Enabling a 2 Watt Mouse Brain Based on BCPNN Model of Cortex as a Cognition Engine for Autonomous Embedded Systems

Ahmed Hemani, Nasim Farahini, Giovanni Grandi, Dimitrios Stathis, S.M.A.H Jafri, School of ICT, KTH, Kista, Sweden, Matthias Jung, Christian Weis, Norbert Wehn, TU Kaiserslautern, Germany, Anders Lansner, School of CSC, KTH, Stockholm, Sweden.

Embedded systems are rapidly moving from interacting with predictable environment to interacting with unpredictable environments. Implementing such systems with traditional programming model is not just difficult but impossible; programming assumes a priori knowledge of the scenarios that the embedded systems need to deal with and this assumption is not valid when dealing with unpredictable situations. The only way we can solve this problem is by learning machines and the biological brain provides the best known model of learning machines. However, implementing brain models, even abstract ones, poses several challenges in terms of computational, engineering and manufacturing efficiencies. This is the challenge we address in this work: we have proposed a structured VLSI design methodology that enables custom hardware synthesis from higher abstractions. Such a design, for an abstract model of cortex of mice size can be implemented with modest engineering cost and consume about 2 watts of power in 22 nm technology node.

The Bayesian Confidence Propagation Neural Network (BCPNN) [2] is an abstract model of cortex that is based on Hebbian learning and is an attractor type of spiking neural network. The synaptic weights represent the probability of co-activation of units and unit biases represent priors in a Bayesian naïve classification scheme. These variables are used to calculate the posterior probability of firing of a unit using Bayes formula. When the probability of firing a unit exceeds a threshold it triggers a spike.

The elementary neuronal unit in BCPNN is a mini-column unit (MCU) that represents the aggregate behavior of 100 neurons in a functional cortical column. Clusters of 100 MCUs are organized in hyper-column units called HCUs. MCUs in the HCUs compete in a soft winner-take-all fashion to generate approximately 100 post-synaptic spikes per second. Each such post-synaptic spike is fanned out to 100 HCUs, thus effectively generating 10 000 post synaptic spikes per second. These post synaptic spikes travel to the destination HCUs where they become the pre-synaptic spikes. Each HCU is connected to 10 000 MCUs, some of which could be in the same HCU and receives on an average input spikes from 10 000 different MCUs per second, approximately poisson distributed in time. The synaptic weight matrix of each HCU has 10 000 input connections represented as rows and 100 columns. Each row represents the input connections from one source MCU to all the MCUs in the HCU and each column holds the input connections to a MCU.

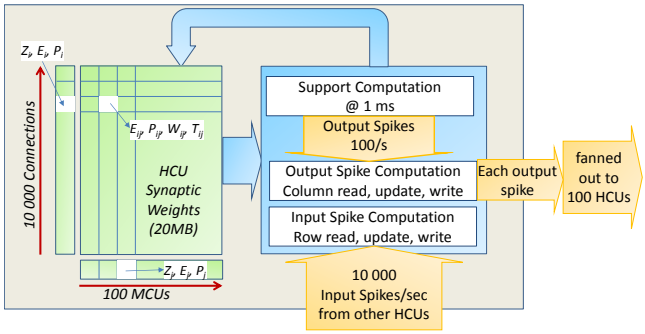

Figure 1. BCPNN Computation Model and Requirements

For a human scale cortex with about 2 million HCUs would require 1 PFlops/s computation, 50 TBs of storage for synaptic weights and even for a lazy evaluation model [4] needs a bandwidth of 250 TBs/s. Implementing such a design on the most efficient supercomputer, Riken, would need 140 kWs. Though 140 kWs is unacceptably high, in reality the actual power consumption would be significantly higher because 140 kWs only factors in 1 PFlops/s of *functional* computation. Managing millions of concurrent interacting functional computing processes induces infrastructural operations for spike propagation, delay management, spike que management, DRAM access etc. When we factor that in on general purpose machines like K at K, SpiNNaker or GPUs, the same computing machines are used for both the heavy number crunching and also the infrastructural operations. This results in under-utilization and sequentialization and is the reason why the real power consumption would be significantly more than 140 kWs. What we need is a custom design to implement concurrent processes to deal with infrastructural and computational processes. However, designing custom machines is perceived to be prohibitively expensive in terms of engineering cost.

We address this problem at a fundamental level by identifying the impediments to achieving computational, engineering and manufacturing efficiencies and proposing a solution that overcomes them to enable sufficiently low power brain models to be deployed as cognition engines for embedded systems. Achieving engineering efficiency in designing custom hardware requires design automation from higher abstractions. The VLSI design community has not made progress beyond RTL synthesis High-level synthesis has been a research topic for last 30 years with intense effort to commercialize it; however it still remains a fringe synthesis technology. We have analyzed the reasons in detail in [5, 6] and the root cause of not making progress beyond synthesis from RTL are a) while the complexity of functionality we want to synthesize has gone up 3-4 orders, the atomic physical design building blocks has remained the same boolean level standard cells in last three decades and b) no new physical design discipline has been introduced since that of the standard cell physical design scheme.

We have proposed a new structured physical design scheme called SiLago that stands for Silicon Large Grain Object. The core concepts are illustrated in Figure 2. A SiLago design is composed of SiLago blocks and laid out on a virtual grid. The SiLago blocks are a) hardened down to the layout level and are not soft RTL or gate level designs, b) characterized with post layout data to export the cost metrics (area, energy and latency in cycles) to synthesis tools and c) they occupy an exact multiple of contiguous virtual grid cells. Furthermore, the SiLago interconnects, clocks, power grids etc. are brought-out to the periphery at right positions and at right metal layers to enable composition by abutment with neighboring SiLago blocks that also have their wires being brought out to the periphery at corresponding and matching positions.

SiLago blocks are 3-4 orders larger than the standard cells and the grid based structured physical design scheme that enables composition by abutment overcomes the limitations of the standard cells that we mentioned above. Effectively, the SiLago physical design scheme reduces the gap between the higher abstraction of models of functionalities like cortex and the physical design target. This speeds up synthesis and makes it predictable because the cost metrics of micro-architectural operations and communication between them are now known with post layout accuracy. The improvement in efficiency and predictability compared to standard

cell based synthesis is 3-4 orders [5, 6]. Further, by enabling synthesis from higher abstractions provides correct-by-construction to the extent that the GDSII design is generated by machine translation and eliminates human error. Similar to elimination of gate level and circuit level verification by enabling synthesis from RTL, SiLago enables synthesis from algorithmic, application (hierarchy of algorithms) and system (multiple concurrent applications) levels and provides a rapid way to reach custom hardware solutions that are correct by construction and whose cost metrics can be predicted with significant accuracy.

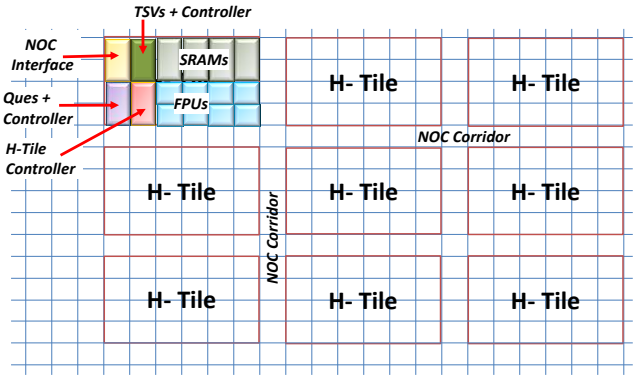

Figure 2. A Fragment of SiLago Design for mice sized BCPNN Cortex

We have applied SiLago to signal processing applications and we are also applying it to synthesis of brain models to automatically generate custom hardware solutions. The BCPNN is modeled at system level as a hierarchy of communicating sequential processes, where the concurrency and communication are explicitly modeled and exposed to the system level synthesis tool. In keeping with the spirit of SiLago - a higher abstract modeling construct should have a correspondence in higher physically design SiLago block – we also have SiLago blocks to support communication and concurrency for system level interaction of communicating concurrent processes. The SiLago vocabulary is considerably richer and more abstract than not only the boolean level standard cells but also compared to the micro-architecture level vocabulary of processors/ASIPs and accelerators. This rich vocabulary is essential for higher abstraction and the implementation of this vocabulary as SiLago blocks in a structured VLSI style is essential to support efficient and predictable synthesis from higher abstractions.

The SiLago Design flow is a multi-layer synthesis flow [6] and was used to explore the design space at system level where the trade-offs are in terms of arithmetic level parallelism, thread level parallelism, bandwidth to the memory, size of the que etc. Figure 1 and the earlier discussion on BCPNN was for a human scale cortex, for the mice sized BCPNN cortex, we need 1.4 TFlops/s and 32K HCUs, where each HCU requires 12 Megabits of synaptic weight storage. The design was organized in terms of 82 mm^2 dies integrated with 8 layers of 3D integrated DRAM. Each die is divided into 32 H-Tiles of 2.52 mm^2 area as shown in Figure 2 and each H-Tile accommodates 64 mice sized HCUs and their synaptic weights in their private 3D integrated DRAMs above them. The 3D DRAM for each H-Tile has two 64 Mb banks per layer organized in a mirror fashion with area for 254 TSVs in between providing 128 bit data bus. Each die implements 2048 HCUs and we need 16 such dies for a mice sized BCPNN cortex that we propose integrating using 2.5D interposer in a package.

We synthesized the design in 40 nm node on the SiLago platform and know the energy and area numbers with the certainty of post layout sign-off quality data. This data was conservatively scaled to 22 nm node. The area and energy data for 3D DRAM was obtained from accurate circuit level models. The total energy consumption for the entire mice BCPNN logic layer is 5.846 Joules composed of 4.032 Joules for the computation delivering close to 1.4 TFlops/s 1.814 Joules for the SRAM buffers and other infrastructural operations involving DRAM controllers, ques, spike propagation and other miscellaneous controllers. The 3D integrated DRAM for the synaptic weights cost 6.912 Joules. The total cost is 9.878 Joules over a period one second, i.e, approximately 10 watts of power consumption. The above is an absolute worst case assuming that each row in each HCU is triggered in each second. In reality, the activity in BCPNN is very sparse and has a high degree of temporal locality, i.e, when a row is triggered, it is highly probable that it will be triggered again in next 10 ms or so. By exploiting these properties, we can easily scale down power by 2-3 X. Further, the present implementation uses single floating point representation. We plan to move to a lower resolution fixed point representation that will reduce the size and energy required for the storage and also arithmetic. Collectively, these factors will bring the overall power consumption down to the target 2 watt range. The estimated value for a human size BCPNN cortex is less than 2 kWs.

REFERENCES

[1] Nasim Farahini, Ahmed Hemani, Anders Lansner, Fabian Clermidy, Christer Svensson, "A Scalable Custom Simulation Machine for the Bayesian Confidence Propagation Neural Network Model of the Brain," 19th ASP-DAC), 2014, Singapore.

[2] C. Johansson, A. Lansner, "Towards Cortex Sized Artificial Neural Systems," in *Neural Networks (Elsevier)*, vol. 20, 2007.

[3] A. Lansner, A. Holst, "A higher order Bayesian neural network with spiking units," *in Int. J. of Neural Systems*, vol. 7, 1996.

[4] A. Lansner, A. Hemani, N. Farahini, "Spiking Brain Models: Computation, Memory and Communication Constraints for Custom Hardware Implementation," in *IEEE ASP-DAC*, 2014.

[5] Nasim Farahini, Ahmed Hemani, Hassan Sohofi, Shuo Li, "Physical Design Aware System Level Synthesis of Hardware," in IEEE Intl. Conf. on Embedded Computer Systems, Architecture, Modeling, and Simulation, SAMOS 2015, Greece.

[6] Ahmed Hemani, Nasim Farahini, Syed M A H Jafri, Hassan Sohofi, Li Shuo, "The SiLago Solution: Architecture and Design Methods for a Heterogeneous Dark Silicon Aware Coarse Grain Reconfigurable Fabric", Chapter in the Book "The Dark Side of Silicon", Springer 2016, ISBN 978-3-319-31594-2