SpiNNaker2 - Towards Extremely Efficient Digital Neuromorphics and Multi-scale Brain Emulation

Sebastian Höppner, Christian Mayr
Technische Universität Dresden, Germany

NICE 2018
Outline

- SpiNNaker Overview
- SpiNNaker2 Hardware
- Neuromorphic Applications
- Conclusion
SpiNNaker

• Communication and memory centric architecture for efficient real-time simulation of spiking neural networks [1]
• Many-core (ARM based) architecture, 18 cores per chip
• SpiNNaker has a broad user base
  • ~40 systems in use around the world
  • Flexibility: adaptable network, neuron model & plasticity
  • Real-time: suits robotics & faster than HPC
  • System capacity of $10^9$ neurons and $10^{12}$ synapses
  • Energy per synaptic event $10^{-8}J$ (HPC: $10^{-4}J$)
• SpiNNaker uses 130nm CMOS technology
• Scope for improvement
  • on modern process (22FDX) [2]
  • Innovative circuit techniques to enhance throughput and energy efficiency for computation and communication
• SpiNNaker2 target: Enhance capacity for brain size network simulation in real time at >10x better efficiency
SpiNNaker2 Hardware
HBP SpiNNaker2 Roadmap

**SpiNNaker2**
- 144 ARM M4F cores
- Power management
- SpiNNaker router
- Low swing serial I/O
- 4x LPDDR4 memory IF
- 8GByte LPDDR4 PoP
- 22nm CMOS

**SpiNNaker2 Simulation of complete human brain**

**JIB1**
- 8 ARM M4F cores
- SpiNNaker router,
- Low swing serial I/O
- 22nm CMOS

**JIB2**

**NanoLink28**
- SerDes Transceiver
- 28nm CMOS

**Santos28**
- 4 ARM M4F cores
- Power management
- SpiNNaker router with SerDes
- LPDDR2 Memory Interface
- 28nm CMOS

**Spinnaker 1:**
- 1% of human brain

**JIB1**

**2023**

**2022**

**2021**

**2020**

**2019**

**2018**

**2017**

**2016**

**2015**

**2014**

**2013**

**2020**

**2019**

**2018**

**2017**

**2016**

**2015**

**2014**

**2013**
SpiNNaker2 Chip Overview
Processing Element

Dynamic Power Management
• DVFS and PSO [3]

Memory sharing
• Synchronous access to neighbor PEs

Multiply-Accumulate accelerator
• MAC array with DMA

Neuromorphic accelerators
• Exp/log [4,7]
• Random numbers (PRNG, TRNG from ADPLL noise) [5]

Network-on-Chip
• On- and off-chip memory access
• SpiNNaker packet (spike) handling

Adaptive Body Biasing
Implementation Strategy

- GLOBALFOUNDRIES 22FDX (FDSOI) technology [2]
- Adaptive body biasing (ABB) solution and foundation IP by Dresden Spinoff Racyics [8] → Enables operation down to 0.40V (0.36V wc)
- Power performance area (PPA) studies for neuromorphic application scenario:

**Low-performance Level (PL1)**
- Operate at Minimum Energy Point (250MHz at 0.50V) or at ultra-low power mode (100MHz at 0.45V)

**High-performance Level (PL2)**
- Operate at 500MHz at 0.60V for maximum peak performance for neuromorphic simulations
Neuromorphic Power Management

- **Dynamic Voltage and Frequency Scaling**
- **Fine-grained** (individually per PE)
- **Fast** DVFS (<100ns) PL change time [6]
- **Self-DVFS** PL change from software based on neuromorphic workload
Neuromorphic Power Management - Example

- Synfire chain network with bursting behavior
- $\approx 90\%$ of simulation cycles are processed at lowest PL
  - $\rightarrow$ maximum energy efficiency
- System performance limit is reached at highest PL (only $\approx 2\%$ of simulation cycles)
  - $\rightarrow$ peak performance for biological real time achieved
- Up to $\approx 50\%$ PE power reduction, while still achieving peak performance for biological real time operation

Note: 28nm testchip Santos supports 3 PLs
**Integrated MAC Accelerator**

- 16x4 MAC array per PE
- Access local-SRAM and NoC
- Offloading matrix multiplication and convolution from the CPU
- Remote controlled operation possible

**Peak Performance @250MHz:**
- 0.032 TOPS/PE $\rightarrow$ 4.6TOPS on *SpiNNaker2* at $\approx 0.72$W PE power consumption $\rightarrow$ **6.4TOPS/W**
Interim Conclusion Hardware

- Energy efficient digital many core approach for neuromorphics
- Motivated by advantages of a mix of current approaches:
  - Processor based $\rightarrow$ flexibility
  - Fixed digital functionality as accelerators $\rightarrow$ performance
  - Low voltage (near threshold) operation enabled by 22FDX and ABB $\rightarrow$ energy efficiency
  - Event driven operation with fine-grained DVFS and energy proportional chip-2-chip links $\rightarrow$ workload adaptivity
- Integrate a SpiNNaker 1 48 node board inside a single chip module
Neuromorphic Applications
# Neuromorphic Applications: Overview

<table>
<thead>
<tr>
<th>Benchmark, Application</th>
<th>PM</th>
<th>EXP</th>
<th>PRNG/TRNG</th>
<th>Float</th>
<th>Benefit of new features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synfire Chain</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-vitro-like Bursting Network</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>85% power reduction</td>
</tr>
<tr>
<td>Asynchronous Irregular States</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reward-Based Synaptic Sampling (with TU Graz)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>&gt; 2x performance</td>
</tr>
<tr>
<td>BCPNN networks (with KTH Stockholm)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Deep Rewiring (with TU Graz)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Spike detection and sorting (realtime biological data processing)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic Vision Sensor Interface, WTA network</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>... evaluation ongoing</td>
</tr>
</tbody>
</table>

Note: All results from first SpiNNaker 2 prototype: Santos chip in 28nm
Reward-Based Synaptic Sampling

- Characteristics [9]:
  - Spiking reward-based learning
  - Synaptic sampling of network configuration
- Benchmarks:
  - Current: Double-T maze, task-dependent routing
  - Future: Pong player
- Task-dependent routing characteristics:
  - 200 input neurons, 20 stochastic neurons
  - 8k stochastic synapses
- Challenge:
  - Reward gradient needs float
  - Exp and random transformation
- Uses random, float&exp, speed-up factor 2
Deep Rewiring

- Synaptic sampling as dynamic rewiring for rate-based neurons (ML networks) [10]
- 96.2% MNIST accuracy for 1.3% connectivity
- Ultra-low memory footprint even during learning
- Uses random, float&exp, speed-up factor 1.5
- Improved fail-soft in comparison to pruning
- Current efforts:
  - Parallelization
  - Low resolution weights
  - ML with power management (exploit spatial and temporal sparseness)

LeNet 300-100

In MNIST 784
Hidden FC 300
Hidden FC 100
Out Softmax 10

![Accuracy vs. Connectivity](chart.png)

- DEEP R
- soft-DEEP R
- $l_1$-shrinkage
- pruning (Han et al. 2015)
- fully connected
- connectivity vs. iteration is shown
Outlook: Multi-Scale Modeling

- Molecular dynamics/Ion channels with random generator, log/exp function accelerator (extend to other functions?)
- Multi-compartment or point neurons based on native spiking network support
- Rate-based neuron with MAC accelerator
- Neural field model via ML-network plus random generator (e.g. mean field equation with added population noise)

Towards a full human brain model sustainable on SpiNNaker 2

1Image source: wikipedia
Conclusion

• **Spinnaker 2 Deployment Approaches**
  - **Small-scale/embedded** for robotics: 48 node board-on-chip with standard interfaces and flexible I/O
  - **Large-scale** full 10 Mio core machine
    - 5PetaFLOPS CPU, 0.6 ExaOPS MAC accelerators
    - → Energy per synaptic update: spike-based 300pJ, rate-based 300fJ

• **Applications**
  - Multi-scale modelling, flexibility by software, wild combinations possible
  - Software stack: SpiNNaker-style for spiking, Tensorflow/Caffe backend for ML, how to merge

• **Outlook: Derived concepts**
  - Tactile internet: Local smart sensor/actor nodes
  - Automotive: ML processing for radar, lidar, visual
  - Closed-loop neural implant: spike sorting, neuromorphic&ML processing
Acknowledgment

The *SpiNNaker2* team

Technische Universität Dresden
Sebastian Höppner, Andreas Dixius, Stefan Scholze, Marco Stolba, Thomas Hocker, Stefan Hänzsche, Florian Kelber, Dennis Walter, Johannes Partzsch, Bernhard Vogginger, Johannes Uhlig, Georg Ellguth, Chen Liu, Ali Zeinolabedin, Yexin Yan, Stefan Schiefer, Stephan Hartmann, Love Cederstroem, Stephan Henker, Felix Neumärker, Sani Md Ismail, Christian Mayr

University of Manchester
Delong Shang, Gengting Liu, Dongwei Hu, Jim Garside, Mantas Mikaitis, Andrew Rowley, Luis Plana, Dave Lester, Simon Davidson, Steve Temple, Steve Furber

Thanks to ARM and Racyics and GLOBALFOUNDRIES
References


[8] www.makeChip.design


Thanks for your attention