Demonstrating BrainScaleS-2 Inter-Chip Pulse Communication using EXTOLL

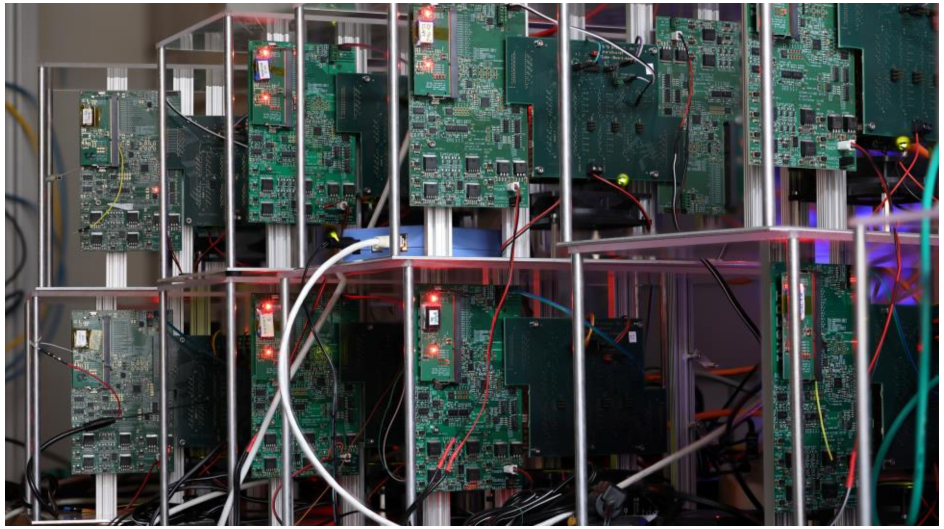
Tobias Thommes, Sven Bordukat, Andreas Grübl, Vitali Karasenko,
Eric Müller, and Johannes Schemmel

Kirchhoff-Institute for Physics, Heidelberg University

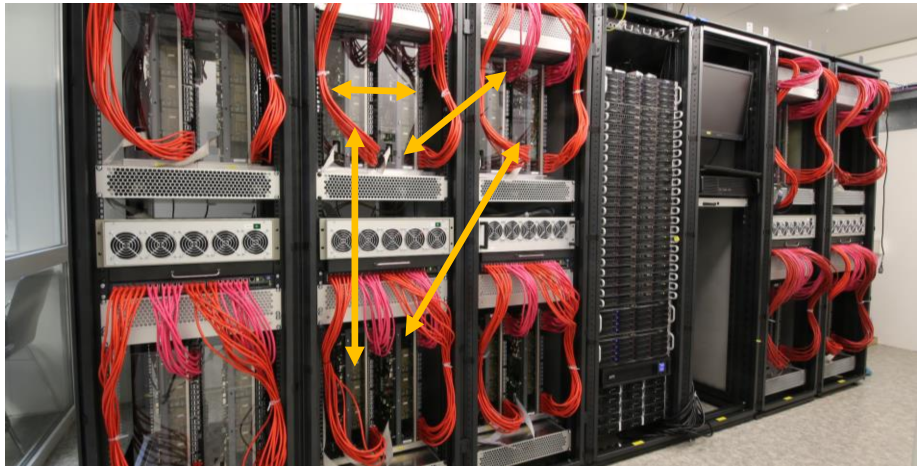
Lightning Talk by Tobias Thommes at NICE2022

The BrainScaleS Architecture

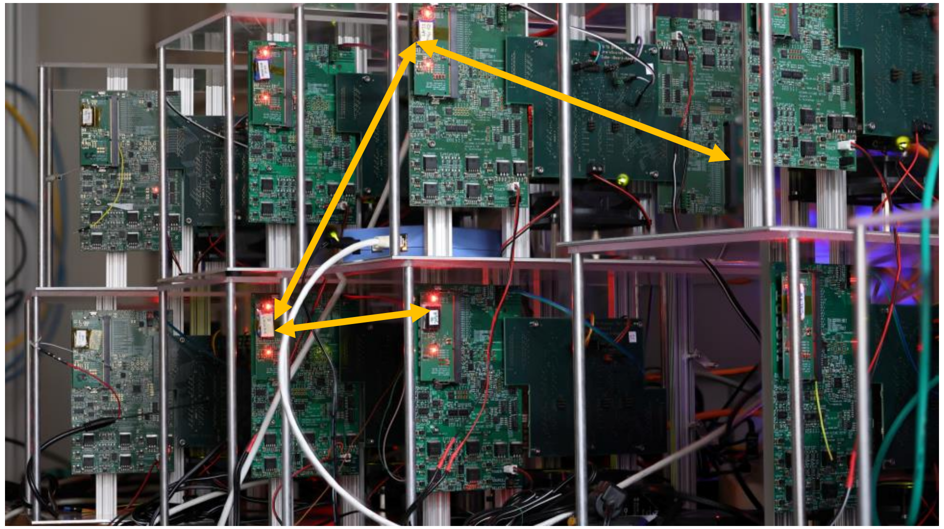
- Emulating biological neuron dynamics in analog electronic circuits
- Spike events are digitised and communicated between neuron circuits
- Accelerated model dynamics (compared to biology)
- BSS-1 ASIC:
  - Acceleration factor: 10^5
- BSS-2 ASIC:
  - Acceleration factor: 10^4
  - 2 embedded SIMD processors



Scaling BrainScaleS



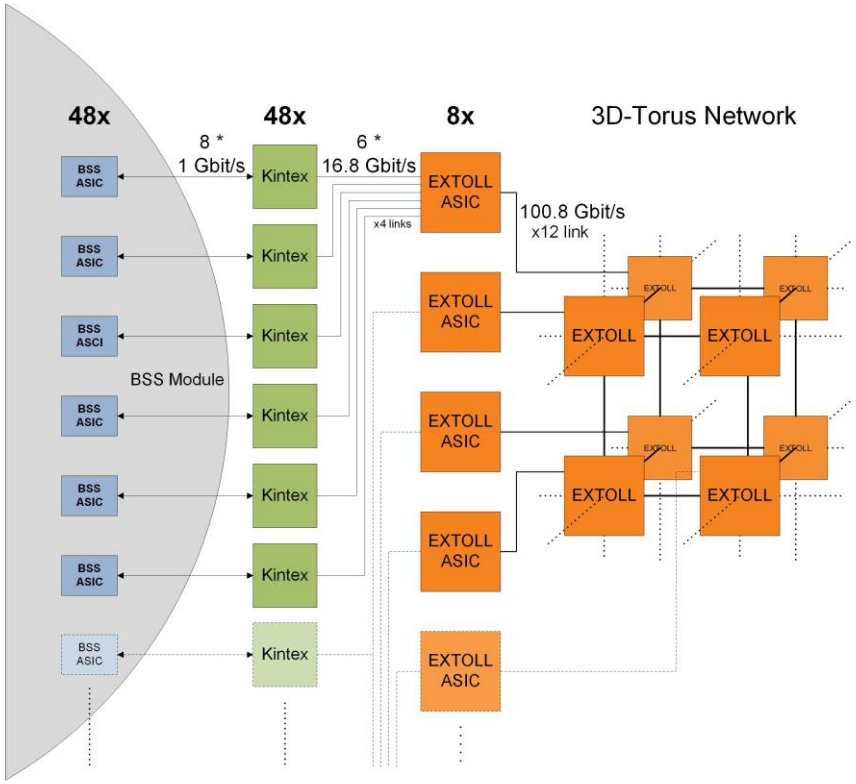
- Connection of Chips / Wafer Modules:
  - Send neural events over packet-based network
- Network requirements (due to speedup)
  - Low latency (O(ms) bio → O(µs) hw)
  - High message rate (O(Hz) bio → O(kHz) hw)
  - small packets (one event ≈ 4 Byte)
  - High bandwidth (≥ 8 Gbps)
- Ethernet
  - high protocol- / header-overhead
    (packet ≥ 64 Bytes for 8 Bytes payload)
- Infiniband
  - Proprietary hardware / IP with limited access



EXTOLL Network

- EXTOLL Company is spin-off from Heidelberg University
- Developed by Computer Architecture Group at ZITI Institute
- Bandwidth:
  - General: 100.8 Gbps
  - We can use: 16 Gbps
- Latency:
  - General: 70ns per hop (@630MHz clk)
  - We can use: 150ns per hop (@300MHz clk)
- Smallest packet:
  - Overall: 40 Byte
  - Payload: 8 Byte
- Topology freedom
  - 7 links

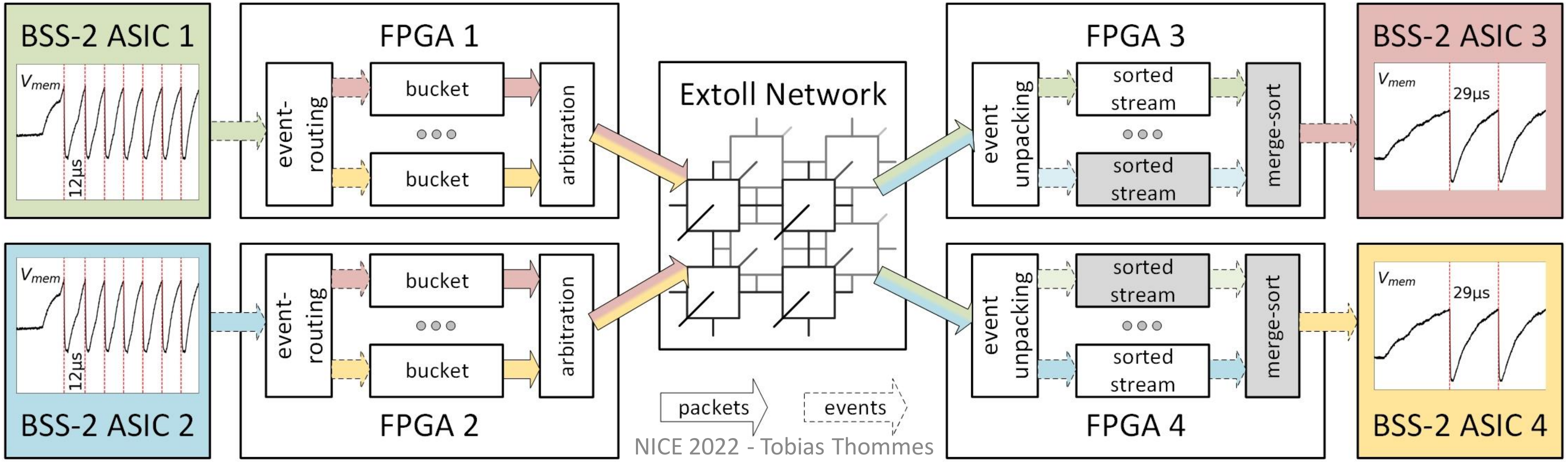
  → 3D-Torus with concentrator-nodes



Tobias Thommes, Niels Buwen, Andreas Grübl, Eric Müller, Ulrich Brüning, and Johannes Schemmel. 2021. BrainScaleS Large Scale Spike Communication using Extoll. arXiv:2111.15296 [cs.AR]

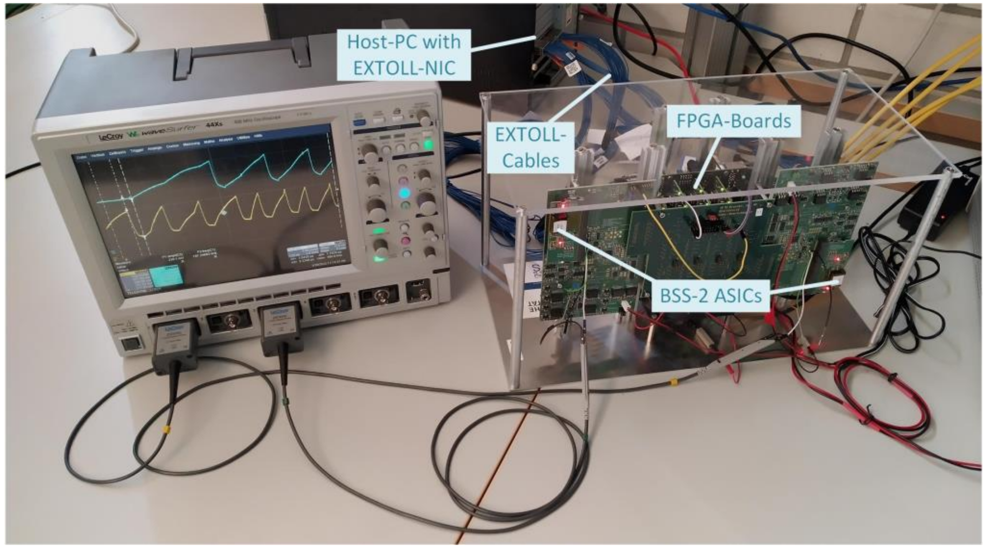
Pulse Event Communication

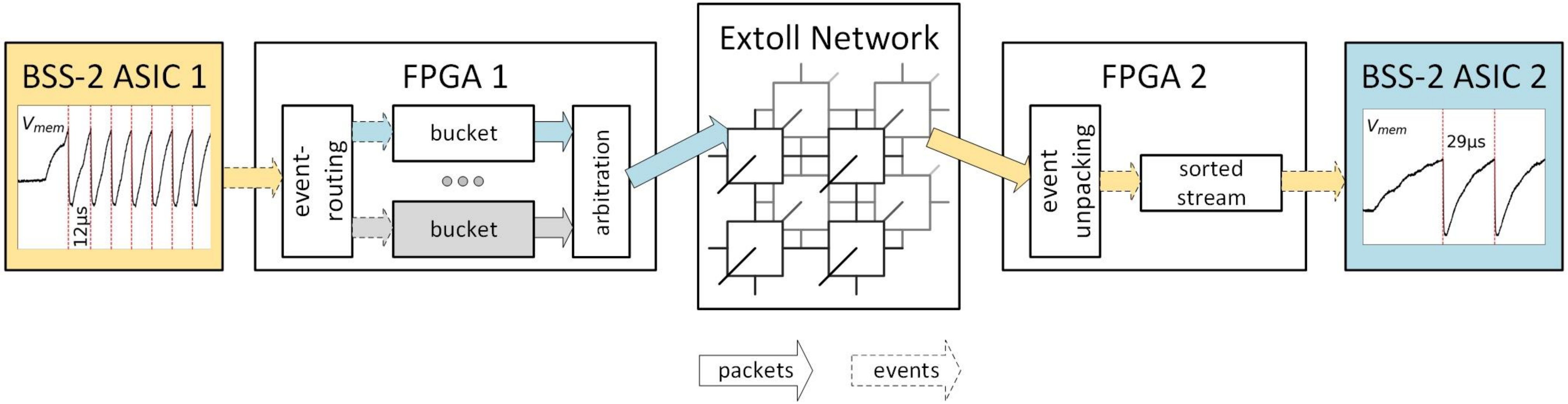
- Buckets *aggregate* events into larger packets at source
  - mitigate header-overhead
  - limit message rate
  - one destination per bucket
- Event-streams need to be *merged* at destination
  - event packets can arrive from different sources
  - streams are pre-sorted by timestamp

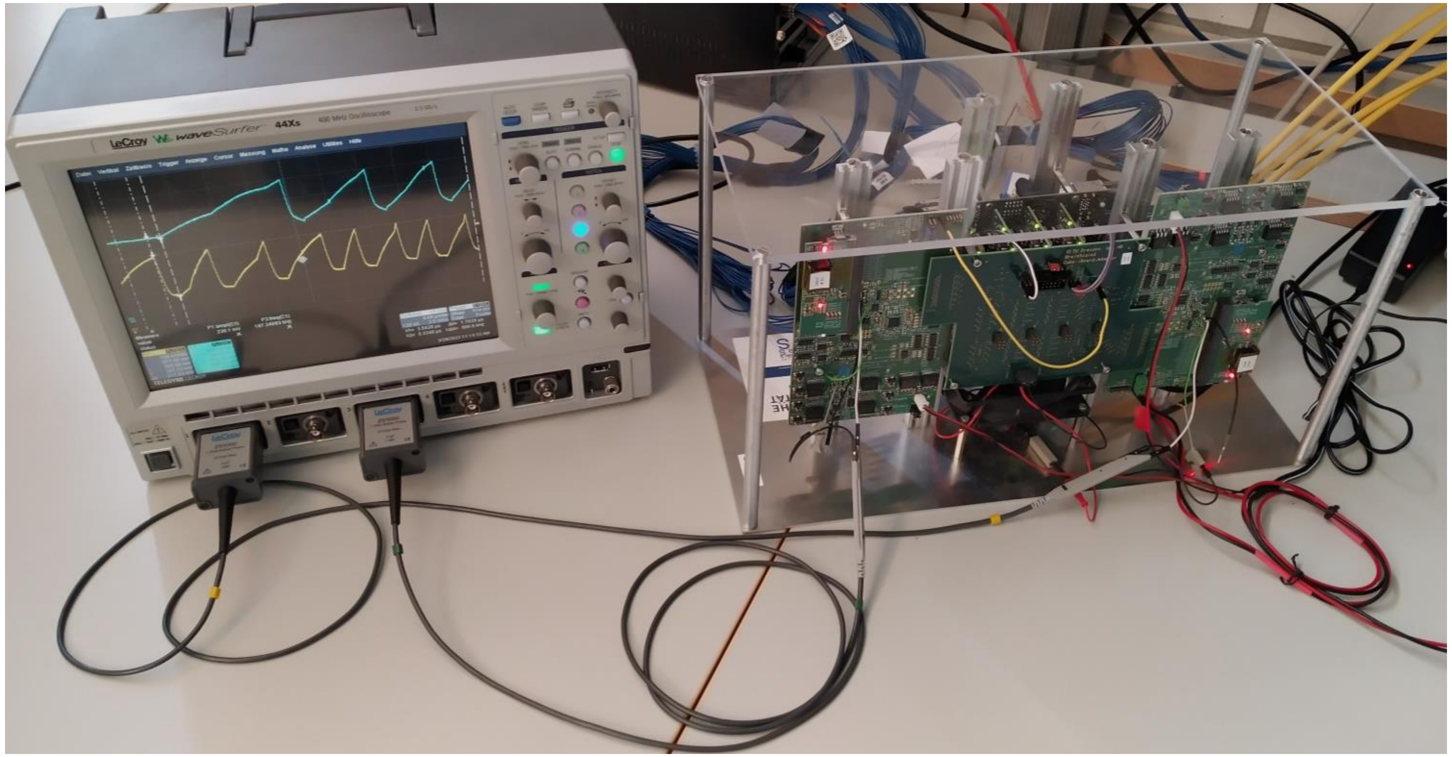


Experiment Setup

- Using simplified FPGA implementation
- Experiment: one source-chip sends events to one destination-chip
- merge-sort not needed yet
- Latency measured Neuron-Neuron:
  $$(1.6 - 2.3)\mu s$$
- Latency range due to interleaving of event traffic with host traffic at network interface







Live Demonstration

Summary:

- We are now able to basically support multi-chip experiments
- First experiments yield latencies of $(1.6 - 2.3)\mu\text{s}$ from neuron to neuron on two chips
- Network size is horizontally scalable

Next Steps:

- Synchronisation of experiment-execution across FPGAs
- Full support for multi-chip experiments in higher software layers
- Scaling up experiments for more than two chips

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Electronic Vision(s)

ziti

computer architecture

EXTOLL... latency matters.