

Demonstrating BrainScaleS-2 Inter-Chip Pulse-Communication using EXTOLL

Tobias Thommes, Sven Bordukat, Andreas Grübl, Vitali Karasenko, Eric Müller, Johannes Schemmel

thommes@kip.uni-heidelberg.de
Kirchhoff-Institute for Physics
Heidelberg, Germany

ABSTRACT

The BrainScaleS-2 (BSS-2) Neuromorphic Computing System currently consists of multiple single-chip setups, which are connected to a compute cluster via Gigabit-Ethernet network technology. This is convenient for small experiments, where the neural networks fit into a single chip. When modeling networks of larger size, neurons have to be connected across chip boundaries. We implement these connections for BSS-2 using the EXTOLL networking technology. This provides high bandwidths and low latencies, as well as high message rates. Here, we describe the targeted pulse-routing implementation and required extensions to the BSS-2 software stack. We as well demonstrate feed-forward pulse-routing on BSS-2 using a scaled-down version without temporal merging.

CCS CONCEPTS

• **Hardware** → **Networking hardware**; *Neural systems*; • **Networks** → *Network protocol design*; *Naming and addressing*; • **Computer systems organization** → *Neural networks*; • **Software and its engineering** → *Software functional properties*.

KEYWORDS

brain-inspired computing, neuromorphic, spike routing, FPGA, low latency, packet-based network

ACM Reference Format:

Tobias Thommes, Sven Bordukat, Andreas Grübl, Vitali Karasenko, Eric Müller, Johannes Schemmel. 2022. Demonstrating BrainScaleS-2 Inter-Chip Pulse-Communication using EXTOLL. In *Neuro-Inspired Computational Elements Conference (NICE 2022)*, March 28-April 1, 2022, Virtual Event, USA. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3517343.3517376>

1 INTRODUCTION

The EXTOLL network technology [2, 5, 8, 9] is based on the Tourmalet Network Interface Card (NIC). It offers 7 links and implements all the switching and interfacing capabilities, necessary to build an HPC network. Each EXTOLL link can comprise up to 12 lanes of 8.4 Gbit s^{-1} each. The NIC can be connected to a host computer through its PCIe x16 Gen3 connector. In an EXTOLL network, the nodes are usually, but not necessarily connected in a 3D-Torus

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

NICE 2022, March 28-April 1, 2022, Virtual Event, USA

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9559-5/22/03.

<https://doi.org/10.1145/3517343.3517376>

topology, which offers good scaling characteristics. Routing of messages through the network is done by the Tourmalet chips and is based on the 16 bit destination node-address.

BSS-2 as a mixed-signal neuromorphic computing system is built upon the HICANN-X (HX) chip which features 512 adaptive exponential integrate and fire (AdEx) neuron-circuits and $512 \times 256 = 131\,072$ synapses [7]. Up to 16 k synaptic inputs per neuron are configurable by combining neuron circuits. Realizing large networks with such neurons requires a multi-chip system. [1, 3, 10, 12] Recently the BSS-2 system development advanced to a multi-chip system featuring 46 HX chips, each connected to a Kintex 7 FPGA through $8 \times 1 \text{ Gbit s}^{-1}$ serial links. These systems make use of the BSS-1 wafer module infrastructure, imitating a full wafer-scale implementation by placing many chips on a large PCB of the exact same size and pin-configuration as a BSS-1 wafer [13, 15]. We consider the topology described in [16] to be optimal for interconnecting multiple FPGAs on wafer modules regarding bandwidth and network diameter.

Figure 1 shows the current lab setup for testing the BSS-2 EXTOLL networking [7, 14]. It is physically connected to the EXTOLL network via USB 3.0 plugs attached to the FPGAs' MGT-ports. Additionally it is still connected to an Ethernet network for FPGA bitfile flashing purposes. This setup contains four FPGAs and two chips.

2 HOST COMMUNICATION

In order to integrate the EXTOLL network with the existing BSS-2 software stack [6], we use the custom protocol layer *Neuromorphic Hardware Transport Layer for EXTOLL (NHTL EXTOLL)*. This layer sits on top of the EXTOLL network's API for Remote Direct Memory Access (RDMA), *librma2* [9], and beneath the FPGA software

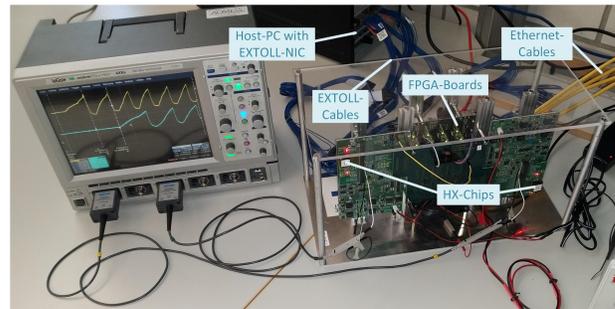


Figure 1: BSS-2 Lab Setup connected to EXTOLL network. Membrane-traces on two connected HX-chips are shown.

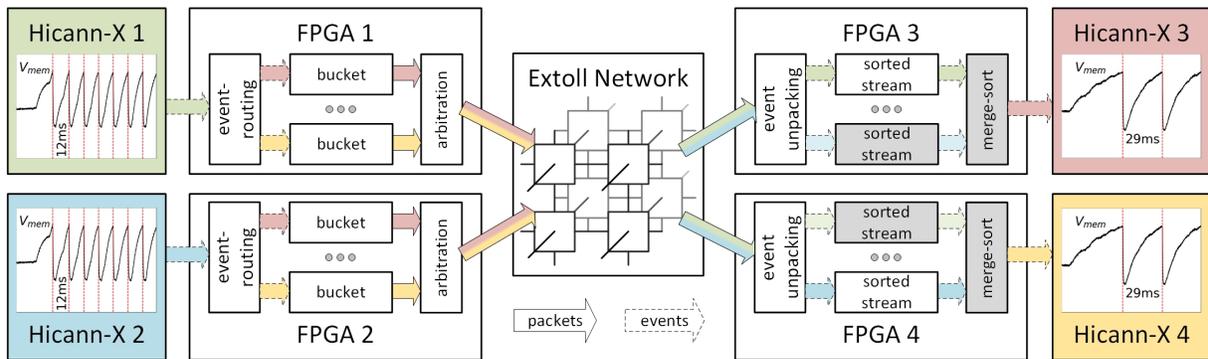


Figure 2: Experiment Setup for an inter-chip feed forward neural network. The Inter-Spike-Interval (ISI), given in bio-time units, is higher at target neurons, as they do not fire with each received input-event.

interface (*hxcomm*) [6]. Thereby, the existing experiment-flow can transparently use the EXTOLL network.

2.1 The EXTOLL Protocol

The EXTOLL network chip uses RDMA to facilitate low-latency communication using its Remote Memory Access (RMA) unit. The RMA unit consists of three sub-units, the Requester, Responder, and Completer. These handle the different aspects of each RDMA *put* or *get* operation. Messages can be issued with flags such that the sub-units produce notifications upon forwarding them. [9] These notifications are used to ensure that API commands are properly executed. Additionally, as they can carry small amounts of payload data, we use notification packets to synchronize the FPGA’s send queue with a ring buffer in the host memory, where the FPGA sends its data using RMA-messages.

2.2 Interaction with the BSS Software Stack

For integration with the BSS-2 software stack, *NHTL EXTOLL* provides two main functionalities: First, it creates and manages the necessary buffers on the host node and configures the FPGAs using the EXTOLL network’s Remote Registerfile Access (RRA) feature. Second, it provides wrapper functions for receiving and sending data via RDMA with the same syntax used by the higher level components of the software stack. With this, the EXTOLL network can be used without having to touch the pre-existing infrastructure provided by the higher abstraction levels [7].

3 INTER-FPGA COMMUNICATION

The pulse-communication architecture described for the BSS-1 system in [16] can easily be adapted for BSS-2. Events from the chip now arrive at the FPGA with rates of up to two events per 125 MHz FPGA clock cycle and comprise of a 14 bit source neuron address and an 8 bit timestamp [4]. The latter has to be converted to an arrival deadline by adding a modeled axonal delay. The lookup table at the source-node now no longer yields a GUID as in [16], since the destination multicast is not needed with a single chip per FPGA. Instead, the lookup now provides a freely remappable destination neuron address.

3.1 Event Aggregation

Figure 2 shows the flow of event-streams through the system with ascending timestamps. As described in [16], pulse events are aggregated into larger network packets [11] using bucket-buffers. The number of events to accumulate is subject to a trade-off between minimizing header-overhead and avoiding congestion when merging packetized event-streams at the destination. Also, to avoid timestamp expiration and resulting event-loss, the possible time for aggregation is limited by the modeled axonal delays.

To keep the first prototype implementation simple, the bucket-renaming proposed in [16] and the merging (gray boxes in Figure 2) are not yet realized. Instead, the destination lookup simply yields a bucket-index and the network addresses are statically configured in the buckets. In this simplified approach, the required numbers of bucket-units and merge-buffers scale with the number of desired destinations and source-streams per chip.

4 INTER-CHIP FEED-FORWARD NEURAL NETWORK

Here we present a technical demonstration of a feed forward neural network spanning two or more HICANN-X chips. A population of neurons on a first chip, driven by external input emits spikes that are then transmitted to a second chip. There, they trigger a second population of neurons to answer this firing. First measurements using an oscilloscope, attached to analog probing pins yield an overall inter-chip-latency of approximately $8\mu\text{s}$ (Figure 1). The membrane-voltage traces in Figure 2 have been recorded using the on-chip Membrane Analog to Digital Converter (MADC). The synapses have been configured for high technical efficacy in this demo to reliably elicit spikes.

5 SUMMARY AND OUTLOOK

We have presented a first hardware implementation and the according software integration of a low-latency, high-bandwidth communication strategy for multi-chip BSS-2 systems. HICANN-X chips can be interconnected by transmitting pulse events between FPGAs through a packet-based HPC network. This is done using the EXTOLL networking technology. At the conference we present a technical demonstration of these first experiments.

ACKNOWLEDGMENTS

We thank all present and former members of the Electronic Vision(s) (KIP) and Computer Architecture (ZITI) research groups contributing to the BrainScaleS EXTOLL communication hardware and software as well as the EXTOLL company for their technical support with their hardware. The research has received funding from the EC Horizon 2020 Framework Programme under Grant Agreements 720270, 785907 and 945539 (HBP), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster), the German Federal Ministry of Education and Research under grant number 16ES1127 as part of the *Pilotinnovationswettbewerb Energieeffizientes KI-System*, by the Helmholtz Association Initiative and Networking Fund (Advanced Computing Architectures, ACA) under project number SO-092 and the Lautenschläger-Forschungspreis 2018 for Karlheinz Meier.

REFERENCES

- [1] Sebastian Billaudelle, Yannik Stradmann, Korbinian Schreiber, Benjamin Cramer, Andreas Baumbach, Dominik Dold, Julian Göltz, Akos F. Kungl, Timo C. Wunderlich, Andreas Hartel, Eric Müller, Oliver Breitwieser, Christian Mauch, Mitja Kleider, Andreas Grübl, David Stöckel, Christian Pehle, Arthur Heimbrecht, Philipp Spilger, Gerd Kriene, Vitali Karasenko, Walter Senn, Mihai A. Petrovici, Johannes Schemmel, and Karlheinz Meier. 2020. Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. <https://doi.org/10.1109/iscas45731.2020.9180741>
- [2] Holger Fröning, Mondrian Nüssle, Heiner Litz, Christian Leber, and Ulrich Brüning. 2013. On achieving high message rates. In *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*. 498–505. <https://doi.org/10.1109/CCGrid.2013.43>
- [3] Andreas Grübl, Sebastian Billaudelle, Benjamin Cramer, Vitali Karasenko, and Johannes Schemmel. 2020. Verification and Design Methods for the BrainScaleS Neuromorphic Hardware System. *Journal of Signal Processing Systems* 92, 11 (2020), 1277–1292. <https://doi.org/10.1007/s11265-020-01558-7>
- [4] Vitali Karasenko. 2020. *Von Neumann bottlenecks in non-von Neumann computing architectures*. Ph. D. Dissertation. Universität Heidelberg. <http://archiv.ub.uni-heidelberg.de/volltextserver/28691/1/KarasenkoPhD.pdf>
- [5] Heiner Litz, Holger Fröning, Mondrian Nüssle, and Ulrich Brüning. 2008. VELO: A novel communication engine for ultra-low latency message transfers. In *2008 37th International Conference on Parallel Processing*. 238–245. <https://doi.org/10.1109/ICPP.2008.85>
- [6] Eric Müller, Elias Arnold, Oliver Breitwieser, Milena Czierlinski, Arne Emmel, Jakob Kaiser, Christian Mauch, Sebastian Schmitt, Philipp Spilger, Raphael Stock, Yannik Stradmann, Johannes Weis, Andreas Baumbach, Sebastian Billaudelle, Benjamin Cramer, Falk Ebert, Julian Göltz, Joscha Ilmberger, Vitali Karasenko, Mitja Kleider, Aron Leibfried, Christian Pehle, and Johannes Schemmel. 2022. A Scalable Approach to Modeling on Accelerated Neuromorphic Hardware. *Frontiers in Neuromorphic Engineering* (2022). Submitted.
- [7] Eric Müller, Christian Mauch, Philipp Spilger, Oliver Julien Breitwieser, Johann Klähn, David Stöckel, Timo Wunderlich, and Johannes Schemmel. 2020. Extending BrainScaleS OS for BrainScaleS-2. *arXiv preprint* (March 2020). <http://arxiv.org/abs/2003.13750>
- [8] Mondrian Nüssle, Benjamin Geib, Holger Fröning, and Ulrich Brüning. 2009. An FPGA-based custom high performance interconnection network. In *2009 International Conference on Reconfigurable Computing and FPGAs*. IEEE, 113–118. <https://doi.org/10.1109/ReConFig.2009.23>
- [9] Mondrian Nüssle, Martin Scherer, and Ulrich Brüning. 2009. A resource optimized remote-memory-access architecture for low-latency communication. In *2009 International Conference on Parallel Processing*. IEEE, 220–227. <https://doi.org/10.1109/ICPP.2009.62>
- [10] Christian Pehle, Sebastian Billaudelle, Benjamin Cramer, Jakob Kaiser, Korbinian Schreiber, Yannik Stradmann, Johannes Weis, Aron Leibfried, Eric Müller, and Johannes Schemmel. 2022. The BrainScaleS-2 Accelerated Neuromorphic System with Hybrid Plasticity. *Frontiers in Neuroscience* 16 (2022). <https://doi.org/10.3389/fnins.2022.795876>
- [11] Ajay Rajkumar and Michael D Turner. 2008. Packet aggregation for real time services on packet data networks. US Patent 7,391,769.
- [12] Johannes Schemmel, Sebastian Billaudelle, Philipp Dauer, and Johannes Weis. 2020. Accelerated Analog Neuromorphic Computing. *arXiv preprint* (2020). <https://arxiv.org/abs/2003.11996>
- [13] Johannes Schemmel, Daniel Brüderle, Andreas Grübl, Matthias Hock, Karlheinz Meier, and Sebastian Millner. 2010. A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling. In *Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1947–1950. <https://doi.org/10.1109/ISCAS.2010.5536970>
- [14] J. Schemmel, A. Grübl, S. Hartmann, A. Kononov, C. Mayr, K. Meier, S. Millner, J. Partzsch, S. Schiefer, S. Scholze, R. Schüffny, and M. Schwartz. 2012. Live demonstration: A scaled-down version of the BrainScaleS wafer-scale neuromorphic system. In *Proceedings of the 2012 IEEE International Symposium on Circuits and Systems (ISCAS)*. 702–702. <https://doi.org/10.1109/ISCAS.2012.6272131>
- [15] Sebastian Schmitt, Johann Klähn, Guillaume Bellec, Andreas Grübl, Maurice Güttler, Andreas Hartel, Stephan Hartmann, Dan Husmann, Kai Husmann, Sebastian Jeltsch, Vitali Karasenko, Mitja Kleider, Christoph Koke, Alexander Kononov, Christian Mauch, Eric Müller, Paul Müller, Johannes Partzsch, Mihai A. Petrovici, Bernhard Vogginger, Stefan Schiefer, Stefan Scholze, Vasilis Thanasoulis, Johannes Schemmel, Robert Legenstein, Wolfgang Maass, Christian Mayr, and Karlheinz Meier. 2017. Neuromorphic Hardware In The Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System. *Proceedings of the 2017 IEEE International Joint Conference on Neural Networks (2017)*. <https://doi.org/10.1109/IJCNN.2017.7966125>
- [16] Tobias Thommes, Niels Buwen, Andreas Grübl, Eric Müller, Ulrich Brüning, and Johannes Schemmel. 2021. BrainScaleS Large Scale Spike Communication using Extoll. In *2021 8th Neuro Inspired Computational Elements Workshop (NICE'2020)*. peer-reviewed extended abstract incl. paper presentation.