IEEE Asian Solid-State Circuits Conference
November 14-16, 2011 / Jeju, Korea

An Asynchronous Mixed-mode Neuro-Fuzzy Controller for Energy Efficient Machine Intelligence SoC

Jinwook Oh, Gyeonghoon Kim, and Hoi-Jun Yoo
Division of Electrical Engineering, School of EE,
KAIST, Daejeon, Republic of Korea
jinwook@eeinfo.kaist.ac.kr

Abstract— This paper presents an asynchronous digital-analog mixed-mode neuro-fuzzy controller that enables energy efficient implementation of machine intelligence SoC. The proposed neuro-fuzzy controller adopts an asynchronous 2-stage pipeline for analog and digital domain operations, which is the main contributor to high throughput and energy efficient machine intelligence SoC. To this end, a neuro-fuzzy controller with a delay prediction unit and a new ISA is introduced to modify the stage depth of mixed-mode pipeline incorporating with highly parallel special function units. Compared to the conventional digital standalone VLSI architecture, power reduction and processing speed acceleration are achieved by 46.6% and 71.4%, respectively, reporting 12.5μJ/epoch energy efficiency.

I. INTRODUCTION

Machine intelligence (MI) recently has been explored to increase system performance in various applications, such as autonomous vehicle control [1], human and computer interface, and computer chromatography [2]. Since they take advantage of MI's inference and learning ability that the next result is predicted from past data history, the accurate prediction is possible under insufficient present criteria or noisy circumstances. Even though well-known MI algorithms have $O(n^2)$ computational complexity with non-linear operations, however, once the learning process is finished, the number of computation can be reduced to be suitable for real-time on-line applications. Most of the previous studies for MI implementation focused on the development of ASICs [3-4] or CPU/DSP simulations [5]. In this work, however, a general programmable processor is proposed for achieving high performance neuro-fuzzy operations as a MI SoC. Especially, with the help of digital-analog mixed mode architecture, it adopts a 2-stage asynchronous pipeline to enhance the energy efficiency of the proposed processor.

Neuro-fuzzy algorithm, or an integrated inference system of neural network and fuzzy logic, is one of the most popular MI due to its high functionality and scalability. Most often, dedicated ASIC is used to implement one specific target system of neural network and fuzzy logic. In case, balancing between the programmability for system generality and the execution time for low energy operation is challenging for limited power and speed [4]. To this end, we proposed the analog-digital mixed-mode neuro-fuzzy accelerator achieving a programmable high performance MI SoC [6] .

As shown in Fig. 1, we present a programmable controller to enable high performance neruo-fuzzy accelerator by the asynchronous analog-digital mixed-mode pipeline in this work. The task delay prediction is performed by the delay prediction unit (DPU) and configuration special function register (CSFR) incorporating with a new neuro-fuzzy ISA. Using the information decoded from instruction in CSFR, the uneven execution time of the analog inference unit (AIU) and digital learning unit (DLU) can be asynchronously optimized respectively by the delay prediction.

II. SYSTEM ARCHITECTURE

A. Inference/Learning Algorithm

The general neuro-fuzzy algorithm is composed of inference and learning as shown in Fig. 2 [7]. Inference is a process to classify observations by comparison with pre-defined fuzzy rules to provide the confidence of input *X*. The

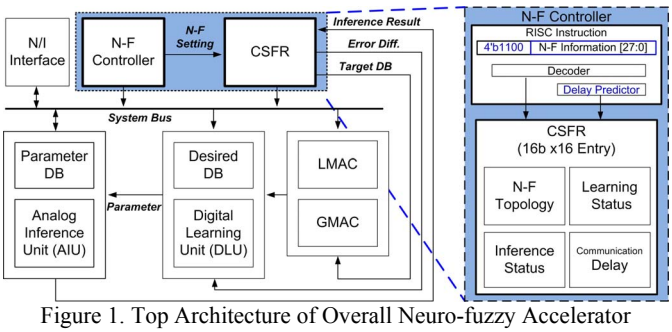


Figure 1. Top Architecture of Overall Neuro-fuzzy Accelerator

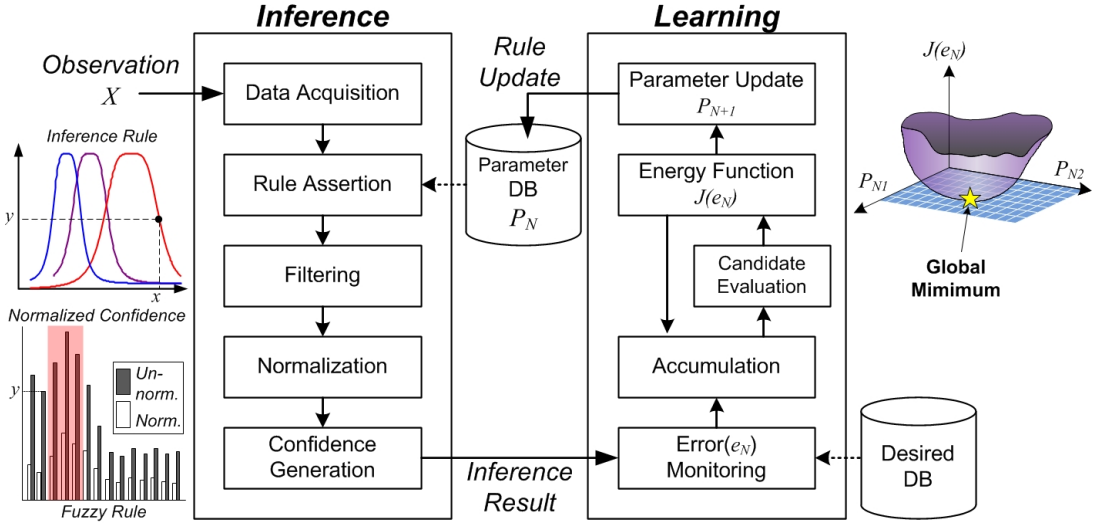


Figure 2. Flow Diagram of Inference/Learning Operation for Machine Intelligence Implementation

predicted result is normalized to attain the immunity to offset noise. Then the confidence is used to update new rule by learning process as inference result. The rule is mostly defined by a multiplication (2) of Gaussian function (1) .

$$f_i(x) = 1/\{1+(x-\frac{c_i}{a_i})^{2b_i}\} \quad for\ \ i=1,2,\cdots,N \qquad (1)$$

$$f_i(x,y,z) = w_i = f_A(x) f_B(y) f_C(z),\ for\ i=1,2,\cdots,N \qquad (2)$$

The a_i, b_i and c_i are the parameters of ith Gaussian function among N fuzzy rule components. One fuzzy rule $f_i(x,y,z)$ is defined for 3 input observations, x, y and z with by the multiplication of each membership function. Otherwise, the learning is a process to update new rule by using the difference between the inference result and desired value in target output y_d. The new rule parameters $w^{n+1}{}_i$ are updated by the energy function $J(w^n{}_i)$ which is the accumulated result of monitored errors $E(w^n{}_i)$ as shown in (3) and (4).

$$w^{n+1}{}_i = w^n{}_i - \eta\{E(w^n{}_i+\delta) - E(w^n{}_i)\}/\delta \qquad (3)$$

$$E(w_i^n) = (y(x_t, w^n{}_i) - y_d)^2 \qquad (4)$$

where η is the decreasing rate and δ is the update strength. Since each task is rooted on computationally complex algorithm, it is hard to achieve real-time operations even in recent CPUs while consuming large amount of energy its process [4]. This has been a barrier for applying large dimension of MI to SoC, restricting extensive use on mobile environment.

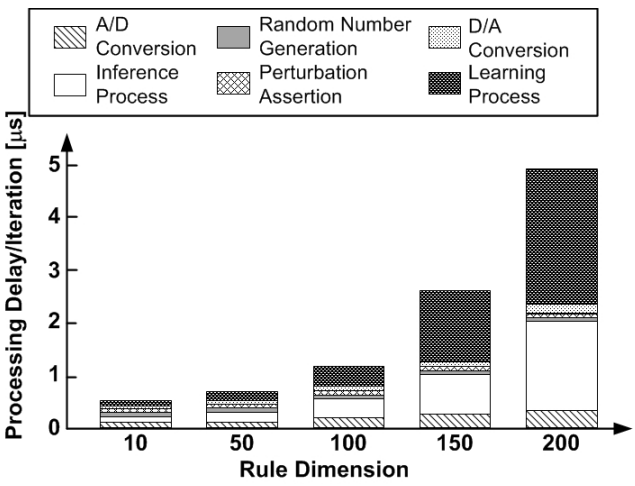
B. Analog-Digital Mixed-mode Architecture

Fig. 3 (a) shows the processing delay of each component of the proposed system architecture of Fig 1. Each stage is affected by the number of rules, especially, the inference operation in AIU and DLU are dominantly extended, thereby reducing overall system performance ~5X between 200 and 10 rules. In terms of power consumption, the case with 200 rules almost consumes 10X powers with 23% variability on measured power, caused by PVT variation on analog circuit arrays as well as temperature noise on current sources, as shown in Fig. 3(b). Increased number of rules is required for the prediction from complex observation patterns, escalating total RC delay of activated membership function circuits and multipliers with respect to n^2, where the n is the number of activated membership function circuits. In addition, the increased latency also can be varied by the relationships of observations and rules, accumulated at the process noise. As a result, the uncertainty of analog processing power and latency soars in accordance with rule's complexity.

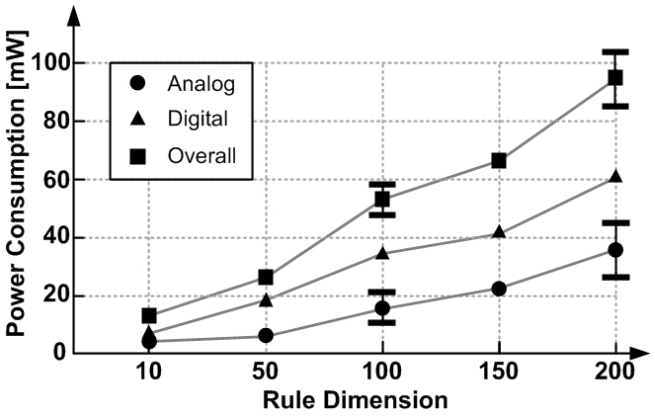
III. Asynchronous Analog-Digital Pipeline

A. Asynchronous Mixed-mode Pipeline

Fig. 4 shows the proposed asynchronous 2-stage mixed-mode pipeline between AIU and DLU. Fig 4 (a) describes 2-stage AIU-DLU pipeline with the fixed duty stage clock, *pclk*, for different two tasks. Since inference and learning processes of successive two iterations incorporate simultaneously, the unbalanced two different operations can induce throughput penalty with *pclk* asserted for worst case processing delay. Therefore, each assigned stage depth contiguously provides



(a)



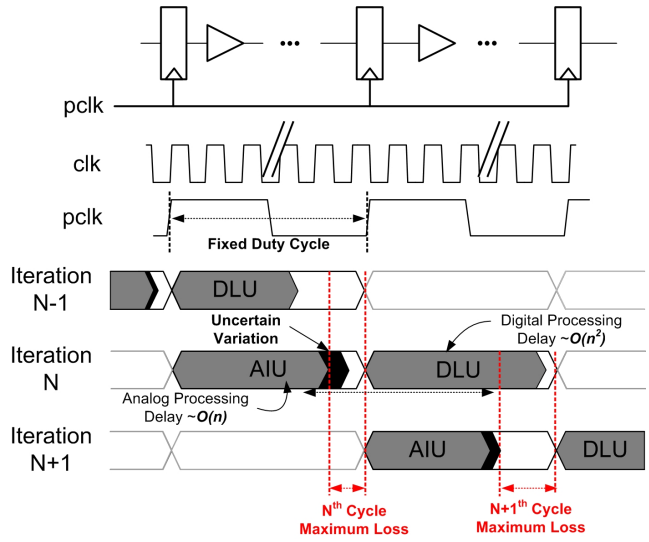
(b)

Figure 3. (a) Processing delay distribution of mixed-mode system (b) Power consumption of analog and digital block

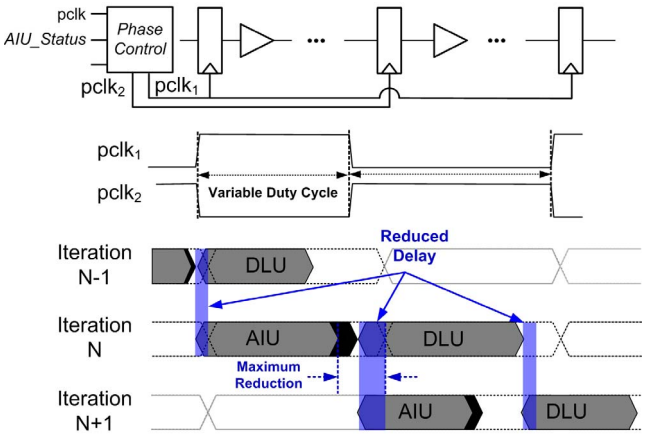
timing loss at every *pclk* cycle. In order to eliminate the stage losses, we proposed asynchronous mixed-mode pipeline by introducing new clock $pclk_1$ and $pclk_2$ which are based on the decoded neuro-fuzzy instructions from the neuro-fuzzy controller and CSFR. To this end, the *phase controller* generates two exclusive phase clocks which have variable duty cycle calculated from AIU status and delay predicted counts, *AIU_status* and *E_delay_cnt*, respectively. In this case, each stage depth is adaptively set to the longest processing delay of two contiguously tasks, achieving processing delay saving in each stage at most 24.4%. Using this scheme for the proposed neuro-fuzzy processor, on average, 14% throughput increase can be achieved for the target algorithm with 100 neuro-fuzzy rules. Furthermore, the proposed pipeline with delay prediction for AIU takes advantage of minimizing uncertain variation effect of processing delay and power increase. If the prediction of AIU operation precisely reflects its physical delay model of analog arrays, the operation power and delay drift can be reduced by about 23%.

B. *Neuro-fuzzy Controller*

In Fig. 5, the proposed neruo-fuzzy controller architecture is shown with 5-stage pipeline. When the AIU and DLU



(a)



(b)

Figure 4. (a) 2-stage pipeline with fixed duty cycle clock (b) Asynchronous mixed-mode pipeline

stages are activated, then execution stage works as 2-stage multi-cycle execution operation. The delay is calculated on the decoding state in parallel by the DPU, and then the information is stored to the CSFR for AIU and DLU. We propose a neuro-fuzzy ISA consisting of inference and learning instructions for neuro-fuzzy operations. These CISC type instructions can increase system throughput for AIU and DLU about 18.3% on average, compared to simple MIPS-based ISA.

C. *Delay Prediction Unit*

Since digital learning latency and analog inference latency are proportional to the number of rules and input observations, each delay can be estimated from previous exploration of physical circuit model and algorithm complexity, respectively. Analog circuit model can be defined in terms of the number of membership functions circuits and its bias current relationship; otherwise, a digital learning latency can be determined by the input observations and rules. In order to realize them, n^3 and n^2 calculation are required and realized by 1kB look-up-tables, as shown in Fig. 6. After the latency is calculated with the decoded data and parameters at the 1st step, the DPU converts the latency to the stage depth count for the asynchronous pipeline at 2nd step. It can have 3 contiguous iteration delays

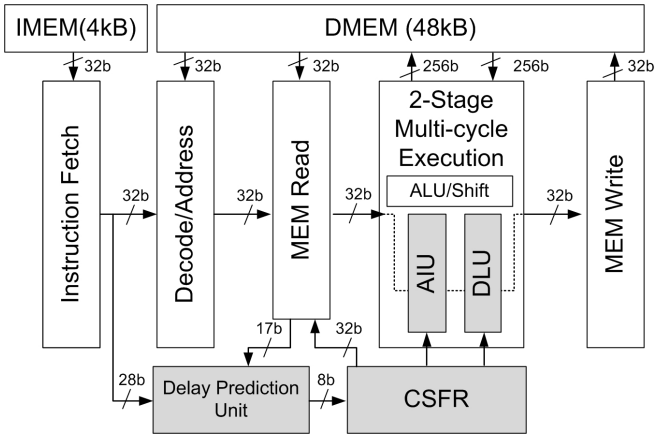


Figure 5. Block diagram of proposed neuro-fuzzy controller

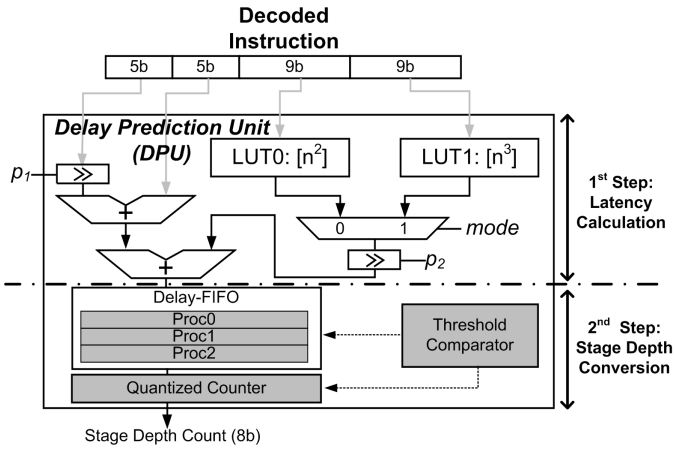


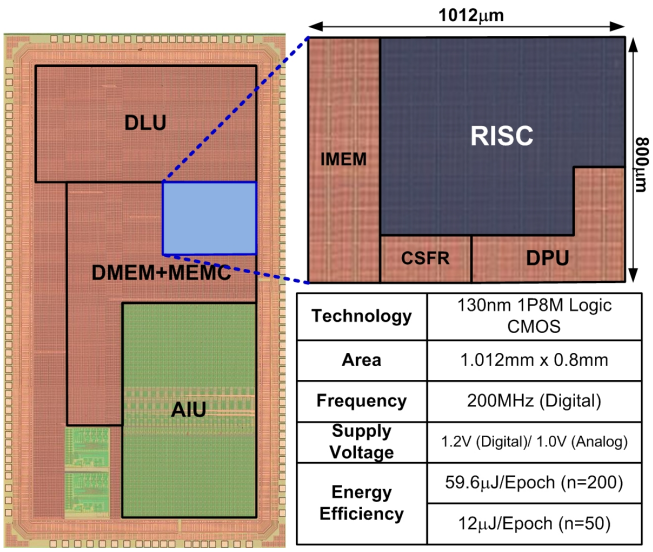
Figure 6. Delay prediction unit architecture

and generate the cycle count of stage depth. As a result, it provides discrete delay prediction, along with 3σ error for 8bit quantization, which is negligible for the system.

IV. Implementation Result

The neuro-fuzzy accelerator chip [6] is implemented in a 0.13µm mixed-mode CMOS process and contains a AIU for highly parallel inference operations and a DLU for complex learning algorithm realization. The location of neuro-fuzzy controller incorporating with CSFR and DPU is shown in Fig 7. With the help of a neruo-fuzzy controller, the SoC achieves 59.6µJ/epoch energy consumption for 200 fuzzy rules and 12.5µJ/epoch for 50 fuzzy rules that are responsible for an average 38.4% increase in the system throughput and 46.6 % reduction in power consumption by mixed-mode architecture compared to the digital equivalent processor [2].

In Fig 8, it shows the measurement result of pipelined operation in SoC performing inference and learning with 200 fuzzy rules. With the help of variable duty cycle, the inference



Technology	130nm 1P8M Logic CMOS
Area	1.012mm x 0.8mm
Frequency	200MHz (Digital)
Supply Voltage	1.2V (Digital)/ 1.0V (Analog)
Energy Efficiency	59.6µJ/Epoch (n=200)
	12µJ/Epoch (n=50)

Figure 7. Microphotograph of proposed neuro-fuzzy controller

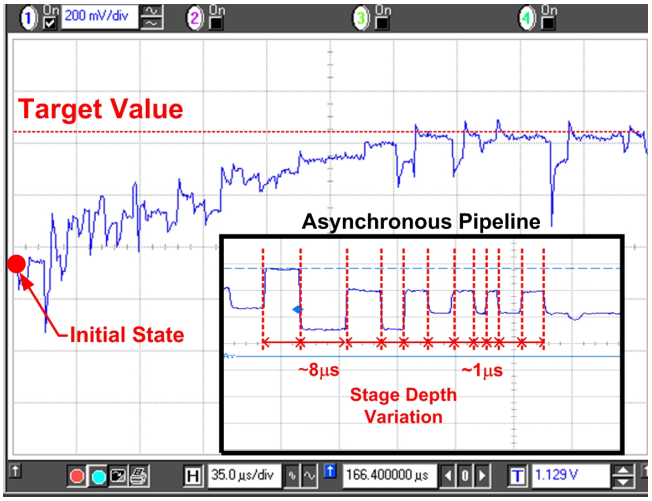


Figure 8. Measured result of the asynchronous mixed- mode pipeline

result, as the measured output voltage, is converged to the target value, The average stage depth is 5.5µs for an update and the cycle variation is maximally 4.6 µs as 83.6% adaption.

V. Conclusion

An asynchronous mixed-mode neuro-fuzzy controller to accelerate inference and learning operations of M.I. SoC has been realized. With the help of DPU and the neuro-fuzzy ISA, the mixed-mode neruo-fuzzy accelerator is deployed with the asynchronous 2-stage mixed-mode pipeline. Thanks to the pipeline and variable state depth, an energy efficient M.I. SoC of 12.5µJ/epoch is realized with an average 38.4% reduction in processing delay and an 46.6% reduction in power consumption compared to the equivalent digital processor.

References

[1] J.-Y. Kim, et al., "A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine," IEEE Journal of Solid-State Circuits, vol. 45, no. 1, pp. 32-45, 2010.

[2] S. Lee, et al., "A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition," IEEE ISSCC 2010 Dig. Tech. Papers, pp.332-333., 2010.

[3] T. Chen, et al., "A Multimedia Semantic Analysis SoC (SASoC) with Machine Learning Engine," ISSCC Dig. Tech. Papers, pp. 338-339, 2010.

[4] J. Oh, et al., "1.2mW On-Line Learning Mixed Mode Intelligent Inference Engine for Robust Object Recognition," Dig. Symp. VLSI Circuits, pp. 17-18, 2010.

[5] H. Graf, et al., "A Massively Parallel Digital Learning Processor," Adv. Neural Information Processing Systems 21, pp. 529-536, 2009.

[6] J. Oh, et al., "A 57mW Embedded Mixed-Mode Neuro-Fuzzy Accelerator for Intelligent Multi-core Processor," IEEE ISSCC 2011 Dig. Tech Papers, pp. 130-131,2011.

[7] S. Russell, el al., "Artificial Intellignece: A Mordan Approach", Prentice Hall, 1995.