7.2 A 20.5TOPS and 217.3GOPS/mm$_2$ Multicore SoC with DNN Accelerator and Image Signal Processor Complying with ISO26262 for Automotive Applications

Yutaka Yamada, Toru Sano, Yasuki Tanabe, Yutaro Ishigaki, Soichiro Hosoda, Fumihiko Hyuga, Akira Moriya, Ryuji Hada, Atsushi Masuda, Masato Uchiyama, Tomohiro Koizumi, Takanori Tamai, Nobuhiro Sato, Jun Tanabe, Katsuyuki Kimura, Ryusuke Murakami, Takashi Yoshikawa

Toshiba Electronic Devices & Storage, Kawasaki, Japan

Vehicles with features for the advanced driver-assistance system (ADAS) are already available on the market. Technologies for image recognition are the essential elements for ADAS applications, and several SoCs, which perform image recognition for ADAS features have been developed [1-3]. As vehicles that support ADAS become more complicated, these technologies need more improvement. Recently, deep neural network (DNN) has achieved higher accuracy and can be applied to more applications such as road detection than traditional feature-based matching algorithms, so several image recognition SoCs with DNN processors, such as [4-5], are being developed. Functional safety is important for automotive applications, so several SoCs such as [2] are complied with ISO26262. The requirements for SoCs that support ADAS features are: high performance for detecting various objects, low power use to keep execution stable in the rapidly-changing environment of moving vehicles, and high safety to avoid serious accidents.

In this paper, we implement an image recognition SoC for ADAS applications. Figure 1 presents a block diagram of the SoC. As the SoC performs both image recognition processes that need high computing power and control processes that need high reliability, we partitioned the SoC into two regions. One region is called “Processing Island” (PI) for image recognition processes, and the other is called “Safe Island” (SI) for control processes. PI includes two clusters of quad ARM Cortex-A53 for application processors, four DSPs for various image processing tasks, eight kinds of hardware dedicated accelerators, a two-channel 32-bit LPDDR4 interface, 16 M bit on-chip SRAM, and several I/O interfaces. The accelerators are: 1) DNN accelerator for classification and segmentation, 2) two channel image signal processors (ISPs) for making color images from raw images, 3) AKAZE for feature point tracking, 4) STMAT for matching stereo images, 5) PYRAM for making pyramid images, 6) MATCH for object tracking, 7) AFFINE for making affine transformation images, and 8) HOX for detecting various objects. SI includes two dual core lock step (DCLS) ARM Cortex-R4s for control processors, an hardware secure module (HSM) for managing security data such as cryptographic keys and performing encryption and decryption, 8 M bit on-chip SRAM, and various I/O interfaces. Additionally, there are memory protection units (MPUs) in each accelerator and each memory unit.

Figure 2 shows an overview of the DNN accelerator. DNN applications require a lot of MAC computations for calculating many network layers, and high memory bandwidth for loading large weight data of layers such as the fully connected layer and for transferring temporary data between layers. The DNN accelerator has 1024 MAC units in its execution unit, and its performance achieves 1.6 TFLOPS. The MAC units support IEEE754 binary 16 format, because automotive applications require more accuracy than consumer applications. This accelerator has two features for reducing bandwidth of external memory: 1) the decompression circuit in the DMA unit can reduce the bandwidth of external memory by loading compressed weight data and decompressing the data; and 2) the local memory to reduce the bandwidth due to transferring temporary data. The execution unit accesses only the local memory, and the DMA unit transfers data between the external memory and the local memory.

Figure 3 shows an overview of the ISP. The ISP supports HDR images because images of automotive applications need high dynamic range to detect objects constantly, day or night. The ISP includes not only a demosaic unit, but also several image quality correction units, an HDR compression unit, an auto exposure unit, an auto white balance correction unit, and a histogram unit. As all of these functions can be performed without storing temporary data to external memory, the ISP achieves low latency and low bandwidth to the external memory. The pixel rate of each ISP is 600 M pixels / sec, and its performance is over 4.4 TOPS.

The goal of functional safety is to reduce the risk of accidents to acceptable levels. The approaches for reducing the risk are to reduce failure rates and increase fault coverage. ISO26262 describes the safety risk level called automotive safety integrity level (ASIL) which is classified as ASIL-A to ASIL-D based on its risk assessment. To reduce risk, introducing safety mechanisms (SMs) is required. As the processes of SI require more integrity than those of PI, we assume ASIL of SI (ASIL-D) is higher than that of PI (ASIL-B). Therefore, the SMs of SI are implemented in order to make them more robust than those of PI. The SMs of PI are: 1) parity/ECC circuits with SRAM. Only SRAM in the ISP has ECC circuits, because output data of the ISP is used by many processes; 2) run time BIST circuits for logic circuits; 3) ECC circuits with the bus for detecting and correcting errors during bus access; and 4) several monitor circuits such as a clock monitor. The SMs of SI are; 1) ECC circuits with SRAM, 2) duplicated circuits and monitor circuits for detecting faults in the random logic, 3) ECC circuits with the bus, and 4) several duplicated monitor circuits. Furthermore, MPUs of DDR I/F and Work RAM of PI protect the memory region allocated by SI to avoid undesired access from PI. The placement and the connection of MPUs and the accessibility of each bus master are shown in Figure 4.

The run time BIST of the accelerators can be controlled by the processors. However, it is so difficult for the processor to control the run time BIST of the ISP, because the process of the ISP begins when a new frame is received from an external imager. Therefore, a dedicated controller for the run time BIST of the ISP is introduced. The run time BIST controller of the ISP is shown in Figure 3. As the internal parameters of the ISP are used in several frames, the backup unit backs up and restores the parameters across run time BIST. The BIST unit executes the run time BIST process in the V blanking period. Figure 5 shows the timing chart of the run time BIST. The sequence of Case A is the following: to finish processing of a frame, the controller 1) backs up the parameters, 2) performs run time BIST, 3) restores the parameters. Then, 4) the ISP waits to execute the main process until receiving the next frame. Tb and Tr, which are the start times of 2) and 3), are preset by the processors. In Case B, when the end time of 1) is later than the start time of 2) due to increasing bus latency, the controller does not execute run time BIST and sends an alarm to the control processors.

Figure 6 and Figure 7 show the performance and die photograph of the SoC implemented in the 16 nm process. Its size is 94.52 mm$_2$. Peak performance achieved is as 20.5 TOPS and total power consumption is 9.78W. Power consumptions in Figure 6 are measured with fully utilized condition by simulation. And, power consumption of PI measured by the evaluation board is 2.73W in the case of executing pedestrian and vehicle detection by 2 cores of ARM Cortex-A53, 4 DSPs, and 6 kinds of accelerators including the DNN and HOX.

References:

[1]J. Tanabe, et al., “A 1.9TOPS and 564GOPS/W Heterogeneous Multicore SoC with Color-based Object Classification Accelerator for Image-Recognition Applications,” ISSCC Dig. Tech. Papers, pp. 328-329, Feb., 2015.

[2]C. Takahashi, et al., "A 16nm FinFET heterogeneous nona-core SoC complying with ISO26262 ASIL-B: Achieving 10−7 random hardware failures per hour reliability, " ISSCC Dig. Tech. Papers. pp. 80-81, Feb., 2016.

[3]K. Jason Lee, et al., "A 502-GOPS and 0.984-mW Dual-Mode Intelligent ADAS SoC With Real-Time Semiglobal Matching and Intention Prediction for Smart Automotive Black Box System," IEEE Journal of Solid-State Circuit, Vol. 52, No. 1, pp139-150, Jan., 2017.

[4]G. Desoli, et al., "A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems," ISSCC Dig. Tech. Papers, pp. 238-239, Feb., 2017.

[5]J. Lee, et al., "UNPU: A 50.6TOPS/W Unified Deep Neural Network Accelerator with 1b-to-16b Fully-Variable Weight Bit-Precision," ISSCC Dig. Tech. Papers, pp. 218-219, Feb., 2018.

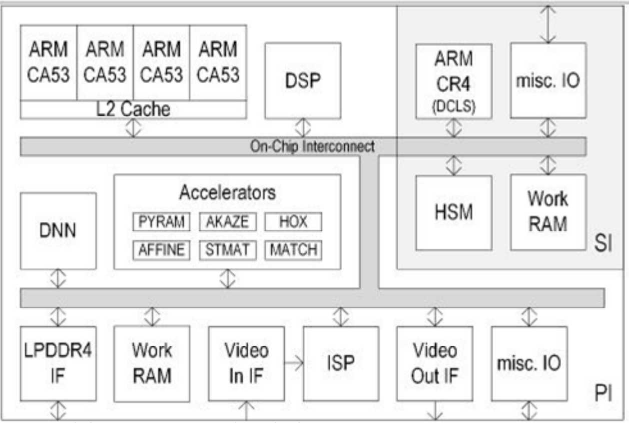


Figure 7.2.1: Block diagram of the SoC

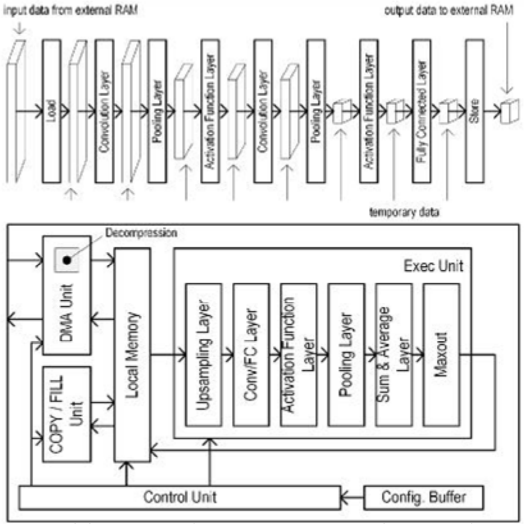


Figure 7.2.2: Example of the DNN process and Overview of the DNN accelerator

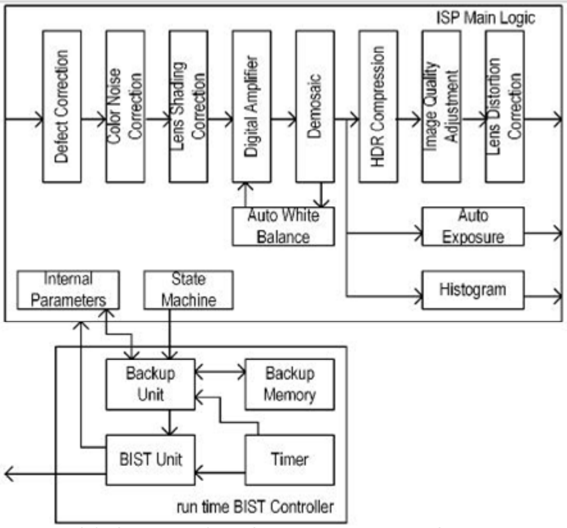
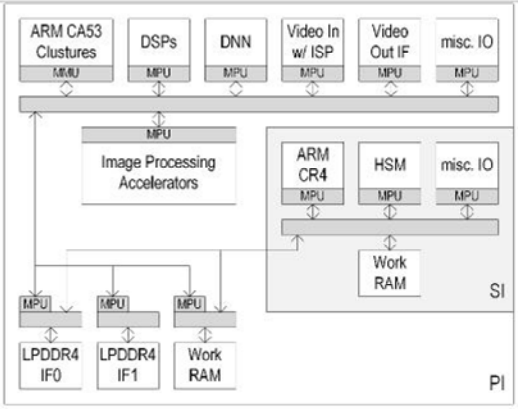


Figure 7.2.3: Overview of the ISP with the run time BIST controller



	Work RAM in SI	Work RAM in PI	LPDDR4 IF0	LPDDR4 IF1
Master in SI	Not protected	Not protected	Not protected	Never accessed
Master in PI	Never accessed	Protected	Protected	Protected

Figure 7.2.4: Connection of the MPUs and the accessibility of bus masters in the SoC

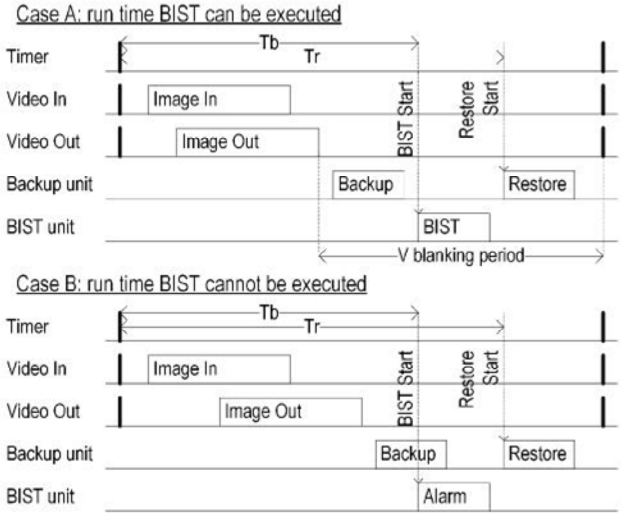


Figure 7.2.5: Timing chart of run time BIST of the ISP

Performance Evaluation of the SoC

	Performance [GOPS]	Power [mW]	Efficiency [GOPS/W]
ARM CA53	256	1,286	199.1
DSP	1,024	1,377	743.6
DNN	1,597	1,593	1,002.8
ISP	8,880	1,579	5,623.8
STMAT	3,480	419	8,305.5
AKAZE	2,681	726	3,692.6
MATCH	1,175	254	4,627.6
PYRAM	396	130	3,046.2
AFFINE	204	281	726.0
HOX	842	1,696	496.7
ARM CR4	1	435	2.8
Total	20,537	9,776	2,100.8

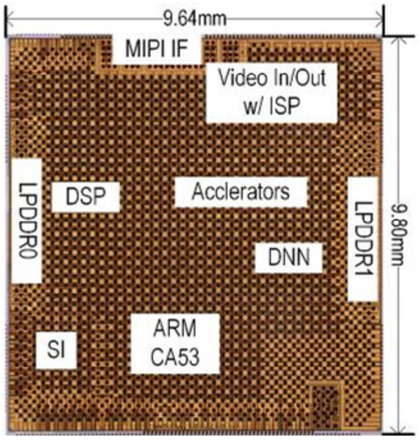
Power consumptions include functional safety logics
*Vdd=0.8V; Process=center; Temp=25°C

Performance Comparison

	This Work	Our prev. work ISSCC 2015 [1]	JSSC 2017 [3]	ISSCC 2017 [4]	ISSCC 2018 [5]
Application	Automotive	Automotive	Automotive	Embedded	IoT
Process	16nm	40nm	65nm	28nm	65nm
Peak Performance [GOPS]	20,537	1,900	502	751	346
Power Efficiency [GOPS/W]	2,101	564	862	2,930	3,080
Area Efficiency [GOPS/mm2]	217.3	18.0	31.4	21.5	21.6
Normalized Area Efficiency*	217.3	65.9	239.3	52.2	347.6

*Normalized Area ratio (16nm : 28nm : 40nm : 65nm) assume to be (1 : 2.43 : 3.66 : 11.08)

Figure 7.2.6: Performance results and comparison



Process	16nm
Chip Size	94.52mm2
On-Chip SRAM	142.9 M bit
Supply Voltage	
SI Core	0.8V
SI I/O	1.8~3.3V
PI Core	0.8V
PI I/O	1.8~3.3V

Figure 7.2.7: Chip photograph of the SoC