## 28.2 A 1.0TOPS/W 36-Core Neocortical Computing Processor with 2.3Tb/s Kautz NoC for Universal Visual Recognition

Chuan-Yung Tsai, Yu-Ju Lee, Chun-Ting Chen, Liang-Gee Chen

National Taiwan University, Taipei, Taiwan

Unlike human brains, where various kinds of visual recognition tasks are carried out with homogeneous neocortical circuits and a unified working mechanism, existing visual recognition processors [1-4] rely on multiple algorithms and heterogeneous multicore architectures to accomplish their narrowly predefined recognition tasks. In contrast, new brain-mimicking recognition algorithms have been proposed recently [5,6]; we label such algorithms as belonging to the Neocortical Computing (NC) model. Based on neuroscience findings, NC algorithms model both the human brain's static and dynamic visual recognition streams (as shown in Fig. 28.2.1) through a series of unified matching/pooling operations. The algorithms exhibit high visual recognition capability and perform accurately on wide range of image/video recognition tasks. However, using the NC model for real-time recognition involves several hundred GOPS of both dense and sparse matrix calculations and requires over 1.5Tb/s inter-stage data bandwidth – requirements that cannot be met efficiently on existing parallel architectures.

In this paper, an NC processor for power-efficient real-time universal visual recognition is proposed with following features: 1) A grey matter-like homogeneous many-core architecture with event-driven hybrid MIMD execution provides 1.0TOPS/W efficient acceleration for NC operations; 2) A white matter-like Kautz NoC provides 2.3Tb/s throughput, fault/congestion avoidance and redundancy-free multicast with 151Tb/s/W NoC power efficiency, which is 2.7-3.9× higher than previous NoC-based visual recognition processors [1,2].

Figure 28.2.2 shows the system architecture of the Kautz NoC-based 36-core NC processor, where the Kautz graph is implemented as an NoC. A Kautz graph can be defined by its degree $d$ and diameter $k$ (both $d$ and $k$ are 3 in this work) and has $N = (d+1)d^{k-1}$ nodes ($N$ = 36 cores in this work). The Kautz NoC provides white matter-like routing capability – the maximum hop count between any 2 cores is less than $\log_d N$. Fault/congestion tolerance is also possible, as there are $d$ disjoint routing paths between any 2 cores. The grey matter-like homogeneous cores support unified NC operations and provide programmability for various recognition workloads. As the examples show in Fig. 28.2.2, such a combination provides better scalability in terms of both application and architecture compared to heterogeneous designs [1-4]. In this processor, each core is addressed by a string of 3 base-4 numbers (i.e. 0/1/2/3; core addresses are 6b in total) and adjacent numbers must be different (e.g. 122 is not a valid address). Each core unidirectionally links to 3 other cores with left-shifted addresses. For example, core 121 links to 210, 212 and 213, and receives links from 012, 212 and 312, accordingly. Through this string-shifting procedure, a core can reach any other core within 3 hops. All cores share a 32b address space, where each core is addressed by the most significant 6b and has a 26b private address space. A memory map, which defines the NC model, is loaded through two 64b system buses as memory pages.

Figure 28.2.3 shows a block diagram of the NC core and its processing element (PE). All computations are event-driven. Specifically, a core is only turned on when a packet is received from another core through the NoC, or from the system bus. Inactive cores and their components are clock gated. Arriving packets are decoded into instructions and queued in the instruction FIFO. The instruction dispatcher can issue up to 2 instructions to the PEs and paging memory units, which cover the instruction's target addresses. Page misses are handled by the memory management unit (MMU) and DMA. In the proposed hybrid MIMD mode, each of the 2 issued instructions can be either SIMD or SISD. Thus, the core can flexibly switch between maximal acceleration for dense NC operations (by SIMD) and minimal power consumption for sparse NC operations (by SISD). With this scheme, a 10.3× recognition speed increase and a 4.43× power reduction can be achieved. Inside each PE, the 16b arithmetic unit can execute up to 4 operations per cycle, i.e. 1GOPS @ 250MHz. Successful instruction execution will cause the resultant datum to be sent to the packet encoder, and fired to its subsequent target addresses through NoC.

Figure 28.2.4 shows a block diagram of the Kautz NoC router and defines its routing rules. The router performs distributed low-radix routing and all input ports share one routing FIFO, which minimizes area and improves power efficiency. All packets can be routed/multicast with the distributed routing string-based procedure, as illustrated in Fig. 28.2.4. Information regarding faulty/congested cores or NoC links can be used to correct routing strings and redirect packets for fault/congestion avoidance with minimal hop count overhead. Based on properties of a Kautz graph, the NC processor can sustain at least 2 faulty cores/links. The proposed redundancy-free multicast scheme exploits the manner in which cores are named by using disallowed name strings as group multicast addresses without header redundancies (e.g. 122 (a disallowed name) can be used to represent the cores 120, 121 and 123 as a subgroup, and 11X represents the entire core group 1). With this scheme, a distributed minimum spanning tree-based multicast is realized and provides a 1.75× recognition speed increase. Compared to a non-fault-tolerant star-/tree-based NoC [1-3], the Kautz NoC provides better efficiency (since centralized high-radix routers are area/power hungry). Compared to a fault-tolerant mesh-based NoC, the Kautz has a smaller diameter and lower routing delay (3 hops vs. 10 hops and 16% less average packet delay when running NC applications vs. a 6×6 mesh).

Figure 28.2.5 shows the single-step communication/execution scheme for this chip, which is called push-based processing. Unlike conventional parallel processors, where data are first placed on, then pulled from, caches multiple times by multiple cores/threads, push-based processing passes data and instructions as multicast packets through the energy-efficient Kautz NoC, which mimics a neuron's firing process, as shown in Fig. 28.2.1. According to each datum's associated NC operation, a SIMD or SISD instruction is executed. Zero-valued datum and datum associated with zero-valued coefficients will not result in any computation being performed, providing further power reductions. Recognition speed is improved by 2.09×, and power is reduced by 1.38×, due to push-based processing. The benefits of the features incorporated are summarized in Fig. 28.2.5. Overall, 37.8× acceleration, and a 6.25× total power reduction are achieved.

Figure 28.2.6 offers a summary of the chip's features. The NC processor is implemented on a 4.5×4.5mm$^2$ die using the TSMC 65nm CMOS process. A wide range of visual recognition applications, including image (object/face/scene) and video (action/sport), is supported and the chip works in real-time with high accuracy. Compared with other programmable visual recognition processors [1-4], high power efficiencies are achieved. The die photo is shown in Fig. 28.2.7.

*Acknowledgments:*
We thank the TSMC University Shuttle Program for chip fabrication and the National Chip Implementation Center for chip testing. This work is funded by the National Science Council. We also thank Tung-Chien Chen, Yu-Han Chen, Chih-Chi Cheng, Shao-Yi Chien, Tzu-Der Chuang, David Cox and Pai-Heng Hsiao for their valuable suggestions.

*References:*
[1] K. Kim, et al., "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-Inspired Visual Attention Engine," *ISSCC Dig. Tech. Papers*, pp. 308-309, 2008.
[2] J.-Y. Kim, et al., "A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine," *ISSCC Dig. Tech. Papers*, pp. 150-151, 2009.
[3] S. Lee, et al., "A 345mW Heterogeneous Many-Core Processor with an Intelligent Inference Engine for Robust Object Recognition," *ISSCC Dig. Tech. Papers*, pp. 332-333, 2010.
[4] T. Kurafuji, et al., "A Scalable Massively Parallel Processor for Real-Time Image Processing," *ISSCC Dig. Tech. Papers*, pp. 334-335, 2010.
[5] H. Jhuang, et al., "A biologically inspired system for action recognition," *IEEE International Conf. on Computer Vision*, pp. 1-8, 2007.
[6] J. Mutch, et al., "Object class recognition and localization using sparse features with limited receptive fields," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 45-57, 2008.

 

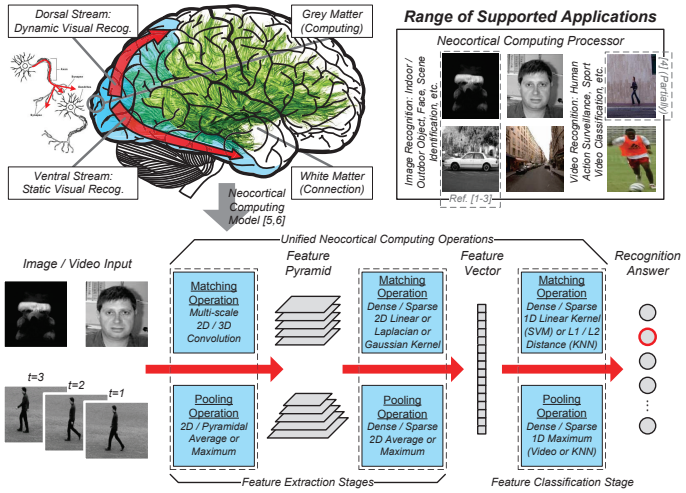


**Figure 28.2.1: NC model and supported applications of NC processor.**

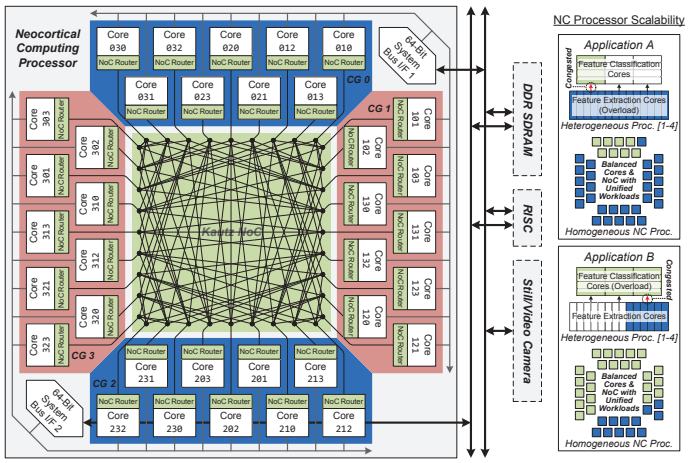


**Figure 28.2.2: NC processor architecture and scalability.**

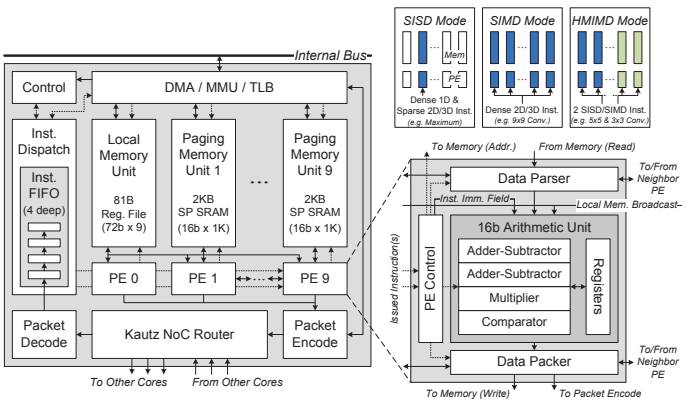


**Figure 28.2.3: Event-driven NC core/PE architectures and execution modes.**

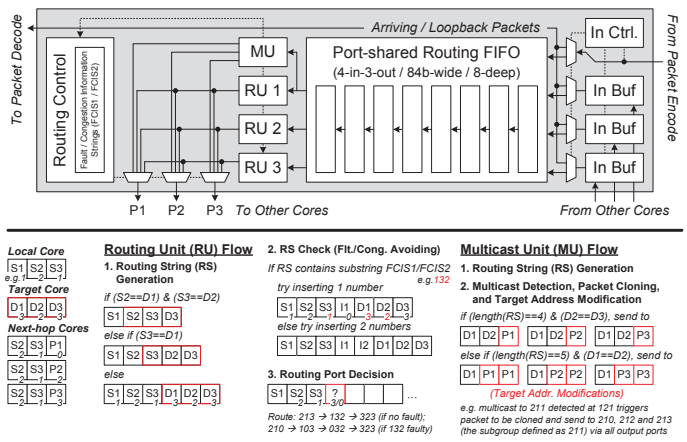


**Figure 28.2.4: Kautz NoC router architecture and routing/multicast flows.**

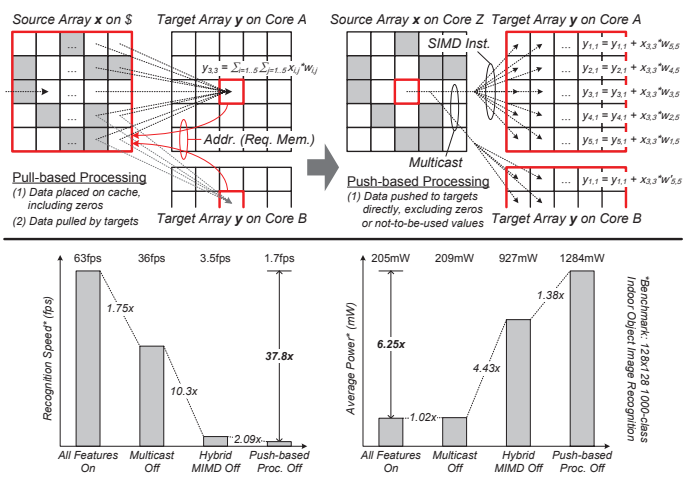


**Figure 28.2.5: Push-based processing mechanism and speed/power summary of proposed features.**

| | |
|---|---|
| Technology | TSMC 65nm 1P9M CMOS |
| Die Size | 20.25mm$^2$ (4.5mm x 4.5mm) |
| Power Supply | Core 1.0V, I/O 2.5V |
| Operating Frequency | 250MHz |
| Gate / SRAM | 2.2M / 648KB |
| Total Power Consumption | 205mW / 351mW (Average / Peak) |
| Peak Performance | 360GOPS |
| Power Efficiency | 1026GOPS/W |
| NoC Throughput | 2.3Tb/s (Aggregated Bandwidth) |

| | | | |
|---|---|---|---|
| NoC Power Consumption | 15mW | | |
| NoC Power Efficiency | 151Tb/s/W | | |
| Fault / Congestion Tolerance | 2 Cores / 2 NoC Links / 1 Core + 1 NoC Link | | |
| Recognition Speed / Feature | 128x128 Image Input (Enhanced NC Model) | 33-68fps | Measured Under Applications Shown in Fig. 28.2.1 (1* Image/Video Database) |
| | 128x128 Image Input | 63-130fps | |
| | 256x256 Image Input | 15-32fps | |
| | 128x128 Video Input | 14-26fps | |
| | Online Instance Learning | Yes (KNN Mode) | |

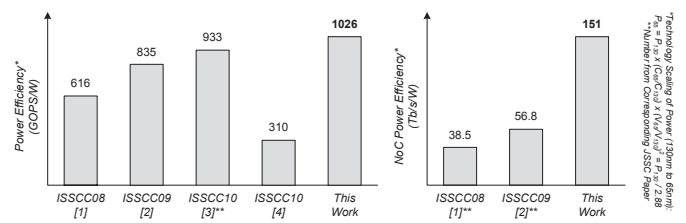


**Figure 28.2.6: Chip features and comparison.**

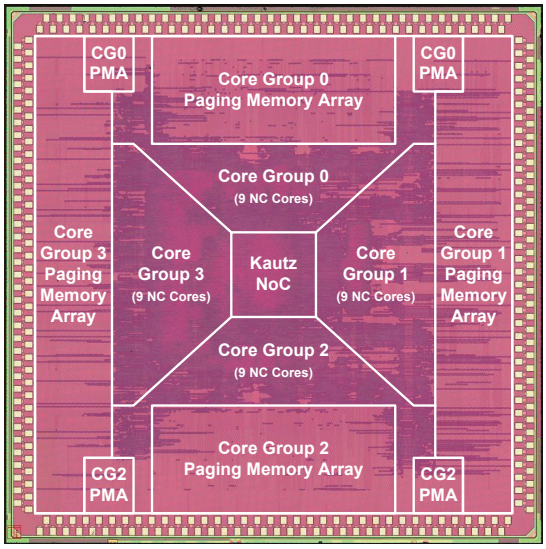


**Figure 28.2.7: Chip micrograph.**


978-1-4673-0377-4/12/$31.00 