Dynamic Voltage and Frequency Scaling for Neuromorphic Many-Core Systems

Sebastian Höppner, Yexin Yan, Bernhard Vogginger, Andreas Dixius,
Johannes Partzsch, Felix Neumärker, Stephan Hartmann,
Stefan Schiefer, Stefan Scholze, Georg Ellguth, Love Cederstroem,
Matthias Eberlein, Christian Mayr
Technische Universität Dresden
Dresden, Germany
Email: {sebastian.hoeppner}@tu-dresden.de

Steve Temple, Luis Plana,
Jim Garside, Simon Davison,
David R. Lester, Steve Furber
University of Manchester
Manchester, UK
Email: {steve.furber}@manchester.ac.uk

***Abstract*—We present a dynamic voltage and frequency scaling technique within SoCs for per-core power management: the architecture allows for individual, self triggered performance-level scaling of the processing elements (PEs) within less than 100ns. This technique enables each core to adjust its local supply voltage and frequency depending on its current computational load. A demonstrator chip has been implemented in 28nm CMOS technology, containing 4 PEs which are operational within the range of 1.1V down to 0.7V at frequencies from 666MHz down to 100MHz; the effectiveness of the power management technique is demonstrated using a standard benchmark from the application domain. The particular domain area of this application specific processor is real-time neuromorphics. Using a standard benchmark - the synfire chain - we show that the total power consumption can be reduced by 45%, with 85% baseline power reduction and a 30% reduction of energy per neuron and synapse computation, all while maintaining biological real-time operation.**

***Index Terms*—MPSoC, neuromorphic computing, power management, DVFS, synfire chain**

I. Introduction

Digital neuromorphic hardware systems [1], [2] allow efficient implementation of neuromorphic computing for technical applications such as image recognition or robotics control applications. Especially purely digital many core architectures allow for energy efficiency implementations which are scalable to nanometer technologies. For those systems energy efficiency is critical especially for mobile, battery powered application scenarios or large scale brain-size scientific computing with system scaling limitations by power supply and cooling.

State-of-the art MPSoCs e.g. for mobile communication [3], [4] contain power management techniques such as DVFS or AVFS to enhance their energy efficiency. Here power management of compute cores is orchestrated by a central management unit which schedules tasks to the cores and issues their execution at a specific supply voltage and clock frequency level. In contrast to this, neuromorphic SoCs typically do not contain a task scheduling unit. Each processing element (PE) executes the neuromorphic computation (neuron state calculation, synaptic updates) based on the connectivity of the network, the assignment of neurons to the PE and the stimulus of the network. Therefore, its actual workload varies both statically with the network mapped to the system and dynamically during the simulation of the experiment.

The application of DVFS for neuromorphics is promising, since neural networks show significant variations in the dynamics of activity, making them inherently energy efficient. This is to be supported by neuromorphic hardware.

This work presents an approach for fined-grained per-core DVFS for neuromorphic SoCs, where each PE can dynamically change its performance level (PL) based on local activity.

II. Neuromorphic SoC Architecture

A. Overview

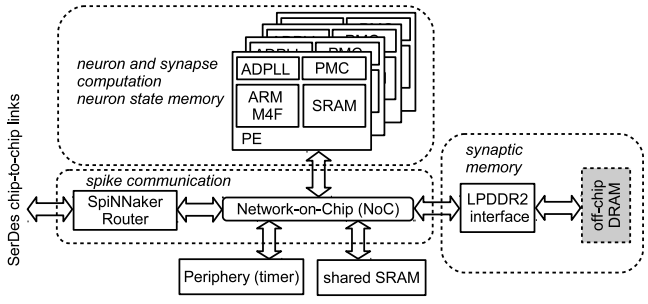


Fig. 1. System architecture

Fig.1 shows the block diagram of a neuromorphic many core SoC. It is based on the architecture from [1]. The processing elements (PEs) contain ARM M4F cores for neuron state calculation and synapse processing. All PEs are clocked in globally asynchronous locally synchronous (GALS) scheme. A peripheral timer is used for time-base generation (e.g. 1ms) derived from the reference clock signal, independent from the actual frequency setting of the PEs. Spike communication is realized by the SpiNNaker router architecture [5], connecting 6 serial off-chip links for chip-to-chip communication. Synaptic memory is realized in off-chip DRAM connected via an LPDDR2 memory interface. All on-chip components are connected by a network-on-chip (NoC).

B. Power Management Hardware Architecture

Fig. 2 shows the power management architecture of the PEs within the proposed neuromorphic many core system. It is adapted from [6] and [4]. Each PE is equipped with a local ADPLL [7] for GALS clocking and can be connected by PMOS header power switches to one out of three on-chip supply rails at different voltage levels. The PE is in power-shut-off if all switches are opened.

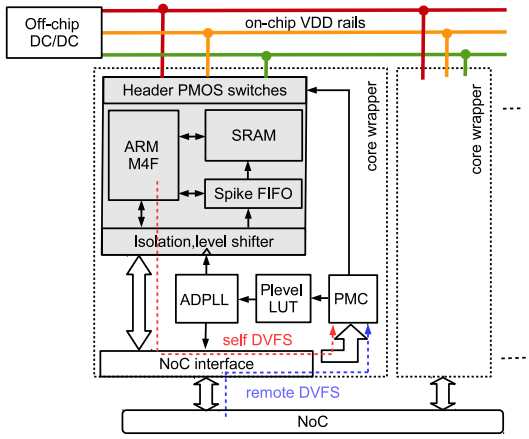


Fig. 2. PE DVFS architecture

Voltage switching and frequency changes are scheduled by the power management controller (PMC) [6]. A performance level (PL) consists of a (V_{DD}, f) pair. For a PL change the PMC controls the sequence of clock disable, supply selection and pre-charge (rush current reduction), frequency selection and clock enable as shown in Fig. 3. All timings are configurable in integer multiples of the reference period of 10ns. Fast PL switching can be achieved within below 100ns. In addition the PMC supports scenarios for power-on and power-shut-off.

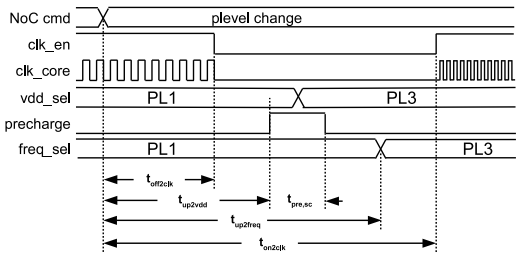


Fig. 3. PE DVFS Timing of performance level change

The PMC receives power management command from the network-on-chip (NoC) interface. In a power-up or remote DVFS scenario, these power management commands can be sent by another core, which for example orchestrates system boot-up. During the neuromorphic application, the PE can change its performance level by issuing a power management command for PL change via NoC packet to itself (self DVFS). Using this architecture the PE can change its PL by software within a very short time frame. Thereby PL changes do not result in significant latency or software overhead. This enables implementation of the application specific power management algorithms completely in software at the local PEs.

C. Power Management Software Architecture

The computational load in neuromorphic simulations is determined by the neuron state updates and synaptic events. While the neuron processing cost is constant in each simulation cycle, the number of synaptic events to be processed per time and core strongly varies with network activity. Our approach for neuromorphic power management exploits this by periodically adapting the power level to the current amount of arrived spikes per PE.

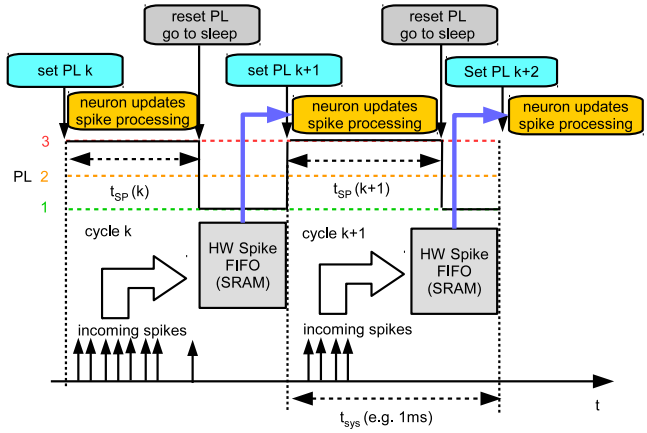


Fig. 4. Software flow for event-driven neuromorphic simulation with DVFS

Fig. 4 visualizes the flow of a neuromorphic simulation with DVFS. It is built upon the event-driven simulation methodology for spiking neural networks on SpiNNaker from [8]. A peripheral timer generates a real-time tick (t_{sys} (e.g. 1ms), which triggers the update of neuron dynamics and synapse processing. Within a simulation cycle of length t_{sys} spikes are received by the PE from the SpiNNaker router over the NoC. Incoming spikes are registered in an event queue assisted by a hardware FIFO connected to the local SRAM as shown in Fig. 2. While spikes of cycle k are received those from cycle $k-1$ are processed without interrupting the processor at incoming spikes. Based on the filling level l of the queue at the beginning of a cycle k its workload can be estimated. From this the PL of cycle k is determined by setting thresholds for the compute performance of the three available PLs, reading:

$$PL(k) = \begin{cases} PL1, \text{if } l < l_{\mathrm{th},1} \\ PL2, \text{if } l_{\mathrm{th},1} \leq l < l_{\mathrm{th},2} \\ PL3, \text{if } l_{\mathrm{th},2} \leq l \end{cases} \quad (1)$$

Then synaptic event processing and neuron state computation is performed at $PL(k)$. When these tasks are completed after the spike processing time $t_{\mathrm{sp}}(k)$ the processor is set back to PL1 and sleep mode is activated. The optimization target for PL selection is to maximize t_{sp} within a single t_{sys} period, since this relates to the usage of the minimum required PL to complete the neuron and synapse processing tasks while

maintaining biological real-time operation. This approach allows application for a wide range of event based simulations of spiking neural networks, as for example BCPNN [9].

III. Results

A. Testchip

A testchip has been implemented in 28nm SLP CMOS. Its chip photo is shown in Fig. 5. It contains 4 PEs based on ARM M4F cores with 128kB local memory and the proposed power management architecture. An LPDDR2 interface to 128MByte off-chip DRAM is used for synaptic memory.

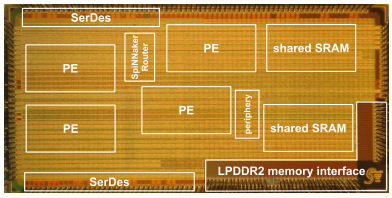
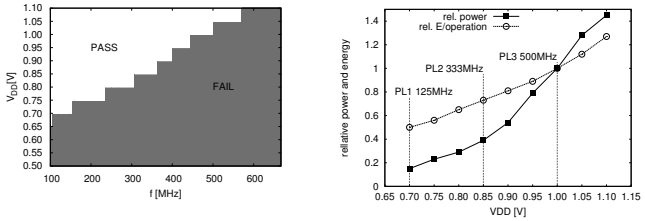


Fig. 5. Chip Photo

B. PE Measurement Results

Fig. 6(a) shows the frequency f vs. supply voltage V_{DD} shmoo plot of the ARM M4F based PE. Safe operation is possible down to 0.7V. Fig. 6(b) shows the scaling of the energy-per-task metric of this PE when scaling V_{DD} and f. The three PLs are defined as PL1 (0.70V,125MHz), PL2 (0.85V,333MHz) and PL3 (1.00V,500MHz), respectively.



(a) PASS/FAIL shmoo plot (b) power and energy per operation

Fig. 6. Processing element (ARM M4F) DVFS measurements

C. Neuromorphic Computation Example

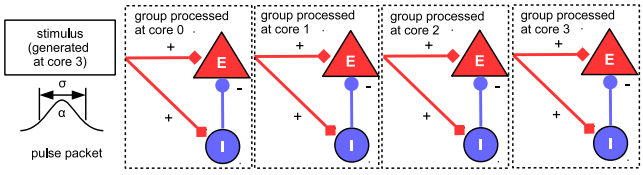


Fig. 7. Synfire chain benchmark, One group E: 200 excitatory neurons I: 50 inhibitory neurons, connectivity: 25 presynaptic connections per neuron from I to E, 60 presynaptic connections per neuron from E of previous group, delays within a group 8ms, between groups 10ms

A synfire chain network [10] serves as benchmark for the power management: Synfire chains are feedforward networks that propagate synchronous firing activity through a chain of

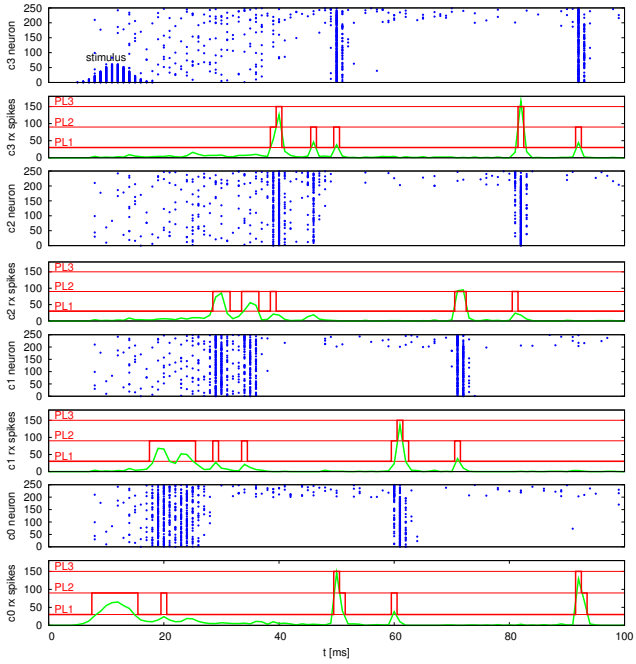


Fig. 8. Synfire chain benchmark spike train example, send spikes (blue), number of received spikes per core (green) and core PL (red)

neuron groups [11]. Compared to other typical benchmarks like sparse random networks, they create a more biologically realistic scenario of switching between phases of asynchronous and synchronous activity, cf. [12]. We implement a synfire chain with feedforward inhibition [10] consisting of 4 groups (Fig. 7), each with 200 excitatory and 50 inhibitory neurons. Excitatory neurons are connected to both excitatory and inhibitory neurons of the next group, while inhibitory neurons only connect to the excitatory population of the same group. There are no recurrent connection within a population. We simulate one group per core and connect the last group to the first one. At start, the first group receives a Gaussian stimulus pulse packet generated on core 3 (400 spikes, $\sigma = 2.4$ms). As shown in Fig. 8, the pulse packet propagates stably from one group to another, where the feedforward inhibition ensures that the network activity does not explode.

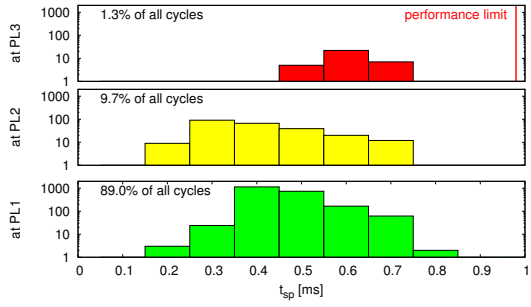


Fig. 9. Histogram of simulation cycles (1ms) processed at different PLs

As shown in Fig. 8, cores adapt their PLs to the number

of incoming spikes within the current 1ms simulation cycle. Fig. 9 shows histograms of the cycles being processed at a particular PL versus t_{sp}. Within some cycles being processed at PL3 spikes occur simultaneously such that their processing t_{sp} requires up to 0.8ms, where 1ms is the real-time constraint. Thus, the system is close to its performance limit. A conventional system without DVFS would have to be operated at PL3. In the DVFS approach only a little percentage of cycles are processed at higher PLs, thereby achieving nearly the energy efficiency of the low voltage operation at PL0.

Tab. I summarizes the power measurement results of the system for the synfire chain benchmark. Power is measured similar to the concept from [13]. Using DVFS, baseline power can be reduced by $\approx 80\%$ and energy consumption for neuron and synapse processing by $\approx 35\%$ without loss of performance of the neuromorphic experiment. Tab. II compares the achieved energy consumptions to other neuromorphic chips.

TABLE I
SYNFIRE CHAIN BENCHMARK POWER RESULTS

4×250 neurons; $4 \times 20k$ synapses; $35k$ spikes/s; $2.8M$ synaptic events/s	only at PL3 (1.0V)	only at PL1 (0.7V) [1]	DVFS
total [mW] [2]	129.6	66.2	70.9
infrastructure [3] [mW]	48.2	48.2	48.2
baseline [4] [mW]	70.2	13.7	15.5
neural [5] [mW]	7.7	3.7	4.8
synaptic [6] [mW]	3.5	0.6	2.4

[1] spike losses occur
[2] excluding unused components
[3] timer, router, LPDDR2
[4] cores active, no calculation
[5] neuron state calculation
[6] synapse processing

TABLE II
EFFICIENCY COMPARISON OF NEUROMORPHIC REAL-TIME CHIPS

Ref.	[14]	[13]	[15]	[2]	this
system type	analog sub-Vt	MPSoC	mixed-signal	custom digital	MPSoC
tech [nm]	800	130	28	28	28
neuron power [nJ/ms]	20 to 100	26	25	0.040	4.82
E/synaptic event [nJ]	0.9	8	n.a.	0.045	0.83

IV. CONCLUSION

A DVFS power management approach for event-based neuromorphic real-time simulations on MPSoCs has been presented. Its effectiveness has been demonstrated with a 28nm CMOS prototype. For a neuromorphic benchmark application, baseline power and energy consumption for neuromorphic processing can be significantly reduced compared to non-DVFS operation while maintaining biological real-time operation.

ACKNOWLEDGMENT

This work was supported by the European Union under Grant Agreements No. 604102 and DLV-720270 (Human Brain Project) and the Center for Advancing Electronics Dresden (cfaed). The authors thank ARM and Synopsis for IP and the Vodafone Chair at Technische Universität Dresden for contributions to RTL design.

REFERENCES

[1] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber, "Spinnaker: A 1-w 18-core system-on-chip for massively-parallel neural network simulation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug 2013.

[2] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, Oct 2015.

[3] M. Winter, S. Kunze, E. Adeva, B. Mennenga, E. Matus, G. Fettweis, H. Eisenreich, G. Ellguth, S. Höppner, S. Scholze, R. Schüffny, and T. Kobori, "A 335Mb/s 3.9mm2 65nm CMOS flexible MIMO detection-decoding engine achieving 4G wireless data rates," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb. 2012, pp. 216 –218.

[4] B. Noethen, O. Arnold, E. Perez Adeva, T. Seifert, E. Fischer, S. Kunze, E. Matus, G. Fettweis, H. Eisenreich, G. Ellguth, S. Hartmann, S. Höppner, S. Schiefer, J.-U. Schlusler, S. Scholze, D. Walter, and R. Schüffny, "A 105GOPS 36mm2 heterogeneous SDR MPSoC with energy-aware dynamic scheduling and iterative detection-decoding for 4G in 65nm CMOS," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, Feb 2014.

[5] J. Navaridas, M. Luján, L. A. Plana, S. Temple, and S. B. Furber, "Spinnaker: Enhanced multicast routing," *Parallel Computing*, vol. 45, pp. 49 – 66, 2015, computing Frontiers 2014: Best Papers. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167819115000095

[6] S. Höppner, C. Shao, H. Eisenreich, G. Ellguth, M. Ander, and R. Schüffny, "A power management architecture for fast per-core DVFS in heterogeneous MPSoCs," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, May 2012, pp. 261 –264.

[7] S. Höppner, S. Haenzsche, G. Ellguth, D. Walter, H. Eisenreich, and R. Schüffny, "A fast-locking ADPLL with instantaneous restart capability in 28-nm CMOS technology," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 60, no. 11, pp. 741–745, Nov 2013.

[8] T. Sharp, L. A. Plana, F. Galluppi, and S. Furber, *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ch. Event-Driven Simulation of Arbitrary Spiking Neural Networks on SpiNNaker, pp. 424–430.

[9] B. Vogginger, R. Schüffny, A. Lansner, L. Cederström, J. Partzsch, and S. Höppner, "Reducing the computational footprint for real-time bcpnn learning," *Frontiers in Neuroscience*, vol. 9, p. 2, 2015. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2015.00002

[10] J. Kremkow, L. U. Perrinet, G. S. Masson, and A. Aertsen, "Functional consequences of correlated excitatory and inhibitory conductances in cortical networks," *Journal of Computational Neuroscience*, vol. 28, no. 3, pp. 579–594, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10827-010-0240-9

[11] M. Abeles, "Synfire chains," vol. 4, no. 7, p. 1441, 2009.

[12] G. Buzsáki and K. Mizuseki, "The log-dynamic brain: how skewed distributions affect network operations," *Nature Reviews Neuroscience*, vol. 15, no. 4, pp. 264–278, 2014.

[13] E. Stromatias, F. Galluppi, C. Patterson, and S. Furber, "Power analysis of large-scale, real-time neural networks on spinnaker," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Aug 2013, pp. 1–8.

[14] G. Indiveri, E. Chicca, and R. Douglas, "A vlsi array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *Trans. Neur. Netw.*, vol. 17, no. 1, pp. 211–221, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1109/TNN.2005.860850

[15] C. Mayr, J. Partzsch, M. Noack, S. Hänzsche, S. Scholze, S. Höppner, G. Ellguth, and R. Schüffny, "A biological-realtime neuromorphic system in 28 nm cmos using low-leakage switched capacitor circuits," *IEEE transactions on biomedical circuits and systems*, vol. 10, no. 1, pp. 243–254, 2016.