4.6 A 1.93TOPS/W Scalable Deep Learning/Inference Processor with Tetra-Parallel MIMD Architecture for Big-Data Applications

Seongwook Park, Kyeongryeol Bong, Dongjoo Shin, Jinmook Lee, Sungpill Choi, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Recently, deep learning (DL) has become a popular approach for big-data analysis in image retrieval with high accuracy [1]. As Fig. 4.6.1 shows, various applications, such as text, 2D image and motion recognition use DL due to its best-in-class recognition accuracy. There are 2 types of DL: supervised DL with labeled data and unsupervised DL with unlabeled data. With unsupervised DL, most of learning time is spent in massively iterative weight updates for a restricted Boltzmann machine [2]. For a ~100MB training dataset, >100 TOP computational capability and ~40GB/s IO and SRAM data bandwidth is required. So, a 3.4GHz CPU needs >10 hours learning time with a ~100K input-vector dataset and takes ~1 second for recognition, which is far from real-time processing. Thus, DL is typically done using cloud servers or high-performance GPU environments with learning-on-server capability. However, the wide use of smart portable devices, such as smartphones and tablets, results in many applications which need big-data processing with machine learning, such as tagging private photos in personal devices. A high-performance and energy-efficient DL/DI (deep inference) processor is required to realize user-centric pattern recognition in portable devices.

This paper presents a high-performance DL/DI processor (code named "K-Brain"). It has 3 key features to realize DL with high performance: 1) a deep-network learning engine (DNLE) with a dual-threaded 4-stage task-level pipeline; 2) a deep-network inference engine (DNIE) with a dynamically reconfigurable systolic PE-array (DRSA); and, 3) a true random number generator (TRNG) and dual-layered architecture (DLA) with a metastable entropy source (ES). The DNLE dramatically increases the DL processing speed by exploiting both task-level parallelism (TLP) and layer-level parallelism (LLP). The DNIE reduces the required SRAM bandwidth and enables per-cycle inference by exploiting both fine-grained parallelism (FGP) and neuron-level parallelism (NLP). The DLA and the TRNG minimize the IP-to-IP communication bandwidth and enhance the accuracy of learning and inference. With this tetra-parallel MIMD architecture, "learning-on-chip" can be realized for personal devices.

Figure 4.6.2 shows a block diagram of the K-Brain, which has a heterogeneous multicore architecture with a 2D mesh NoC connecting 4 DL cores and 2 DI cores (but not the TRNG). The TRNG is connected to the DNLE and DNIE through an independent communication path. The DNLE, with 4 DL cores inside, performs massively iterative unsupervised and supervised learning, and a DL core has 4 dual-threaded task-level pipelined datapaths (DTPD) inside for scalability. The DNIE with 2 DI cores performs a feed-forward inference task, and a DI core has 2 fine-grained pipelined per-cycle datapaths (FPPD) inside for scalability. The TRNG performs 16b RN generation.

Figure 4.6.3 shows a block diagram of the DTPD. It trains a mini-batch of data with 5 main dedicated task blocks in a 4-stage pipeline, comprised of: 1) positive inference (PI); 2) layer reconstruction (LR); 3) negative inference (NI); and, 4) negative visible-hidden layer product (NVHP). The positive visible-hidden-layer product (PVHP) task can be performed in data parallel with the other tasks, and after the PI stage, the 4 tasks can be divided into two threads; one is the PVHP thread and the other is the LR-NI-NVHP thread. On the other hand, in the supervised mode, only PI and NI are required and are performed in the PIU and the NIU; the PVHP, NVHP and LR functions are not required in supervised mode. The number of PEs in a PVHP unit can be reduced to 1/3, because the PVHP thread can take 3-pipeline-stages processing time. Since the LR unit requires only register-register operation with a bit-check scheme, it does not need a complicated multiplier, reducing area and design complexity. With the function-dedicated DTPD, the 4-stage task-level pipeline can achieve >95.3% SIMD PE utilization, 3.5× increased throughput and a 29.7% reduction in PEs.

Figure 4.6.4 shows a block diagram of the FPPD together with the inference algorithm and its results. The FPPD comprises 4 task blocks organized into a 4-stage pipeline to perform a per-cycle inference task: 1) a reconfigurable integrate unit (RIU) with DRSA for the integrate operation; 2) a biasing unit (BU) for biasing; 3) a non-linear unit (NLU) for non-linear activation; and, 4) a Gibbs-sampling unit (GSU) for stochastic neuron firing. The probability value of the nonlinear activation result is compared with the RN generated from TRNG, and if it is larger than the random number, the neuron fires. A maximum of 256 integrate-and-fire neurons are processed in each core with the combination of 2 kinds of parallelism (FGP and NLP). The DNIE-processed images of 3 intermediate layers are shown in the top right side of Fig. 4.6.4. Each layer encodes an abstract aspect of the trained dataset: layer 1 for edge-like features, layer 2 for the parts of an object, and layer 3 for the whole object. For a 3-layer deep neural network, inference takes ~20ms for 640×480 images, achieving real-time operation.

The K-Brain has 2 layers for IP-to-IP communication, as shown in Fig. 4.6.5. In the 1st layer, 4 DL cores (LC_0-LC_3) of the DNLE and 2 DI cores (IC_0-IC_1) of the DNIE are connected together to run the DL and DI algorithm for different layers in parallel. To reduce the communication bandwidth of the 1st layer for maximizing the LLP, the 2nd layer for RNs broadcasting to all IPs inside the DNLE and DNIE is separated from the 1st layer. With the DLA, the bandwidth in the 1st layer is reduced by the maximum 18.1Gb/s, and we can achieve 3.6× and 1.9× throughput increases for the DL and DI algorithms, respectively. The use of RNs is essential for 3 tasks: 1) initialization, 2) learning, and 3) inference in stochastic neuron models. The high-speed TRNG (instead of a pseudo RNG) enables more accurate DL and DI algorithms with non-deterministic properties [3]. In this work, the TRNG is implemented with the modified ES for 200MHz operation [4]. The state is determined by thermal noise in the system and outputs a random bit (RB). After RB generation, the buffered RBs are processed by digital post-processing (DP). The 256b TRN is grouped into 16 chunks of 16b TRNs, which are sent to the DNLE and DNIE in the 16-cycle burst mode, while ES-DP is clock-gated. This scheme reduces the TRNG power consumption by 38.23%.

The K-Brain performs DL and DI algorithms for image recognition, as shown in Fig. 4.6.6. The inputs are 32×32 24b color hand images. First, the DNLE trains the deep neural network with ~40K randomly selected images in unsupervised and supervised learning for hand-shape classification. After learning converges, 10 states of the top layer inferred by DNIE are compared with each other to classify the shapes of the hand. The database, the learned weights by DNLE, and the learning convergence graph are shown in the top side of Fig. 4.6.6. With TRNG, the convergence of the cost function shows a 1.8% lower error rate than that of learning without TRNG. In addition, for testing the scalability, three K-Brain chips were assembled on a test board and connected through an external NoC in an FPGA. The maximum performance of learning and inference increases by 2.2× and 2.6× compared to the single-chip implementation.

The high-performance deep learning/inference processor is fabricated using 65nm 8-metal CMOS technology, integrating 3.75M equivalent gates and 216KB of SRAM for battery-powered personal devices. It achieves 42.1× and 1.3× faster deep learning than a CPU and GPU, respectively, and consumes the 213.1mW peak power when running at 200 MHz with 1.2V supply voltage. With 411.3GOPS peak performance, the K-Brain achieves 1.93TOPS/W power efficiency – 185.6% improvement over a state-of-the-art on-chip deep-learning processor [5], and its high scalability enables multi-chip implementation to realize real-time learning and inference for deep neural networks.

References:

[1] H. Lee, *et al.*, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations," *International Conf. on Machine Learning*, pp. 609-616, 2009.

[2] G. Hinton, *et al.*, "A Fast Learning Algorithm for Deep Belief Nets," *J. of Neural Computation*, vol. 18, Issue 7, pp.1527-1554, 2006.

[3] J. G. Liao, "Variance Reduction in Gibbs Sampler Using Quasi Random Number," *J. of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 253-266, 1998.

[4] M. Hamburg, *et al.*, "Analysis of Intel's Ivy Bridge Digital Random Number Generator," *Cryptography Research Inc.*, pp. 1-22, 2012.

[5] J. Lu, *et al.*, "A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13um CMOS," *ISSCC Dig Tech. Papers*, pp. 504-505, Feb. 2014.

[6] J. Kim, *et al.*, "A 6.67mW Sparse Coding ASIC Enabling On-Chip Learning and Inference," *IEEE Symp. on VLSI Circuits*, pp. 1-2, 2014.

[7] P. Pham, *et al.*, "NeuFlow: Dataflow Vision Processing System-on-a-Chip," *IEEE Midwest Symp. on Circuits and Systems*, pp. 1044-1047, 2012.

 978-1-4799-6224-2/15/$31.00

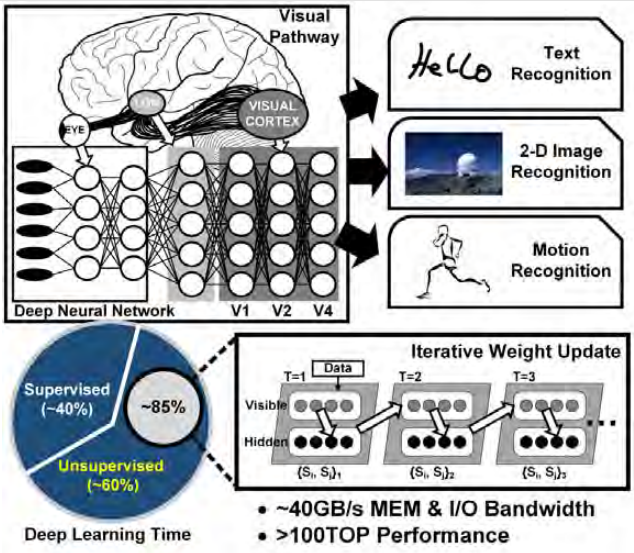


Figure 4.6.1: Deep learning algorithm, and its requirements.

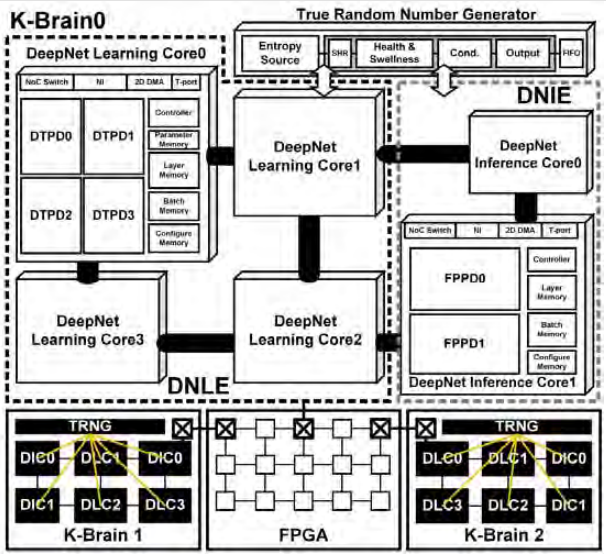


Figure 4.6.2: Overall architecture of the K-Brain system.

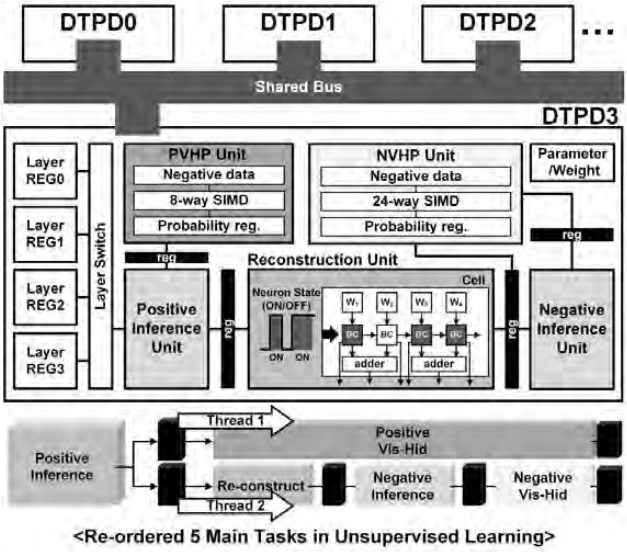


Figure 4.6.3: A 4-stage task-level pipelined deep-learning core.

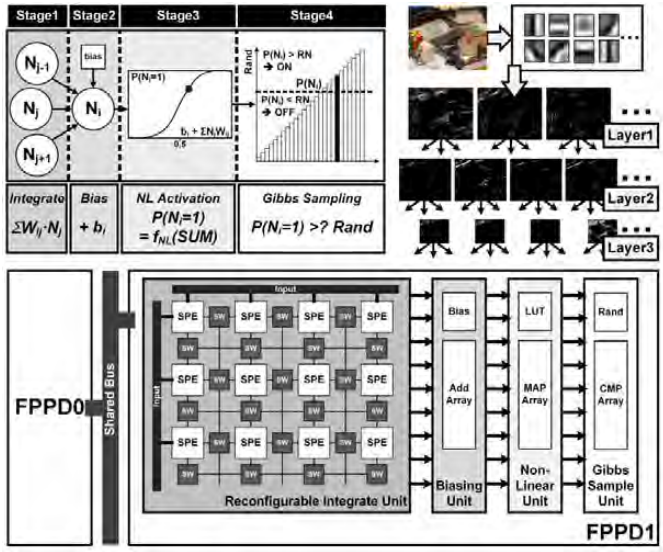


Figure 4.6.4: A 4-stage fine-grained pipelined deep-inference core.

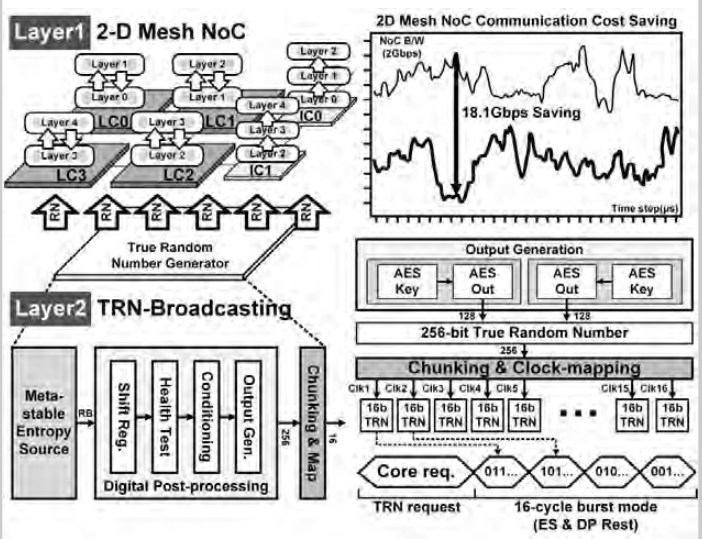


Figure 4.6.5: A TRNG and dual-layered architecture.

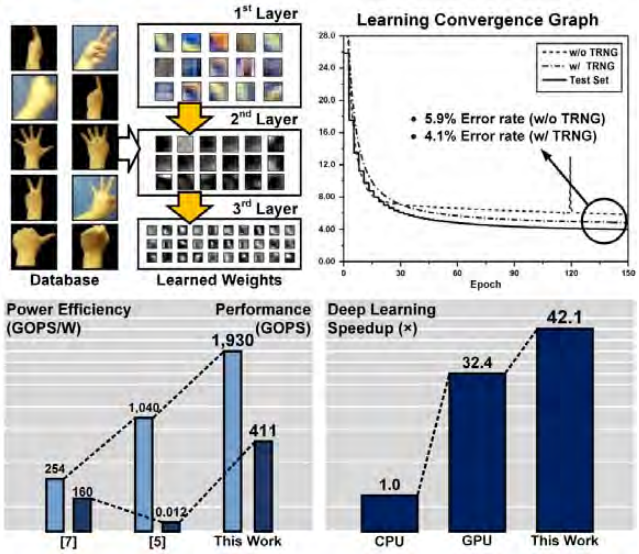
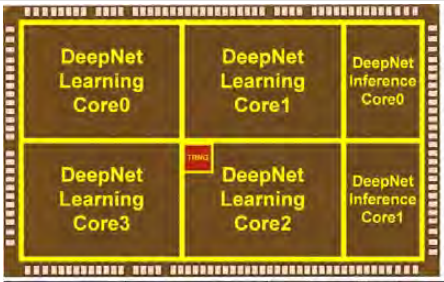


Figure 4.6.6: Measurement results with 40k hand image database.



Technology	65nm 1P8M Logic CMOS
Chip Size	2.5mm x 4.0mm
Gate Count	3.75M
SRAM	216KB
Supply Voltage	1.2V (Nominal)
CLK Frequency	200MHz (90FO4)
Power Consumption	213.1mW (Peak) 185.3mW (Average)
Peak Performance	DNLE 328.3 GOPS DNIE 82.1 GOPS TRNG 0.9 GOPS
Power Efficiency	1.93 TOPS/W (learning+inference)
Area Efficiency	41.13 GOPS/mm^2 (learning+inference)

	This Work	ISSCC'14[5]	SoVC'14[6]	MWSCAS'12[7]
Application	Big Data Analysis	Pattern Recognition	Image Processing	Image Segmentation
Function	Deep Neural Network	Unsupervised Online Clustering	Sparse Coding	Deep Neural Network
Model	Deep Network	Deep Network	Shallow Network	Deep Network
Throughput	11.8 Gb/s (learning+inference)	N/A	1.5 Gb/s (learning) 9.9 Gb/s (inference)	12.04fps @ 500×375
Technology	65nm	0.13um	65nm	45nm
Area	10.00mm^2	0.36mm^2	4.45mm^2	12.50mm^2
Performance	411.3 GOPS	0.012 GOPS	N/A	160 GOPS
Power	213.1mW (learning+inference)	27uW (learning) 11.4uW (inference)	228.1mW (learning) 218.0mW (inference)	570.0mW
Power Efficiency	1.93 TOPS/W (learning+inference)	1.04 TOPS/W (inference)	6.58 Gb/s/W (learning)	254.00 GOPS/W
Area Efficiency	41.13 GOPS/mm^2 (learning+inference)	0.03 GOPS/mm^2 (inference)	0.34 Gb/s/mm^2 (learning)	12.8 GOPS/mm^2

Figure 4.6.7: Chip micrograph and performance summary.

978-1-4799-6224-2/15/$31.00 ©2015 IEEE