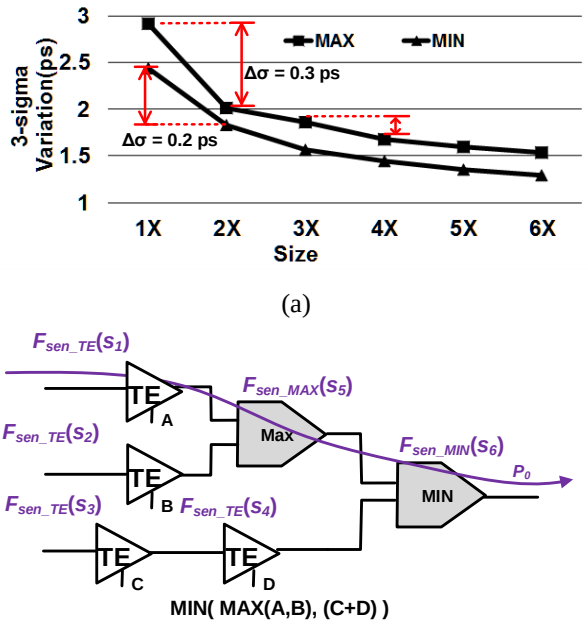


Fig. 8. (a) Variation of MAX (NAND2) and MIN (NOR2) at various sizes; (b) Example schematic.

Fig.8 (a) shows the variation-area(size) curve of MAX (NAND) and MIN (NOR) gates. Since most time-domain cells are standard-cell like, we follow standard cell sizing convention of 1X, 2X, 3X, etc. An example of TC operation, MIN(MAX(A, B), (C+D)), is shown in Fig. 8 (b). Given a simple task of decreasing the variation of critical path P_0, it is obvious that we can gain more variation improvement if we give the sizing priority to the module whose current variation sensitivity is larger than the rest. Based on this, we propose our netlist optimization in following.

Assume we have totally n modules, X_1, X_2, …X_n, the size of each module is s_1, s_2, … s_n. Besides, there are P paths need to be considered in the placement. The optimization problem of netlist is then formed in (4) and (5):

$$\text{Minimize } \sum_{i=1}^{n} A_i(s_i) \tag{4}$$

$$\forall\, paths \in P, s.t. \sqrt{\sum_{i=1}^{n} {\sigma_{pi}}^2(s_i)} \leq \sigma_T \tag{5}$$

where $\sigma_i(s_i)$ is the variation of X_i, and $A_i(s_i)$ is the module area of X_i. The optimization algorithm is described as follows:

Given the initial schematic/netlist with minimum size for each module, we first check if the variation of critical path meets the budget σ_T. If yes, the optimization is completed. Otherwise, the second step is performed where we traverse the netlist to find out the module in the critical path with maximum variation sensitivity with their current size. The size of module is then increased by 1X. In the following steps, we continue to check whether the variation budget is satisfied. If not, the optimization repeats by upsizing the most effective module, i.e. highest sensitivity. The pseudo code is shown in below.

Algorithm 1 Variation-Awareness Optimization Algorithm

Input: Initial schematic/netlist of module $X_1, X_2, \ldots X_n$, with minimum sizing $s_1, s_2, \ldots s_n$.

Output: Netlist which satisfies variation budget with minimum estimated area

while $\forall\, paths \in P, \sqrt{\sum_{i=1}^{n} {\sigma_i}^2(s_i)} > \sigma_T$ **do**
  for i = 1 to n **do**
    find the module $j = i$,with maximum $F_{sen\_j}(s_j)$
  end
  Increase the size of module j by 1X, update s_j
end
Return the schematic/netlist with current sizing

IV. Exploiting Complex TC Non-linear Operations

Many image processing applications such as pattern classification and facial recognition, require a large amount of non-linear signal processing operations, e.g. comparison, sorting, minimum, maximum, etc. [16, 17]. Among them, winner-take-all (WTA) and median filter (MF) are two of the most critical building blocks commonly used for pattern classification. In WTA and MF, a deterministic decision is made based on excessive compare and sorting operation which are highly expensive to be implemented in standard CMOS ASIC design and even more difficult for a CPU operation [18, 19]. In this section, we introduce the time-domain WTA and MF algorithm as well as their circuit implementation. As introduced in Section II-A, many nonlinear operations can be efficiently implemented in TC. Fig. 9 (a) shows the circuit diagram of TC implementation for operation MIN(MAX(A, B), (C+D)), which was introduced in previous section. As we can see, by using several NAND and NOR gates, such a complex operation can be easily implemented in TC rendering an energy saving of 6X. Therefore, we can improve the area efficiency by increasing the utilization of such efficient TC non-linear functions, e.g. MAX, CMP, in the design.

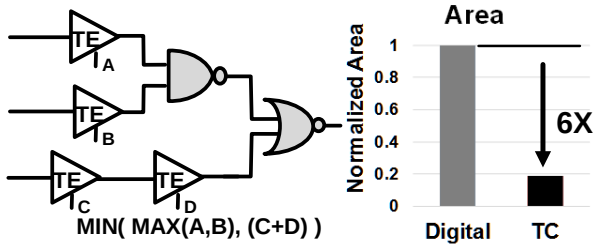


Fig. 9 (a) TC implementation of a non-linear operation; (b) Area comparison between digital and TC approaches.

A. *Winner-Take-All*

1) *Time-domain Based Algorithm*

We derived our time-domain WTA algorithm from a binary comparison tree scheme which takes advantages of the efficient MAN/MIN and CMP TC operations (Fig. 10).

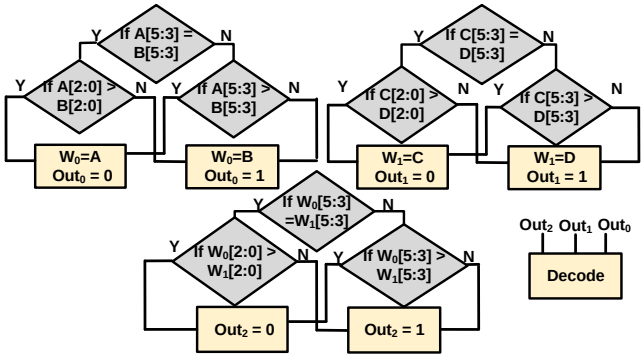
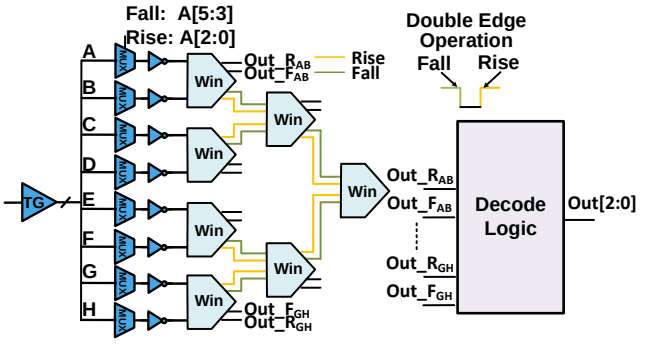


Fig. 10. Algorithm of 6-bit 4-input WTA design with 3-bit split in MSBs and LSBs.

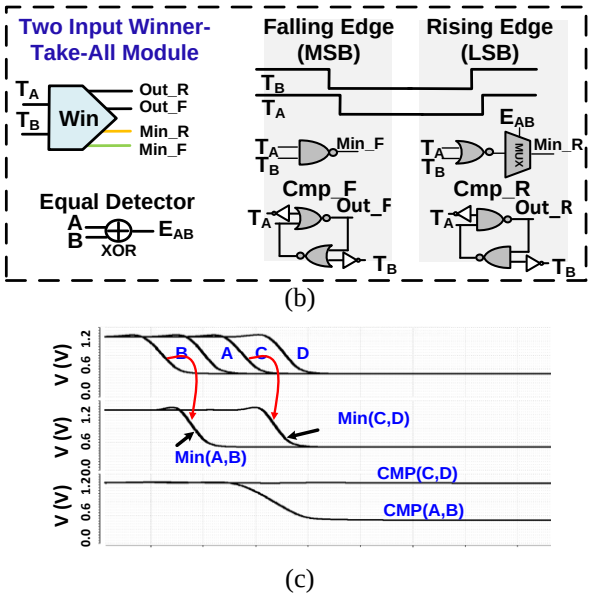
2) Circuit implementation

Fig.11 (a) shows the circuit implementation of the proposed 8-input 6-bit winner-take-all accelerator. The winners or MIN value from each branch in each stage are selected in parallel and propagated to the subsequent stage to be compared again. After converting the input digital value into time-domain, the comparison can be simply made by using time-domain CMP. The MIN function which is built by a single NAND/NOR gate directly propagates the winner to next stage without intermediate restoration or regeneration. As a result, a massive parallel operation with mostly NAND or NOR gates, is realized in TC. All comparison results are finally decoded in digital domain to find the final winner. Shared TE, double-edge operation and bit-split technique are also utilized in this design.

Fig. 11 (b) shows the example of double-edge operation and bit-split techniques utilized in the WTA design: (1) the 6-bit input are divided into two groups as in[5:3] and in[2:0]; (2) in[5:3] (referred as MSBs) and in[2:0] (referred as LSBs) are encoded into falling and rising edge of the same clock cycle respectively; (3) During falling transition of signal, MSBs are processed while LSBs are processed during rising transition. Fig.11 (c) shows the example waveform of MIN function used in this design. As a result, the TC WTA achieves lower area consumption with faster speed. The area of proposed time-based WTA is improved by 42% compare to digital implementation as shown in Fig. 12.



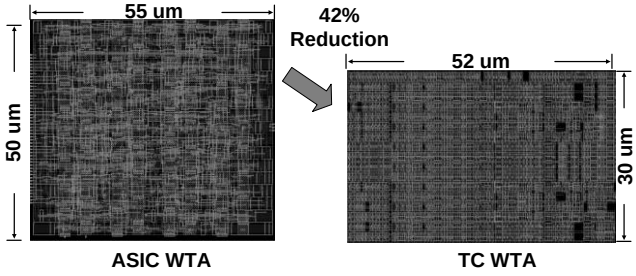
(a)



(b)

(c)

Fig. 11. (a) Circuit design of WTA, (b) example of double-edge and bit-split technique, (c) example waveform of MIN function used in this design.



(a) (b)

Fig. 12. Layout comparison between (a) conventional digital WTA and (b) TC WTA.

B. Median Filter

1) Time-domain Based Algorithm

As a core building block in applications such as facial recognition, median filter (MF) consumes up to 70% of total CPU cycles due to its enormous amount of CMP and swapping operation in a typical bubble sorting algorithm [20]. In order to remove the bottlenecks of the application, several efficient MF algorithms have been proposed such as [21, 22]. The majority voting algorithm [22] improves energy efficient, but still suffers from the drawbacks of analog-based design such as cannot scale with technology which requires substantial amount of effort in tuning the circuit and designing the layout.

We propose an energy efficient time-based MF with high performance. The core idea of the algorithm is to have a massive comparison between each of the two inputs and order the inputs from high to low. The final median value is filtered/selected by the proposed decoder. Fig.13 shows the algorithm and detailed implementation of the proposed 12-input 8-bit time-domain median filter design as following

steps: (1) each pair of the 12 inputs is compared, thus a total number of 66 comparisons are processed. The comparison result is recorded as "0" or "1", e.g. if A > B, OUT_{AB} = 1, OUT_{BA} = 0; (2) The related comparison results of each input are summed up, e.g. $OUT_A = OUT_{AB} + OUT_{AC+...} + OUT_{AL}$; (3) Finally, all the summation results are compared with N/2, where N is the number of inputs. The input whose summation result of the comparisons equals to N/2 is marked as the median value.

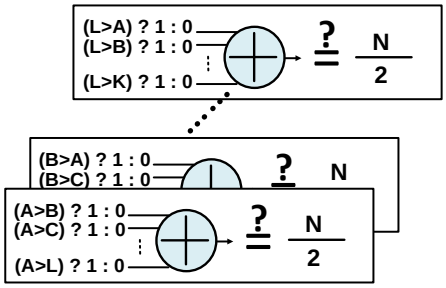


Fig. 13. Algorithm of 12-input MF.

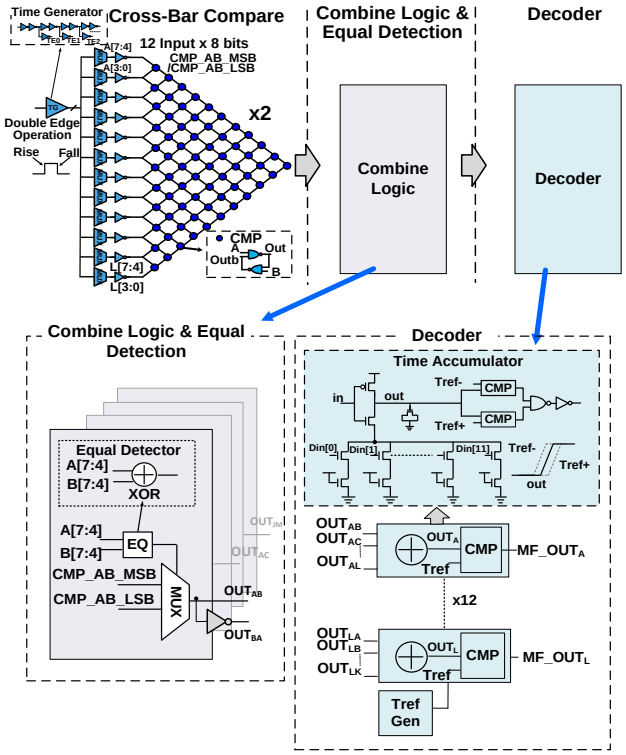


Fig. 14. Circuit diagram of MF with detailed circuit schematic of combine logic & equal detection and the circuit detail of decoder.

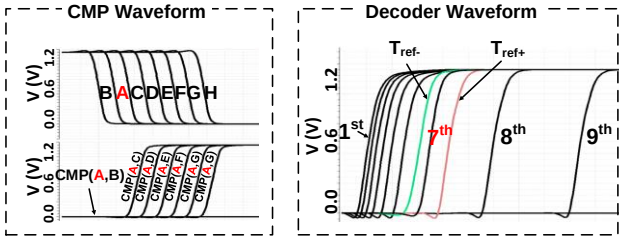


Fig. 15. Example of waveform of MF operation.

Figure. 14 shows the corresponding TD circuit diagram of MF. During the comparison stage, the digital inputs are first converted into time-domain by the proposed shared TEs. Each pair of input is compared in parallel in time-domain with double-edge and bit-split design, with overall 66 comparisons for both MSBs and LSBs for all 12 input vectors. The 66 comparisons are processed parallelly in the cross-bar compare module as shown in the left-hand side of Fig. 14. During the combine & equal-detection stage, all the comparison results are then processed by the following digital logic for purpose of equal detection and MSBs/LSBs grouping to obtain the comparison results for all input vectors. In the decoder stage, to decode the comparison results into final median value, a special time accumulator/adder design is implemented where all 11 digital comparison results from each input are summed in time-domain and compared with a reference median time-domain signal.

2) Combine Logic and Time-accumulator Based Decoder

Although the 66 comparison results can be further decoded into final result in conventional digital design, the decoding logic will incur large overhead due to the complex operations. To reduce the overhead, we proposed a time-accumulator based time-domain decoding logic. The bottom-left side of Fig. 14 shows the detailed circuit implementation of the combine & equal detection stage. The bottom-right side of Fig. 14 shows the detailed circuit implementation of decoder. The core idea of the time-domain decoder is to form a detection window by T_{ref-} and T_{ref+}. As the 12 inputs are ordered and represented by the delay of the rising edge, the median value is carried by the 7^{th} rising edge. The T_{ref-} is set to be located in middle of 6^{th} and 7^{th} signal while T_{ref+} is set to be located in middle of 7^{th} and 8^{th} signal. In this way, the 7^{th} signal which represents the median value can be captured by the detection window as shown in the decoder waveform in Fig. 15. Compared with digital decoder design, the time-based decoder dramatically reduced the area by 3X.

Overall, the final area of proposed time-based MF is improved by 24% compared to conventional digital implementation as shown in Fig. 16.

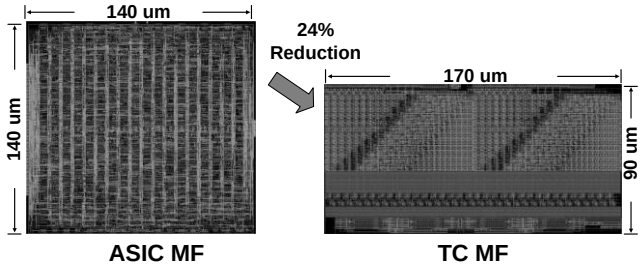


Fig. 16. Layout comparison between (a) conventional digital MF and (b) TC MF.

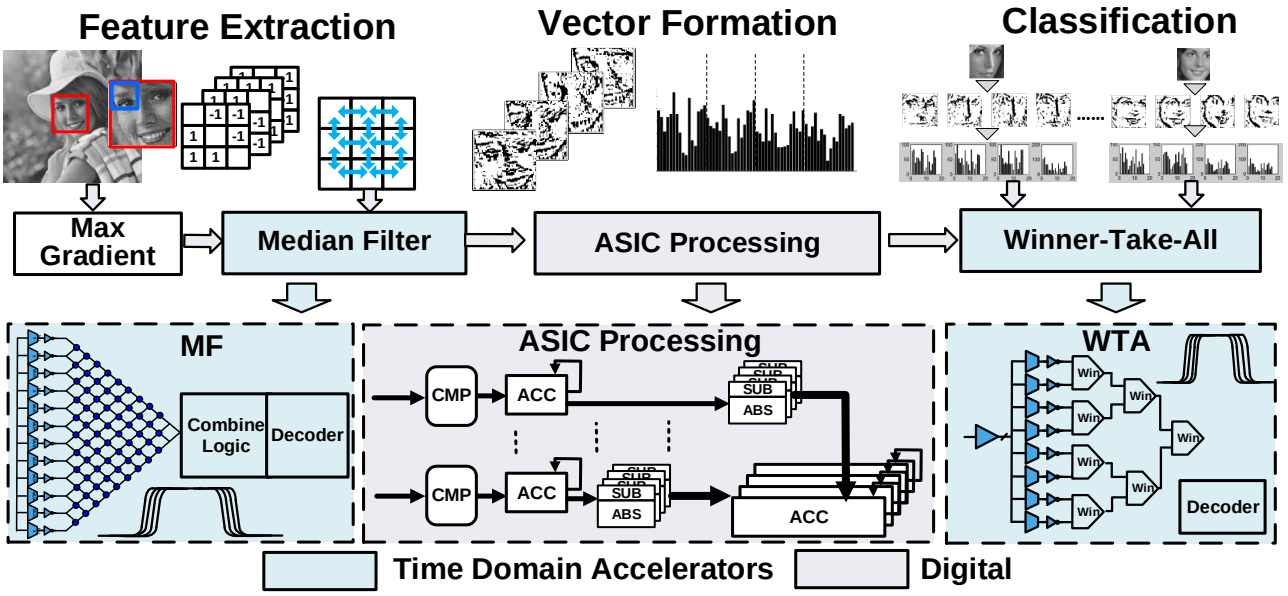


Fig. 17. Overview of image recognition algorithm used in this work.

V. TIME-DOMAIN COMPUTING ACCELERATED IMAGE RECOGNITION PROCESSOR

To demonstrate the proposed circuit techniques, we adopt a basic image recognition algorithm as shown in Fig. 17 into a hybrid ASIC design with time-domain accelerators [22].

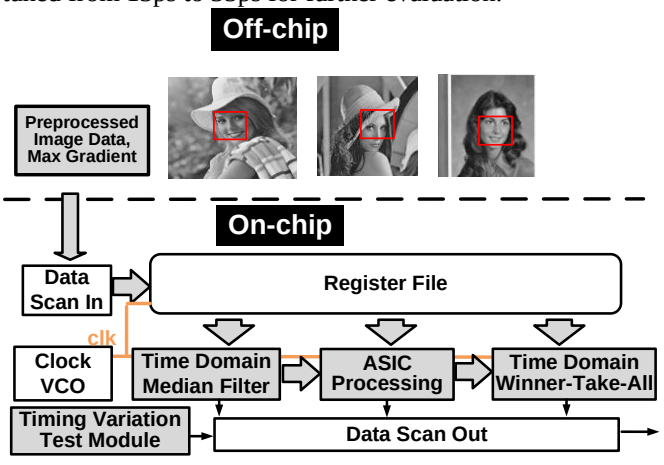
A. Implemented Image Processing Algorithm

As shown in Fig. 17, the operations of the image recognition algorithm involve three main steps: (1) feature extraction which detects edges in four directions: horizontal, vertical, +45˚ and -45˚. In order to determine the threshold value for edge detection, all the absolute-value differences between each two neighboring pixels are calculated in the 3×3 kernel and the median detection of the 12 difference-value is adopted as the threshold; (2) Vector formation where edge flags in all directions are counted and the spatial distribution of edge flags is represented by a vector of 64 elements; (3) Classification: the generated feature vector is then classified by a winner-take-all (WTA) classifier.

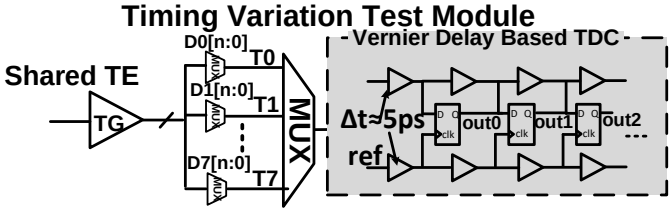
The subtractors and absolute value circuits shown in Fig. 17 are used for calculating the distance between template feature vector and input feature vectors, e.g. $D_{xy} = |x - y|$. The compare (CMP) and 1st stage accumulator (ACC) compute the 64 elements. Then the subtractor and absolute value circuits calculate the distance of each element between the template feature vector and input feature vectors. At the end, the 2nd stage accumulator calculate the accumulated distance of the 64 elements between template feature vector and each input feature vectors. The heavily used nonlinear computations such as comparison (CMP), MIN/MAX function, are expensive for CPU/GPU based design or even state-of-art ASIC design. In this work, TC based accelerators are used to remove the bottlenecks of the algorithm, i.e. MF and WTA with significant speedup as shown in Section IV.

B. Test Chip Implementation

Fig.18 (a) shows the test chip implementation of the proposed image recognition processor in a 55nm low power CMOS process at 1.2V. Scan chains are used to fetch image data to on-chip register files and read out all internal register/comparator values for test verification. A special timing test module (Fig. 18 (b)) is built to exam the linearity and robustness of the proposed shared TE design. A Vernier-delay-chain based TDC with ~5ps bit-resolution is used to characterize the timing variation of TE. The TE used throughout this work is implemented with a ~25ps single-bit resolution which can be tuned from 13ps to 35ps for further evaluation.



(a)

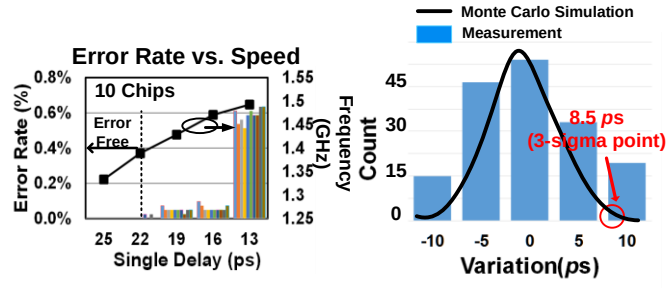


(b)

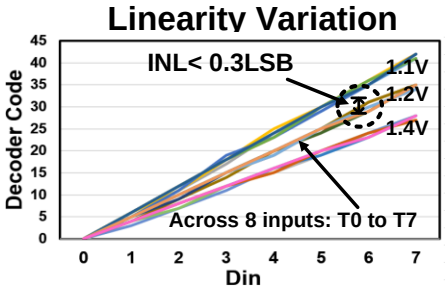
Fig. 18. (a) Top level implementation of the proposed test chip; (b) Circuit diagram of timing variation test module.

VI. MEASUREMENT

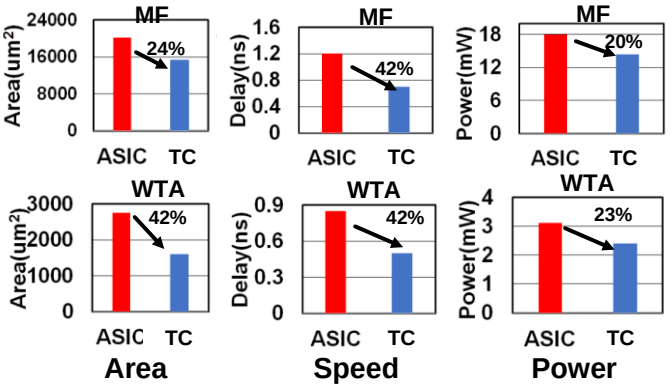
In the test chip setup, there is a separate power supply for TE modules to change the single-bit delay from 13ps to 35ps. Note that, we cannot directly measure the single-bit delay on the test chip due to the limited measurement resolution. However, based on extracted simulation from SPICE, we can estimate the single-bit delay on the test chip of the current supply voltage. Fig. 19 shows measurement results. Robustness of the design was verified across 10 chips. As shown in Fig.19 (a), by default, no error was observed at the design target speed of 1.33GHz. When pushing the TE resolution beyond 22ps (estimated based on simulation), small error was observed at the MF's output at LSBs while no error was observed at the final WTA output. The error rate from MF reached 0.6% when reducing the resolution to 13ps which led to an operating speed of 1.5GHz, a 13% boost of performance without observing error at the final output. This shows the strength of TC where small errors may be generated at LSBs at stringent timing condition but does not lead to significant error at final output.



(a) (b)



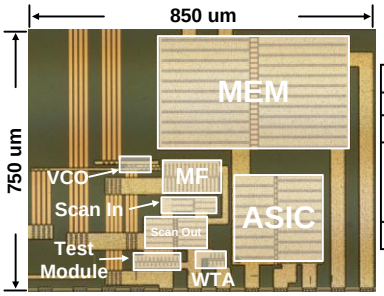
(c)



(d)

Fig. 19. Measurement results on (a)performance, (b) measured (blue histogram) vs. simulated variation through chips, (c) linearity of the TE, (d) area, speed and power compared with ASIC.

The linearity of TE was also measured for all eight inputs across 10 chips and supply voltages from 1.1V to 1.4V as shown in Fig. 19 (c). Only small deviation (~8ps) from ideal value was observed across all the measurement leading to an integral-nonlinearity (INL) of less than 0.3 LSB. The measured (blue histogram) vs. the simulated variation from SPICE Monte Carlo simulation of TE across chips are shown in Fig. 19 (b). The results match the expectation from the simulation. Note that, the measurement results in the figure are from all chips and all paths with Din set as 7. As the matching among paths on the same chip is more critical, the measured variation on a single chip is quite small, which is within 5ps.



Process		Low Power 55 nm
Area		0.64 mm²
Frequency		1.33 GHz
Power	ASIC	59 mW @1.33 GHz
	MF	14 mW @1.33 GHz
	WTA	2 mW @1.33 GHz
Throughput		72 Frames/s

Fig. 20. Die micrograph and specifications.

The design is compared with conventional ASIC design in the same process with standard synthesis and place & route implementation. As shown in Fig. 19 (d), 24% to 42% area saving is observed in MF and WTA accelerators compared with ASIC implementation. A 1.7X speedup and 20% to 23% power saving are also observed using TC. The overall image recognition processor operates at 1.33GHz with a throughput of 72 frames per second (fps). Fig. 20 shows the die micrograph and the detailed design specifications. As the focus of this work is on robust and efficient techniques for time-based design, a direct comparison with prior work is difficult. We made comparison in two aspects:

1) Time-based Work

As shown in Table I and Fig. 1, compared with prior time-based work [9, 10, 11], (1) we achieved the fastest operation speed with a single-bit delay which is 2X~4X shorter; (2) We encoded largest number of bits by the bit-split technique; (3) We achieved lowest mismatch/variation which is 3X smaller compared with [9, 10, 11]; (4) We had the least encoding effort with lowest transistor count.

2) Image Recognition Processors

As shown in Table I and Fig. 21, compared with image recognition processors with similar algorithms, e.g. feature vector based, we achieved (1) the highest throughput per core and throughput per area, (2) highest energy efficiency for single processor core with more than 9X improvement. However, it is notable that prior work involves more configurations and numbers of processing units [22, 23, 24]. Note that, (1) compared with our implementation, only [22] implements very similar design. We have significant advantages due to both time-domain design as well as technology scaling from 180nm to 65nm. (2) For designs in [23, 24], their algorithms are much more complex and support more throughput. For example, the work presented in [23] is based on vector parallel image

recognition algorithm. Work presented in [24] is based on principal component analysis (PCA), and the proposed hardware utilizes the technique described above to reduce data dimensionality and uses support vector machine (SVM) as a final classifier for face recognition. Our comparison is focused on "single core/PE".

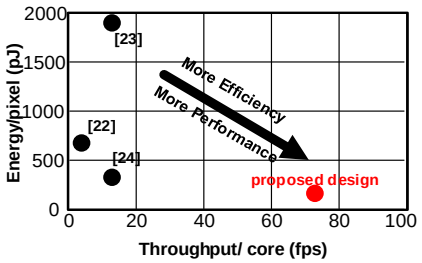


Fig. 21. Performance vs. energy for prior image processing designs [22, 23, 11].

TABLE I
PERFORMANCE COMPARISON

		[22] JSSC 2007	[23] JSSC 2014	[24] JSSC 2017	[9] JSSC 2014	[10] CICC 2017	[11] ASSCC 2016	**This work**
Image Recognition	Technology	180nm	180nm	40nm	65nm	65nm	55nm	55nm
	Voltage (V)	1.8	1.8	0.6	1.2	1.2	1.2	1.2
	Area (mm^2)	33.64	82.3	5.9	0.063	0.24	3.61	0.64
	Power (mW)	85	630	23	4	0.3	-	75
	Accuracy (No. of bit)	8	8	10	-	-	-	8
	Frequency (GHz)	0.1	0.05	0.1	-	0.1	-	1.33
	Throughput/core (fps) *	6.1	16	14.7	-	-	-	72
	Throughput/area (fps/mm^2)	11.6	12.4	80.3	-	-	-	116
	Energy/pixel (pJ)	756	2126	84	-	-	-	54
Time-based	Single bit delay (ps)	-	-	-	100	50	-	25
	Maximum No. of bit encoded	-	-	-	3	3	5	8
	Timing Mismatch**				6.5	6.5	4	2.8
	No. of equiv. inverters to generate 1-bit delay (4-bit TE)	-	-	-	22	10	7	5

* Based on 256×256 image with 64×64 scan window.

** Based on 4-bit time-encoding. Normalized to sigma (σ) of single delay cell.

VII. Conclusion

This paper proposed a series of highly efficient time-domain computing techniques including: shared time generator, double-edge operation scheme, bit-split technique and high efficient time-domain non-linear operations. In our approach, the use of TC-based accelerates the pipeline operation bottleneck by 40% due to the limitation of MF and WTA operations. The strength of TC including error resiliency, highly efficient non-linear operations and better energy/area efficiency compared with digital counterpart is demonstrated by a test chip. The test chip on image recognition processor is fabricated in 55-nm low power CMOS showing state-of-art energy efficiency and throughput with significant improvement from time-domain techniques compared with conventional digital implementation.

References

[1] Yong Shim, et al, "Low-Power Approximate Convolution Computing Unit with Domain-Wall Motion Based "Spin-Memristor" for Image Processing Applications", *IEEE DAC*, 2016

[2] Soheil Hashemi, et al, "Approximate Computing for Biometric Security Systems: A Case Study on Iris Scanning", *IEEE DATE*, 2018.

[3] Naresh R. Shanbhag, et al., "Stochastic Computation", *IEEE DAC*, pp. 859-864, 2010.

[4] Rami A. Abdallah, et al., "An Energy-Efficient ECG Processor in 45-nm CMOS Using Statistical Error Compensation", *IEEE JSSC*, vol. 48, no. 11, pp. 2882-2893, Nov. 2013.

[5] P. Godoy, et al., "Chopper Stabilization of Analog Multipliers, Variable Gain Amplifiers, and Mixers", *IEEE JSSC*, vol. 43, no. 10, pp. 2311-2321, 2008.

[6] Stephen T. Kim, et al., "Subthreshold Current Mode Matrix Determinant Computation for Analog Signal Processing", *IEEE ISCAS*, 2010.

[7] Karim Abdelhalim, et al, "915-MHz FSK/OOK Wireless Neural Recording SoC With 64 Mixed-Signal FIR Filters," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2478-2493, 2013.

[8] M. Gu, et al., "A 100 pJ/bit (32, 8) CMOS Analog Low-Density Parity-Check Decoder Based on Margin Propagation," *IEEE JSSC*, vol. 46, no. 6, 2011.

[9] Daisuke Miyashita, Ryo Yamaki, Kazunori Hashiyoshi, Hiroyuki Kobayashi, Shouhei Kousai, Yukihito Oowaki, Yasuo Unekawa," An LDPC Decoder With Time-Domain Analog and Digital Mixed-Signal Processing", *IEEE JSSC*, vol. 49, no. 1, pp. 73-83, 2014.

[10] Muqing Liu, Luke R. Everson, Chris H. Kim, " A Scalable Time-based Integrate-and-Fire Neuromorphic Core with Brain-Inspired Leak and Local Lateral Inhibition Capabilities", *IEEE CICC*, 2017.

[11] Daisuke Miyashita, Shouhei Kousai, Tomoya Suzuki, Jun Deguchi, Toshiba Corporation, Kawasaki, Japan, "Time-Domain Neural Network: A 48.5 TSOp/s/W Neuromorphic Chip Optimized for Deep Learning and CMOS Technology", *IEEE ASSCC*, 2016.

[12] Anvesha Amravati, Saad Bin Nasir, Sivaram Thangadurai, Insik Yoon, Arijit Raychowdhury," A 55nm Time-domain Mixed-signal Neuromorphic Accelerator with Stochastic Synapses and Embedded Reinforcement Learning for Autonomous Micro-Robots", *IEEE ISSCC*, 2018.

[13] Zhengyu Chen, Jie Gu, "Analysis and Design of Energy Efficient Time Domain Signal Processing", *IEEE ISLPED*, 2016.

[14] Young-Hun Seo, et al., "A 1.25ps Resolution 8b Cyclic TDC in 0.13µm CMOS", *IEEE JSSC*, vol. 47, no. 3, 2012.

[15] Stephan Henzler, et al., "A Local Passive Time Interpolation Concept for Variation-Tolerant High-Resolution Time-to-Digital Conversion", *IEEE JSSC*, vol. 43, no. 7, pp. 1666-1676, 2014.

[16] Elisabetta Chicca, et al., "Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems," *Proceedings of the IEEE*, vol. 102, no. 9, pp. 1367-1388, 2014.

[17] Hung, Y.-C. & Liu, B.-D. (2004), "A high-reliability programmable CMOS WTA/LTA circuit of O(N) complexity using a single comparator", *IEE Proc.—Circuits Devices and Syst.*, vol. 151, Dec. 2004, pp. 579-586.

[18] Shih-Chii Liu and Matthias Oster, "Feature Competition in a Spike-Based Winner-Take-All VLSI Network", *IEEE ISCAS*, 2006.

[19] Yuguang Fang, et al., "Dynamic Analysis of a General Class of Winner-Take-All Competitive Neural Networks", *Transactions on Neural Networks*, vol. 21, no. 5, pp. 771-783, 2010.

[20] G.M. Blair, "Low Cost Sorting Circuit for VLSI", *IEEE Transactions on Circuits and Systems*, 1996.

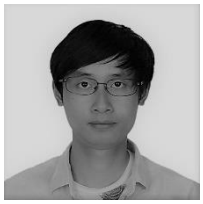
[21] Shiqiang Chen, et al, "sWMF: Separable Weighted Median Filter for Efficient Large-Disparity Stereo Matching", *IEEE ISCAS*, 2017.

[22] Hideo Yamasaki, Tadashi Shibata, "A Real-Time Image-Feature-Extraction and Vector-Generation VLSI Employing Arrayed-Shift-Register Architecture", *IEEE JSSC*, vol. 42, no. 9, pp. 2046-2053, 2007.

[23] Cong Shi, Jie Yang, et al, " A 1000 fps Vision Chip Based on a Dynamically Reconfigurable Hybrid Architecture Comprising a PE Array Processor and Self-Organizing Map Neural Network", *IEEE JSSC*, VOL. 49, NO. 9, 2014.

[24] Dongsuk Jeon, et al, "A 23-mW Face Recognition Processor with Mostly-Read 5T Memory in 40-nm CMOS" *IEEE JSSC*, 2017.

[25] Zhengyu Chen, Jie Gu, "An Image Recognition Processor with Time-domain Accelerators using Efficient Time Encoding and Non-linear Logic Operation", *IEEE ASSCC*, 2018.



Zhengyu Chen (S'16) received the B.S. degree in electrical engineering from Southeast University, Nanjing, China, in 2013, the M.S. degree in computer engineering from the Cornell University, Ithaca, NY, in 2015, and is now a Ph.D. candidate in computer engineering from Northwestern University, Chicago, IL. He is an aspiring researcher doing research in the area of ultra-low power design/algorithm for VLSI, mixed-signal ICs and emerging device. Now he is focusing on the low power algorithm design like time-domain signal processing and accelerator design of machine learning algorithms.

Jie Gu (M'10) received the B.S. degree from Tsinghua University, China, the M.S. degree from Texas A&M University and the Ph.D. degree from the University of Minnesota. He worked as an IC design engineer in Texas Instruments from 2008 to 2010 focusing on ultra-low-voltage mobile processor design and integrated power management techniques. He was a senior staff engineer in Maxlinear, Inc from 2011 to 2014 focusing on low power mixed-signal broadband SoC design. He is currently an assistant professor in Northwestern University. He has served as program committees and conference co-chairs for numerous low power design conference and journals, such as ISPLED, DAC, ICCAD, ICCD, etc. His research interests include ultra-low power mixed-signal VLSI circuit design, integrated power and clock management with hardware and software co-design, and emerging device/technology integration.