# A Simultaneous Multithreading Heterogeneous Object Recognition Processor with Machine Learning Based Dynamic Resource Management

Jinwook Oh, Gyeonghoon Kim, Junyoung Park, Injoon Hong, Seungjin Lee, Joo-Young Kim, and Hoi-Jun Yoo

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)
373-1, Guseong-dong, Yuseong-gu,
Daejeon, 305-701, Republic of Korea
+82-42-350-8068, Email: jinwook@eeinfo.kaist.ac.kr

Abstract: A simultaneous multithreading multicore processor is proposed to accelerate object recognition for 720p HD video streams. The multithreading architecture with Q-learning based dynamic resource management enables concurrent processing of 8 region-of-interests with 5-stage fine grained recognition pipeline outperforming previous object recognition processors with 342GOPS computing power. In addition, the dynamic resource management contributes to increase of energy efficiency by applying the on-line learning DVFS and dynamic tile allocation based on task variance and hardware utilization to achieve 9.6mJ/frame with 1280x720 pixel image. It achieves 2.72x throughput and 3.7x energy efficiency compared to previous recognition processors.
(Keywords: Simultaneous multithreading, multicore processor, dynamic resource management, object recognition)

## 1. Introduction

Object recognition is becoming a prevalent technology for advanced video applications such as augmented reality, unmanned vehicle control and active surveillance. Especially for mobile devices, the object recognition requires extensive processing throughput and energy efficiency to support recent 720p camera/display of smartphones. The previous approach adopted multicore-like collection of single instruction and multiple data (SIMD) processing element (PE) and/or multiple instruction and multiple data (MIMD) PEs to obtain high parallelism for different tasks of object recognition [1,2]. However, this approach fails to process 720p video stream in real-time due to its insufficient computing power and low PE utilization suffering from its task dependency. This paper presents a heterogeneous multicore processor which employs a simultaneous multithreading architecture with newly proposed 5-stage pipeline to achieve more thread/data/instruction level parallelism for higher computing power. In addition, to increase the system throughput and energy efficiency of the processor, the Q-learning based dynamic resource management [3] is implemented with tile allocation for full PE utilization and with utilization-aware on-line learning DVFS for time-varying task.

## 2. Simultaneous Multithreading Heterogeneous Multicore Processor Architecture

Fig. 1 shows the proposed the simultaneous multithreading multicore processor and its fine grained pipeline as well as previous object recognition processor architectures. The homogenous SIMD architecture [1] utilizes whole 16 SIMD PEs for the processing of feature extraction (FE) stage, and it processes 2-stage recognition pipeline in combination with feature matching (FM) stage of matching processor. The heterogeneous PE architecture [2], consisting of 4 SIMD and 32 MIMD PEs, has 3-stage pipeline since each SIMD PE and MIMD PE is separately occupied by FE and feature description (FD) stages, respectively, divided from previous FE stage. Thanks to increased throughput of 3-stage pipeline, this architecture [4] can obtain 1.5x computing power improvement. However, it suffers from performance degradation due to comparatively long SIMD operation delay and low PE utilization. Therefore, in this paper the simultaneous multithreading multicore processor is proposed to increase not only system throughput with multithreading and 5-stage fine grained pipeline but also full-utilization of PE with machine learning based dynamic resource allocation. It is realized with 3 key enablers of 5-stage SMT-enabled heterogeneous architecture based on the hierarchical ring-star NoC with 6.4GB/sec/port speed as shown in Fig. 2. First, the simultaneous multithreading feature extraction cluster (SFEC) is proposed to obtain energy-efficient thread-level parallelism for FE and FD stages. Second, full-utilization of SFEC is achieved with the help of 5-stage pipeline based PE utilization control and distance-based dynamic tile allocation. Finally, the machine learning based dynamic resource controller (DRC) is implemented to apply power/throughput management dynamically for high performance and low power consumption of the proposed processor.

 978-1-4673-1202-8/12/$31.00 

### 3. Simultaneous multithreading Feature Extraction Cluster

Fig.3 shows the 16-way SIMD dual vector processing engine (DVPE) architecture and the dynamic tile allocation technique of SFEC. The DVPE integrates 3 16-bit special function units (SFU) for Gaussian function (GF), difference of Gaussian (DoG) and local min/max extraction (LME), respectively. It issues two threads at the same time for 3 SFUs and executes with 3-stage pipeline in order. Since the SFEC consists of 1 DVPE and 4 MIMD scalar processing element (SPE), connected through the local NoC with 4GB/sec/port speed, the utilization of each PE is critical to achieve highest SFEC throughput. In order to increase utilization of 4DVPEs and 16SPEs, this processor adopts the dynamic tile allocation based on the distance between individual ROIs. At most 6 of 9-neighboring tiles of ROI can be shared with consecutive ROIs, the shared tiles can be transferred directly to another thread of the same DVPE or other SFECs by using the distance information in the DRC. As a result, it can reduce the external channel occupancy to increase accumulated throughput by 17%. In addition, since the DRC prefetches the required data of SPE at the second stage DVPE pipeline to the other idling SPE, the pipeline throughput is enhanced to 47700 tiles/sec by increasing the SFEC utilization.

### 4. Q-Learning based Dynamic Resource Management

Fig. 4 shows the details of how the DRC controls the thread and power consumption of the proposed processor. The DRC adjusts the power management configuration of SFEC based on the amount of ROIs and utilization per frame. Since the dynamic resource management can estimate the optimized state transition point by adopting reconfigurable thresholds of $\Delta$ROI, $\rho$, and utilization, $\upsilon$, respectively, the optimum energy and throughput management can be selected by one of three different configurations, C0:DVFS, C1:DVFS+multithreading (MT), C2:DVFS+MT+dynamic tile allocation. According to the different resource management configuration, achievable throughput and power efficiency can vary while keeping the execution time constant with 30fps with lowest energy consumption. The 4 parameters, namely, 2 state transition thresholds and 2 configuration ROI numbers, $P_{DRM}$, can be updated by on-line learning operations with Q-learning model [3]. As a result, 9.6mJ/frame or 10.4nJ/pixel energy efficiency can be obtained from 95% utilization and 320mW power consumption.

### 5. Implementation Result

The proposed processor of Fig. 2 is fabricated in 0.13μm CMOS technology. Using COIL-100-based synthesizable database, the recognition performance is examined on the 720p HD video input stream containing high amounts of ROIs. The proposed architectures for object recognition can achieve 30fps throughput, which are 1.3x higher than the processor without DRC and 2.72x higher than the previous recognition processor. Compared to other object recognition processors with different architectures, this chip shows the highest power and area efficiency with 640GOPS/W and 10.7 GOPS/mm$^2$, respectively.

### 6. Conclusion

Simultaneous multithreading heterogeneous multicore processor is proposed for 720p HD object recognition. With the help of Q-learning based DRC, 4 SFEC process at most 8 threads in parallel and integrate 5-stage fine-grained recognition pipeline to obtain 3.45x higher tile/sec pipeline throughput. The DRC can prefetche the required SPE data and allocate required tiles dynamically for 17% reduction of top NoC occupancy. Meanwhile, the DRC performs power/throughput management by applying DVFS, MT and dynamic tile allocation gradually to obtain highest energy efficiency with full PE utilization. Thanks to the increased throughput, the object recognition SoC can achieve 30fps throughput and 10.47nJ/pixel energy efficiency, which is 2.72x and 3.7x improvement compared to the previous architectures.


[1] J.-Y. Kim, Minsu Kim, Seungjin Lee, Jinwook Oh, Kwanho Kim, and Hoi-Jun Yoo, "A 201.4 GOPS 496mW real-time multi-object recognition processor with bio-inspired neural perception engine," IEEE J. Solid-State Circuits, vol. 45, no. 1, pp. 32-45, Jan. 2010.

[2] Seungjin Lee, Jinwook Oh, Junyoung Park, Joonsoo Kwon, Minsu Kim, and Hoi-Jun Yoo, "A 345mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," IEEE J. Solid-State Circuits, vol.46, no.1, pp. 42-51, Jan., 2011.

[3] E. Ipek, O. Mutlu, J.F. Martinez, R. Caruana, "Self-optimizing memory controllers: A reinforcement learning approach," IEEE International Symposium on Computer Architecture (ISCA), pp.39-50, 2008.

[4] Jinwook Oh, Gyeonghoon Kim, Junyoung Park, Injoon Hong, Seungjin Lee, and Hoi-Jun Yoo, "A 320mW 342GOPS Real-Time Moving Object Recognition Processor for HD 720p Video Streams," ISSCC Dig. Tech. Papers 12.4, Feb., 2012.


 978-1-4673-1202-8/12/$31.00 

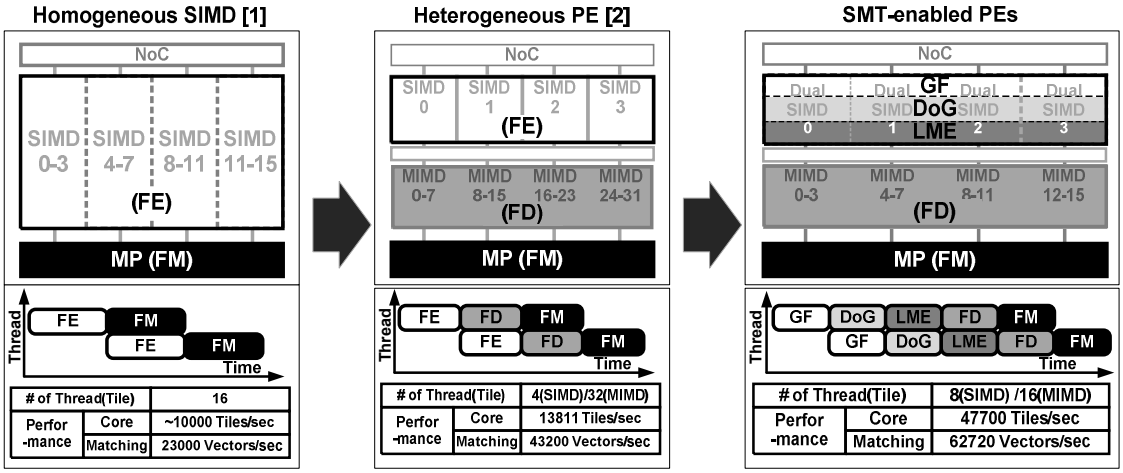



Fig. 1 Proposed simultaneous multithreading heterogeneous multicore processor for object recognition

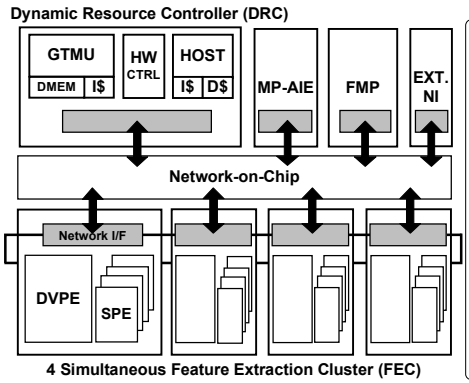



Fig. 2 Block diagram of overall processor

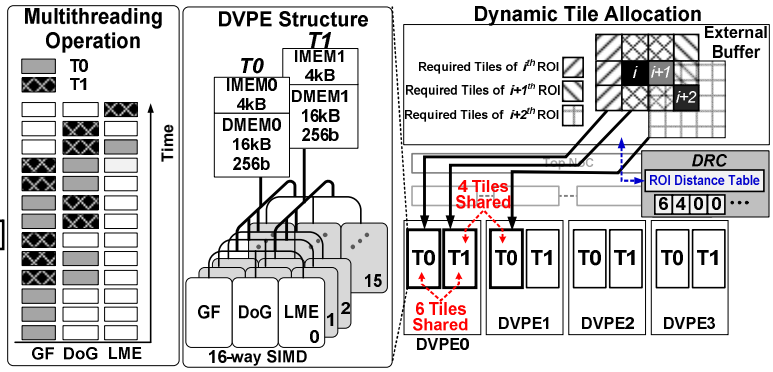



Fig. 3 DVPE architecture and dynamic tile allocation

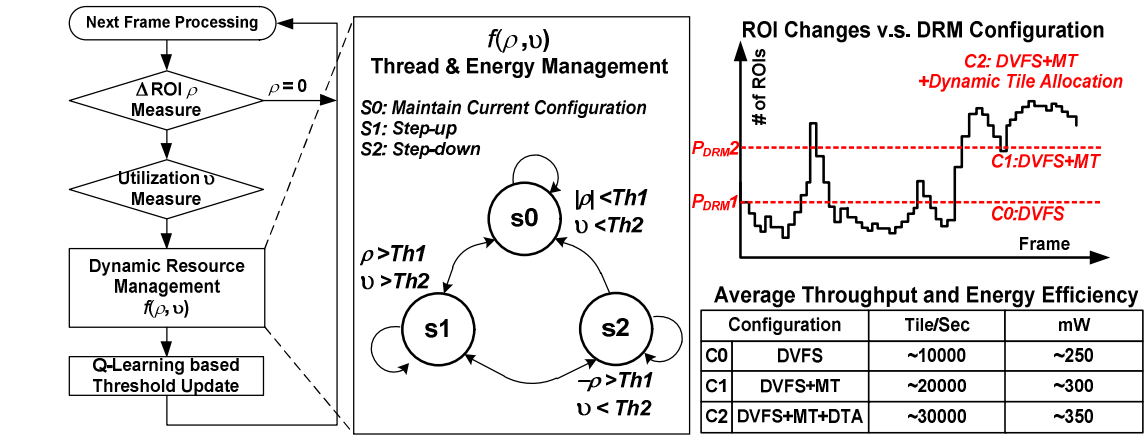
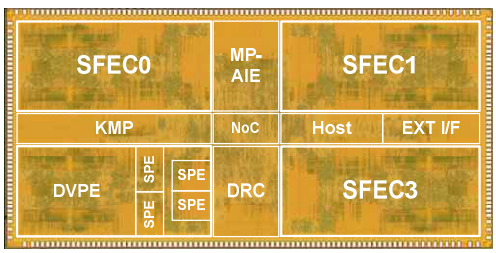



Fig. 4 Q-learning based dynamic resource management




| | | | | | |
|---|---|---|---|---|---|
| Technology | 0.13μm 1P6M CMOS | | Peak Performance | 342 GOPS | |
| Chip Size | 4.0mm x 8.0mm | | Area Efficiency | 10.69 GOPS/mm² | |
| Voltage/ Frequency | 0.7 ~1.2 V | 50~200MHz | Power Efficiency | 640GOPS/W | |
| Power Consumption | Peak | 534mW | Energy Efficiency | Frame | 9.6mJ |
| | Average | 320mW | | Pixel | 10.47nJ |

Fig. 5 Chip photograph and summary

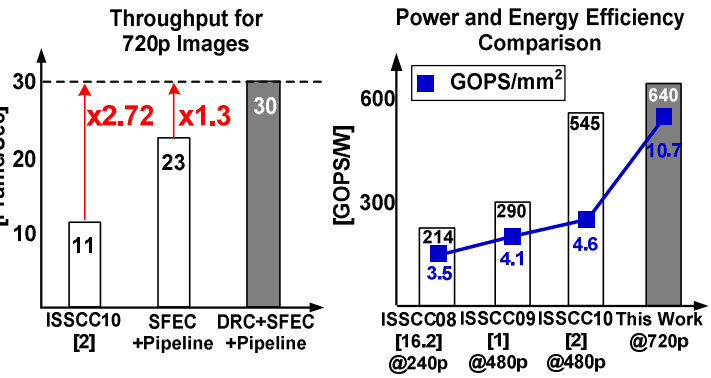



Fig. 6 Experimental results and performance comparison


978-1-4673-1202-8/12/$31.00 ©2012 IEEE