Neuromorphic Computing and Engineering

PAPER • OPEN ACCESS

From clean room to machine room: commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system

To cite this article: Hartmut Schmidt *et al* 2023 *Neuromorph. Comput. Eng.* **3** 034013

NEUROMORPHIC
Computing and Engineering

OPEN ACCESS

RECEIVED
27 April 2023

REVISED
10 August 2023

ACCEPTED FOR PUBLICATION
8 September 2023

PUBLISHED
25 September 2023

PAPER

From clean room to machine room: commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system

Hartmut Schmidt[1,3], José Montes[1,3], Andreas Grübl[1], Maurice Güttler[1], Dan Husmann[1], Joscha Ilmberger[1], Jakob Kaiser[1], Christian Mauch[1], Eric Müller[1], Lars Sterzenbach[1], Johannes Schemmel[1,*] and Sebastian Schmitt[2]

[1] Kirchhoff-Institute for Physics, Heidelberg University, Heidelberg, Germany
[2] Department for Neuro- and Sensory Physiology, University Medical Center Göttingen, Göttingen, Germany
[3] Contributed equally.
[*] Author to whom any correspondence should be addressed.

E-mail: schemmel@kip.uni-heidelberg.de

Keywords: neuromorphic hardware, wafer-scale integration, spiking neural networks, emulated networks, analog neuromorphic devices, synfire chains

Supplementary material for this article is available online

Abstract

The first-generation of BrainScaleS, also referred to as BrainScaleS-1, is a neuromorphic system for emulating large-scale networks of spiking neurons. Following a 'physical modeling' principle, its VLSI circuits are designed to emulate the dynamics of biological examples: analog circuits implement neurons and synapses with time constants that arise from their electronic components' intrinsic properties. It operates in continuous time, with dynamics typically matching an acceleration factor of 10 000 compared to the biological regime. A fault-tolerant design allows it to achieve wafer-scale integration despite unavoidable analog variability and component failures. In this paper, we present the commissioning process of a BrainScaleS-1 wafer module, providing a short description of the system's physical components, illustrating the steps taken during its assembly and the measures taken to operate it. Furthermore, we reflect on the system's development process and the lessons learned to conclude with a demonstration of its functionality by emulating a wafer-scale synchronous firing chain, the largest spiking network emulation ran with analog components and individual synapses to date.

1. Introduction

Simulating the dynamic properties of large-scale spiking neural networks is challenging due to the massively parallel interactions of their neurons and synapses. Neuromorphic architectures propose a solution to this dilemma by providing inherently parallel computation at nodes operating as neurons and synapses and communicating through action potentials, also termed spikes [1–3]. Different neuromorphic hardware systems have been suggested and created, each distinguishing itself based on its architecture, scalability, learning capabilities, and whether it adheres to an analog/mixed-signal or purely digital methodology [4–9].

The BrainScaleS neuromorphic architecture implements physical models of neurons and synapses on a CMOS substrate with analog circuits, while the spike communication is digital. On the one hand, the physical models inherently provide solutions to neuron and synapse dynamics in continuous time, in contrast to the time-discretized and numerically integrated solutions of digital systems and software simulations. On the other hand, the programmable digital communication of action potentials allows for flexible network topologies and the possibility of using digital logic to feed and read spike events from outside the system. Furthermore, circuits are operated in strong inversion, targeting dynamics with a typical speedup factor of 10 000 compared to biological real-time.

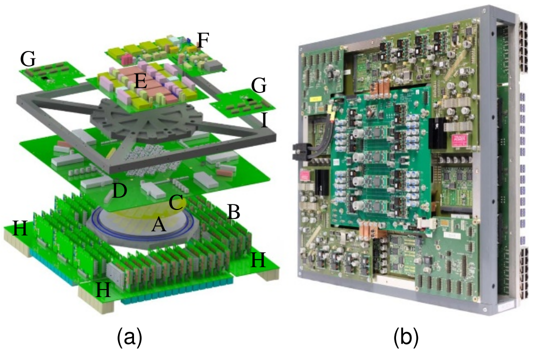

Figure 1. (a) 3D-schematic of a BrainScaleS-1 wafer module (dimensions: 50 cm × 50 cm × 15 cm) hosting the wafer (A) and 48 communication boards (B). The positioning mask (C) aligns elastomeric connectors that link the wafer to the large main PCB (D). Support PCBs provide power supply (E & F) for the on-wafer circuits as well as access (G) to analog dynamic variables such as neuron membrane voltages. The connectors for inter-wafer and off-wafer/host connectivity (48 × Gigabit-Ethernet) are distributed over all four edges (H) of the main PCB. Mechanical stability is provided by an aluminum frame (I) © [2017] IEEE. Reprinted, with permission, from [14]. (b) Photograph of a fully assembled wafer module. © [2017] IEEE. Reprinted, with permission, from [14].

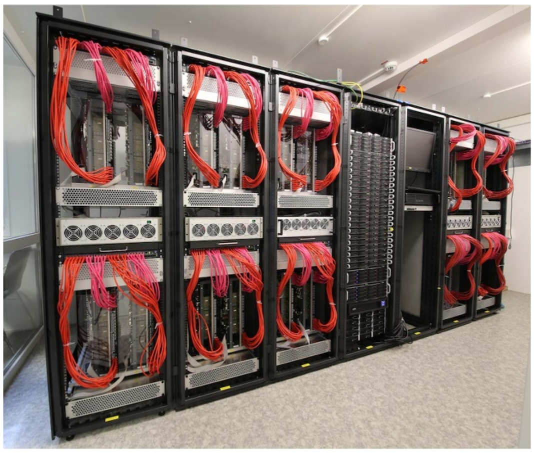

Figure 2. The BrainScaleS-1 machine room comprising 20 wafer modules organized in 5 racks. A slot in the middle of each rack hosts the analog readout module and the main control units of its neighboring wafer modules. Gigabit-Ethernet cables connect each wafer module via aggregation switches to the control cluster positioned in the middle rack. © [2017] IEEE. Reprinted, with permission, from [14].

The BrainScaleS-1 system utilizes wafer-scale integration to achieve large application-specific integrated circuit (ASIC) counts with energy efficiency and high communication bandwidth. The structure of its underlying neuromorphic chip and the technology to achieve its wafer-scale integration are introduced in [10–13]. Turning the silicon wafer into a ready-to-use system, though, implicates bringing several additional components, shown in figure 1, to work hand in hand. For that cause, a commissioning chain is established, which is this paper's focus.

The paper is divided into three sections. In the first part, we illustrate the different components that constitute the system and how they are tested. We show the steps to assemble the module before it is finally placed in the machine room, as shown in figure 2. In the second part of the paper, we describe the methods devised to bring such a system into a reliable substrate for neuromorphic experiments: a large number of very large-scale integration (VLSI) analog components inevitably leads to malfunctioning parts and analog variability, for which an underlying fault-tolerant design and suitable handling have to be put in place. In the third part, we finally demonstrate its operation and the successful implementation of these measures by emulating a biologically-motivated network of spiking neurons, a synchronous firing chain, on a fully commissioned BrainScaleS-1 wafer module.

The system belongs to the still-nascent field of neuromorphic computing and remains under continuous development. Having pioneered a neuromorphic wafer-scale integration of VLSI analog and digital circuits, we also discuss the lessons learned while solving or circumventing the challenges faced along the way.

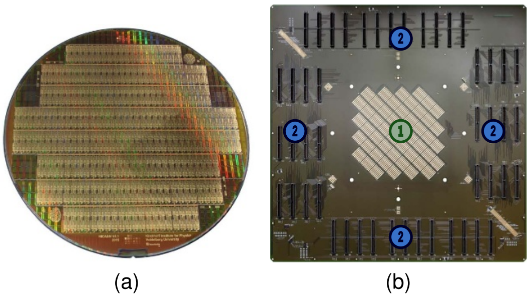

Figure 3. (a) The BrainScaleS-1 wafer with applied postprocessing to achieve wafer-scale integration and to establish its connection to the (b) bottom side of the main PCB. There, the wafer connects through elastometic connectors to the center, marked with 1. Along the edges, 48 connectors, marked with 2, accommodate the communication boards.

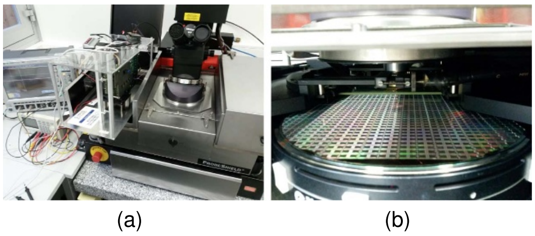

Figure 4. (a) Photograph of the wafer prober and (b) a close-up of a wafer under test. Different needle cards have been developed and used for tests carried out before wafer post-processing (section 2.1, visible in this setup) and before wafer module assembly (section 3.2), respectively.

2. System components and individual tests

A BrainScaleS-1 wafer module is depicted in figure 1. Each of its constituent boards is individually tested before its integration into the system, which permits differentiating errors in the parts from those arising from the assembly. A short description of each component and the tests it undergoes is given in the following.

2.1. The BrainScaleS-1 wafer

The heart of each module is an uncut 20 cm wafer, displayed in figure 3(a), fabricated in UMC 180 nm technology comprising 384 high input count analog neural network (HICANN) ASICs. Each HICANN contains 512 analog neuron circuits implementing the adaptive exponential integrate-and-fire model [13, 15]. Single neuron circuits receive input from up to 220 analog synapses. Since neuron membranes can interconnect in groups of up to 64, a maximum of 14 080 synapses can provide input to each of these composite neurons. Synapse weights are stored with 4-bit resolution in local SRAM at each synapse. To set, e.g. analog neuron model parameters and other on-chip bias voltages and currents, the HICANN stores 12 384 analog quantities in single-poly floating gate (FG) CMOS cells that retain their operation levels in accordance to their isolated gate's accumulated charge [16, 17].

Wafer-wide communication is achieved with a custom-developed redistribution layer applied post-wafer-production, creating around 160 000 lateral connections across chip boundaries [18]. These connections provide the modules with on-wafer spike event communication through low-voltage differential signaling (LVDS) buses utilizing an asynchronous serial event transmission protocol. Furthermore, connections through top-layer pads on the wafer provide the modules with parallel per HICANN off-wafer communication. In conjunction with a high number of instances of each component, which can be deactivated independently, this constitutes the system's fault tolerant design [10].

Testing: in order to assess the effect of wafer post-processing on the digital yield of an entire wafer, initial needle card tests were carried out on two unprocessed[4] wafers to determine their yield immediately after production. Since the wafers undergoing these tests cannot be further processed, comparing results on the same wafers before and after the post-processing is, however, not possible.

The setup for these tests in the institute's clean room is shown in figure 4, and the procedure is as follows. The needle card is used to contact each individual ASIC. Immediately after contacting and powering up, the

[4] 'unprocessed' in this context means untested wafers straight from the manufacturer, before the custom redistribution layers have been added.

total current on the used lab supply is measured to detect potential power shorts. Henceforward, all digital memory cells on the HICANN circuits are tested using a built-in joint test action group (JTAG) access mode. During these tests, 448 HICANNs on each of the two wafers were tested, and 93% of them showed no single digital error. To compare, UMC's calculator estimates a yield of approximately 85% by taking into account the process parameters and circuit size. However, our results are only an estimation: on the one hand, the tested digital memory cells only cover a fraction of the whole silicon area, which is dominated by analog circuitry. Therefore, the digital test yield could be assumed to be too optimistic. On the other hand, perfect power and signal integrity could not be ensured while connecting the circuits through the needles, leading to a possible detection of false negatives, caused for example by slightly underpowered memory cells. In addition, only wafers from the initial engineering sample production have been available for testing. No documentation has been available to relate the production yield data from UMC to small batch-size engineering runs. Nonetheless, the results match the expectations taking the high level of uncertainty into account. Also, a yield in the order of, e.g. 85% would not indicate that 15% of the dies cannot be used. Instead, advantaging from the fault-tolerant design, and depending on the defect type, it could suffice to disable single neuron or synapse circuits, for example, on affected HICANNs that are otherwise fully functional and can remain available for experiments.

2.2. Main PCB

The main printed circuit board (PCB), displayed in figure 3(b), is a 43 cm × 43 cm passive interconnector board for most parts of the wafer-scale integration system. Seven of its 14 layers are used to distribute 23 power rails carrying up to 200 A of current. The rest of the layers are used to route 1152 power monitoring, 1472 high-speed differential communication, and different sideband signals. Auxiliary boards, communication infrastructure, and the silicon wafer are connected via various kinds of detachable connectors. These enable system modularity for development and upgrades, desirable for research and development in dynamic environments over longer timespans.

Testing: the manufacturer[5] performs complete optical inspection and electrical tests of the main PCB. The BrainScaleS-1 wafer modules are assembled using exclusively fully validated, error-free main PCBs.

2.3. Auxiliary boards

The wafer module is completed by populating it with 48 communication boards and auxiliary boards for power delivery, control, monitoring, and inter-module communication.

2.3.1. Communication boards

Each communication board[6] contains a Kintex7-XC7K160T field-programmable gate array (FPGA) and connects to one HICANN group consisting of eight HICANNs. These boards communicate through separate high-speed LVDS interfaces with each of the connected HICANNs to configure, monitor, and coordinate the experiment runs; they feed and collect generated spikes into/from the experiments. Furthermore, they synchronize the start of experiments to allow for wafer wide execution. Trigger signals generated on these boards also align experiments with analog recordings using the analog readout module (AnaRM).

Testing: the communication boards are tested on a standalone setup that implements loopback connections for the high-speed interfaces. For this purpose, a test board accommodates and tests four PCBs in parallel, as shown in figure 5(a). Primarily automated and controlled via software, the tests switch the power supply via general purpose interface bus. Programming is performed via JTAG and power management bus. Tests comprising current consumption measurements, loading and communicating with the FPGA design, as well as memory tests are conducted. In addition, communication with the host computer as well as the links to the wafer and neighboring communication boards are tested. As per data logs, only 18 out of 1404 produced PCBs had to be discarded after failed tests.

Improvements: in a previous version of the system, a Virtex5-based FPGA board, together with four digital network chips (DNCs) containing additional pulse routing logic, were combined into a communication subgroup [19–21]. Twelve of these subgroups were connected with four fine-pitch connectors each to a predecessor version of the main PCB. This communication subgroup had heat-pipe-based thermal management, and the DNCs were mounted on mezzanine boards in between the subgroup assembly and the main PCB. The challenging mechanical alignment of the four fine-pitch connectors combined with the weight of the subgroup assembly, caused failing connections on the fine-pitch connectors, also compromising JTAG signals which are required for system initialization and control. None of the communication subgroups could be connected without at least one failing connection, requiring a re-design of this part of the system. In

[5] Manufactured by Würth Elektronik, Germany.

[6] Developed at the chair of Hochparallele VLSI-Systeme und Neuromikroelektronik at TU Dresden.

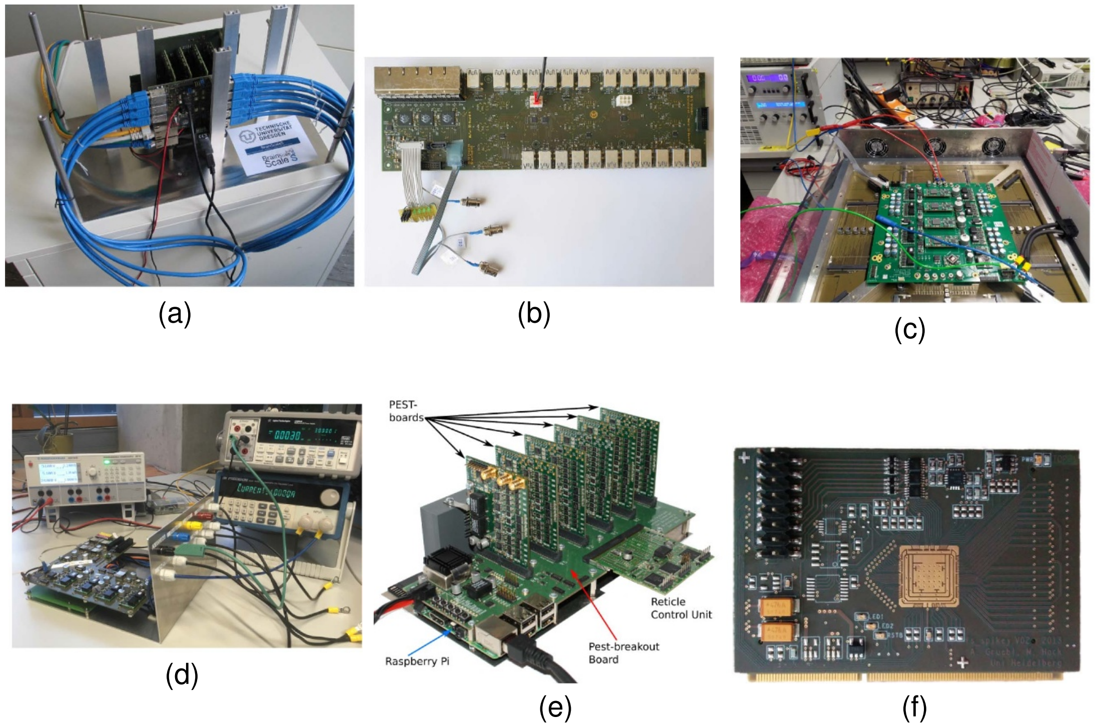

Figure 5. Auxiliary boards under test. (a) Communication boards test setup and (b) Wafer I/O PCB board. (c) Main power supply connected to programmable power supply and electronic load. (d) Auxiliary power supply PCB test stand. (e) Control unit for reticles test stand. Each power emulation systems for testing (PEST) board emulates the supply voltages of one HICANN group. (f) FPGA board of the analog readout module. During the calibration, the pins on the top left are connected via a 50 Ω impedance to an external source meter, while the module is connected via USB to the host computer. Figures (a) and (b) made available by S Schiefer, TU-Dresden.

the presented system, all digital logic could be merged into the more modern Kintex7 FPGA, eliminating the need for individual DNCs. Furthermore, connectivity of individual FPGA boards is established through board-edge connections with larger pitch and much more tolerance to slight mechanical stress. This assembly, combining main PCB ⇔ communication board ⇔ and Wafer I/O PCBs on top (see following section), improved the connection reliability, with currently no failing connections in this part of the system.

2.3.2. Wafer I/O PCB

Each one of the module's four Wafer I/O PCBs (WIOs)[6] attaches to twelve communication boards, aggregating Gbit-Ethernet and connections to other communication boards.

Testing: a manual approach is followed as the number of boards is smaller than that of the communication boards. The board, shown in figure 5(b), is supplied with power, and the proper functioning of the DC/DC converters is checked with a multimeter. Individual communication ports are tested. In addition, the proper transmission of signals using a signal generator and differential probes is measured. A partial test of the JTAG pins is also carried out. As per data logs, only 2 out of 120 produced WIOs were discarded after failed tests.

2.3.3. Main power supply

The main power supply (PowerIt) has three output channels: two 1.8 V outputs as main analog and digital supplies of the wafer with a current limit of 200 A each, as well as a 9.6 V output capable of up to 110 A to supply the communication boards. Multiple custom-milled copper parts ensure a low-resistance screw connection between the PowerIt and the main PCB. Additionally, digital control of the voltages and sensors as near to the wafer as possible allow for compensation of IR-drop. An integrated microcontroller can measure input and output currents and voltages via shunt resistors, hall sensors, and voltage dividers.

Testing: commissioning of the PowerIt involves basic functionality tests and calibration of the current and voltage measurement circuits using an external electronic load capable of sinking 4.8 kW and precision multimeters, see figure 5(c).

2.3.4. Auxiliary power supply

The auxiliary power supply PCB (AuxPwr) designed in [22], receives 9.6 V from the PowerIt and provides ten different voltage outputs for the wafer module. The currents drawn at the derived voltages vary from

50 mA, for the common-mode voltage of the LVDS on-wafer communication, to 60 A for the synapse driver output. The board has an L-shape with linear and switching regulators placed on different axes to reduce the coils' electromagnetic-noise induction. In addition, the usage of intermediate voltages reduces the power dissipation for the voltage scaling. An onboard microcontroller monitors all the voltages and currents. Four voltages can be controlled digitally through the inter-integrated circuit (I^2C) protocol.

Testing: the AuxPwr components' functionality is tested during the calibration process of the board, during which an external voltmeter permits adjusting voltage offsets. A two-point linear calibration under load is performed for the currents. The test stand can be seen in figure 5(d).

2.3.5. Control unit for reticles

Since the BrainScaleS-1 wafer is not cut into individual chips, the wafer module must be fault-tolerant to individual HICANN problems. For this purpose, the main PCB features power-FETs for the supply rails of each HICANN group of the wafer; overcurrents manifest as a large voltage drop across these power transistors. The control unit for reticles (CURe) controls the gates of these transistors and monitors the supply voltages of the wafer. Three microcontrollers manage the measured data and react to fault conditions by shutting off the power of the affected HICANN groups. Thus, the CURe allows to identify individual fatal faults and to exclude the respective HICANN groups from the usable components. The term reticle refers to the semiconductor manufacturing process and consists of one HICANN group.

Testing: the CURe is tested using a custom setup producing the voltages expected inside the actual BrainScaleS-1 wafer module, simulating all possible fault conditions while the response time is measured. Likewise, the drive strength of the control signals for the power transistors on the main PCB is quantified. The test setup is displayed in figure 5(e).

2.3.6. Analog readout module

Further insight into the neuron dynamics can be obtained via measurements of its membrane potential, allowing for a better understanding of experiment results and the implementation of calibration routines. To this end, each neuron contains a switchable analog output amplifier that connects to one of two 50 Ω output buffers per die. These two outputs are each short-circuited across dies in the same HICANN group. Therefore, each of these groups has two analog outputs, totaling 96 independent analog channels available on each wafer module.

The AnaRM system consists of twelve FPGA-controlled 12-bit ADC modules that allow for the digitization of the membrane voltages on one wafer module per BrainScaleS-1 system rack. Each of the modules in the AnaRM system connects through a ribbon cable to one of two analog breakout PCBs mounted on the main PCB, receiving eight analog signals that are multiplexed into the ADC. An additional digital signal acts as a trigger; four HICANN groups share one, allowing synchronization during an experiment between the involved communication boards, HICANNs and the AnaRM system. Overall, the AnaRM system can simultaneously sample 12 membrane traces per wafer module.

Testing: the FPGA board in the AnaRM, displayed in figure 5(f), undergoes DRAM memory tests and basic functional testing of all its peripheral components. The analog front end is tested during the calibration of the modules. This calibration is performed using a source meter to generate a series of ground-truth voltages, which are subsequently measured using each input channel. A 50 Ω series impedance is used at the output of the source meter to match the impedance of the output buffers on the HICANN. This voltage divider formed by the output and input impedances halves the 1.8 V span of the HICANN output to the 0.9 V maximum input of the AnaRM. A linear function fits the recorded signal to the source meter voltages, and the per board offset and gains are stored in a database.

2.4. Main control unit

The main control unit (MaCU) consists of a Raspberry Pi powered by the standby voltage of the PowerIt. Using the I^2C protocol to communicate with all other wafer module components, it controls the start-up sequence of the system. Additionally, it monitors the multitude of components of a wafer module, which is crucial to ensure robust operation. With this in mind, the MaCU aggregates over 1800 metrics per wafer, e.g. supply voltages, temperatures, or the active/inactive status of components. Most data is of a time-series nature and stored via Graphite [23], with visualization through Grafana dashboards [24]. These dashboards are hierarchically structured, allowing an intuitive drill-down navigation of the data. As it is not practical to manually oversee such a large amount of metrics, alerts are set up to check for unexpected events. For example, supply voltages are checked to be in a valid range and to remain constant over time. Furthermore, event data, e.g. powering up components, is handled via the ELK stack [25] but also integrated into Grafana and displayed as marks. These allow easily matching the events with changes in the time-series data.

Testing: the Raspberry Pi computers used for the MaCUs are purchased and commissioned without further tests. However, the maintenance and deployment of the control and monitoring software they run is part of the system's continuous integration development methodology [26].

3. System assembly and integration tests

In addition to the tests devised for the individual components, the BrainScaleS-1 wafer module assembly process is carried out along with additional tests that allow pinpointing problems to the individual steps. In the following, we discuss the module assembly method and the different tests it undergoes during this phase.

3.1. Wafer to main PCB marriage and module integration

The wafer is connected to a total of 11 904 pads on the main PCB via 384 elastomeric connectors, shown in figure 6(a). Mounting the main PCB and the silicon wafer in custom-milled aluminum brackets allows reaching the compression forces required by the connectors. The station used to align the two components is shown in figure 6(b). There, the wafer is placed face up into the wafer bracket that is mounted to the center of the alignment station. Subsequently, the elastomeric mask is attached to the bottom of the main PCB, which is kept floating 1 cm above the wafer fixed by the springs of the station. The position of the main PCB is then adjusted via micrometer screws. During the process, fiducial marks on the wafer are aligned with two wires forming a cross on the main PCB. Fitting accuracy is checked using a digital camera in combination with a macro lens that allows to optically inspect the alignment of the markers through the top cover of the main PCB. Initially, additional structures were planned on wafer and main PCB to verify the alignment electrically to speed up the whole process. However, since the most time consuming part is the adjustment of the screws these structures were found to be not beneficial and therefore omitted.

Once the markers are aligned, the screws connecting the wafer and main PCB get tightened. Electrical resistance tests, described in section 3.2.2, are performed while compressing the elastomeric connectors to ensure correct positioning and even pressure distribution. Only if all HICANNs pass these tests the marriage is complete. Else, the whole process is repeated.

Afterward, the wafer module is populated with the auxiliary boards and, when fully assembled, connected to the MaCU. Then, it is put on a test stand for initial full-system tests using the same communication chain later used for experiments. Following this step, the wafer module is placed in a rack in the machine room and attached to the AnaRM system.

3.2. Tests at different assembly stages

Stage-specific tests allow mapping arising errors to individual assembly steps of the BrainScaleS-1 wafer module, which enables evaluating and improving the procedure. This section shows the test results obtained for one wafer as an example.

3.2.1. Pre-assembly tests of all HICANNs on the wafer

Before placing a wafer in a module, digital and analog tests are performed on a wafer prober in the institute's clean room, see figure 4. These tests distinguish production problems from those arising in the wafer module assembly procedure.

Similar to the initial needle card tests on the unprocessed wafers, described in section 2.1, a test system was built using a different needle card connecting to the redistribution layer of a pair of HICANNs on a wafer with post-processing. Extended analog and digital tests are run on the connected dies, a process that is repeated until the entire wafer is analyzed. These tests serve two purposes: first, to sort out wafers with a high error count that might arise from disrupted connections in the post-processing, and second, to establish a base level for the following assembly tests. Figure 7(a) shows the results of a high-level test for all HICANNs of one wafer. The image shows more test results than the number of dies on the picture of the assembled wafer module. The reason for this was design constraints and limited routing resources on the main PCB, by which not all HICANN groups could be electrically connected and thus used within the module context; those at the edge of the wafer were left out. For the same reason, the two HICANN groups at the center are without high-speed connection.

3.2.2. Tests during the assembly phase

For these additional tests the main PCB is equipped with test PCBs[7], shown in figure 6(c), which measure ESD diode currents and termination resistances between the LVDS lines on the wafer. The tests determine whether a good connection of the wafer to the main PCB exists. Figure 7(b) shows the result of one of these

[7] Developed by the group of Yasar Gürbüz at Sabanci University, Istanbul.

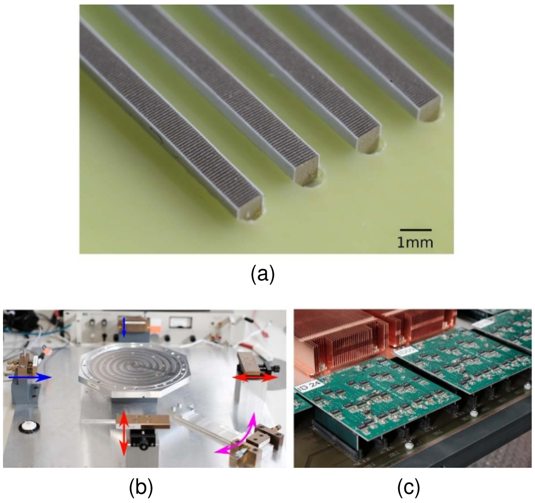

Figure 6. (a) Detail view of the elastomeric connectors that connect the pads on the BrainScaleS-1 wafer with the main PCB. (b) Station used to align the main PCB to the silicon wafer. The main PCB is fixed by springs that apply a constant force (blue arrows). Its position is controlled with a micrometer linear stage (red arrows). Angular errors can be corrected by rotating the wafer (purple arrows). (c) Test PCBs mounted on the main PCB to measure the connectivity to the wafer during assembly.

tests, where only the same faulty device on HICANN group 29, also detected in the needle card test, can be seen. No additional faulty devices validate that the wafer to main PCB marriage was appropriate.

3.2.3. Post-assembly tests of all HICANNs on the wafer

After the assembly of the wafer module is completed, the same tests run on the pre-assembly phase are conducted, and results are compared. The results for one test are shown in figure 7(c). The errors in HICANN groups 15 and 29 are still present, while the errors in groups 36 and 42 are not. Further investigations could trace these last errors to connection problems of the needle card used in the wafer prober.

4. Commissioning software

After assembly, additional steps are necessary to bring the BrainScaleS-1 wafer module into readiness for experiments. These include digital tests to find and exclude malfunctioning components and calibrating the individual neurons to address manufacturing-process-induced circuit mismatches. Databases store the results from these two steps, allowing serialized data storage to disk. See [26] for details. Furthermore, all steps are fully automated and periodically executed after installation of the module in the machine room to track the systems' current state.

4.1. Communication tests

The first test that is executed on a newly assembled wafer module is the communication test, which is used to find unresponsive HICANNs. Communication problems most likely arise from insufficient connection quality between the main PCB and the wafer, see [18], or from scratches or similar defects on the post-processing layers.

During the test, an individual connection is established to each of the 384 HICANNs of one wafer. The test is split into a high-speed test and a JTAG test, which reflects the two possibilities to communicate with the HICANN. Failures are stored separately in the availability database. The result of a communication test is shown in figure 7(d). In this example, the result comparison between the test stand and the rack-mounted fully assembled wafer module shows one additional HICANN group and three individual HICANNs that cannot communicate via JTAG.

4.2. Memory tests

Using a whole uncut wafer, each BrainScaleS-1 wafer module profits from better energy efficiency and higher bandwidth for communication between its ASICs as if these were produced separately and then integrated. This approach presents a challenge, though, as producing an error-free wafer-scale system in such a way is not possible, as ASICs with manufacturing-induced problems cannot be removed. The BrainScaleS-1 system addresses this through a digital memory test. In conjunction with a high number of instances of each component making up redundant functional units that operate independently from each other, which is part of the fault-tolerant system design, this enables dynamic handling of malfunctioning components. Executed

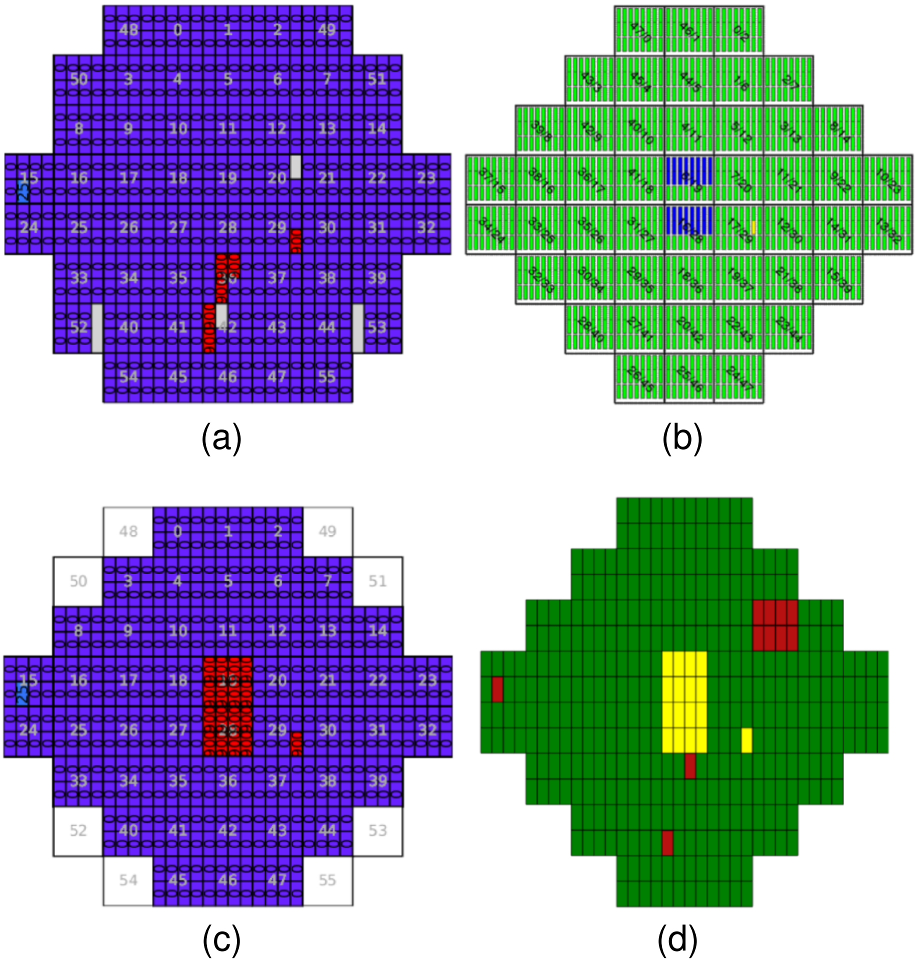

Figure 7. Test results of one BrainScaleS-1 wafer for the different assembly steps: (a) before assembly (b) during assembly (c) after assembly. In (a) and (c), the number in the smallest rectangles shows the amount of errors found on the corresponding HICANN. Purple or red indicate that all tests were successful or failed, respectively. For grey HICANNs the test was skipped since no connection could be established using the wafer prober. In (b), test results are shown per elastomeric connector and a yellow rectangle indicates a problem in the high-speed communication of one HICANN. (d) Communication test result. HICANNs without high-speed communication are marked yellow, without JTAG communication red. The center two HICANN groups have no high-speed interface by design. Consequently, they are marked faulty in all tests requiring high-speed communication to the main PCB.

after assembly as well as periodically, the test also tracks the state of the systems over time. Therefore, it allows to operate wafer modules despite a subset of malfunctioning components or connections, consequently increasing the yield of functional systems.

The test builds upon the communication test and establishes a connection to a HICANN group. First, it initializes the connected communication board and the HICANN under test. Subsequently, each digital memory is repeatedly write/read-tested using random values. If a mismatch is found, all components that cannot be used without the tested one must be excluded from the availability database. Here, the hierarchical structure of the system ensures the existence of a functional unit that contains all these components but does not include usable components. Exclusively excluding this functional unit a sparse resource representation is found, which ensures that no malfunctioning or dependent component is utilized in experiments. For example, if due to a malfunctioning register an individual synapse cannot be programmed to listen to a defined stimulus, all synapses sharing the same signal path cannot be used without said synapse injecting wrong signals into its neuron. As the signal for all affected synapses is generated in the same synapse driver, excluding the functional unit consisting of this synapse driver is sufficient. Thereby, the driver's attached non-usable synapses, both functioning and malfunctioning, are effectively excluded.

HICANNs that can communicate only via JTAG are exclusively used for spike route-through to and from neighboring HICANNs on the same wafer. For these, a routing-specific reduced memory test minimizes the runtime using the slower connection. In total, more than 42 MiB of digital memory get tested per wafer. Results for a fully assembled wafer module are shown in table 1. Tested components and their position on the HICANN are visualized in figure 8.

Table 1. Overview of excluded components extracted from a fully assembled BrainScaleS-1 wafer module. 'Components' shows the number of components taken into account for the tests and the effective exclusion. If two numbers are given, the first one is the number of tested components and the second one is the number of components evaluated for the effective exclusion. 'Individual' lists the communication and memory test results. Buses are marked with '-' because they have no digital memory that could be tested. 'Effective' shows the results of the effective exclusion of components. Here, all components that should not be used during an experiment are included. They not necessarily failed a test.

		Excluded	
Resource	Components	Individual	Effective
JTAG comm.	384	2.86%	3.39%
High-speed comm.	368/384	3.26%	7.81%
Synapse drivers	78 320	0.04%	0.04%
Synapse arrays	712	1.97%	1.97%
Synapse rows	159 488	0.11%	0.11%
Synapses	40 099 840	0.68%	0.68%
FG blocks	1492	0.34%	0.34%
External input mergers	2848/2984	0.0%	4.83%
Analog outputs	712	0.0%	0.0%
Background-generators	2848	0.0%	0.0%
Mergers	5340	0.0%	0.0%
Buses	—/119 360	—	2.2%
Repeaters	119 360	0.21%	0.22%
Switches	2864 640	0.02%	0.02%

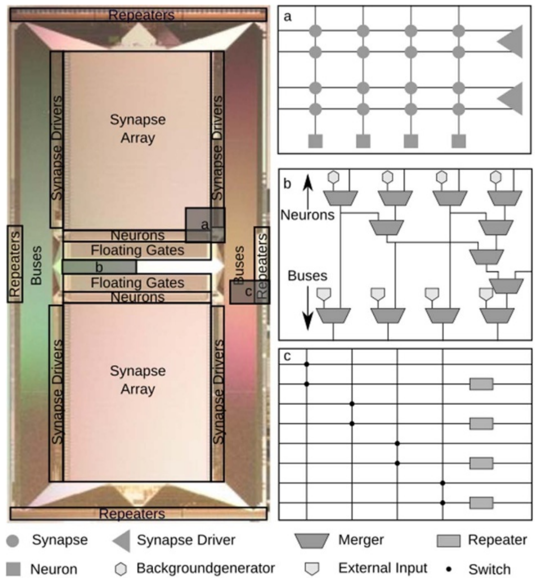

Figure 8. Left: picture of the HICANN with labeled components and marked areas shown on the right side. Top right: detail of the synapse array. Two synapse rows are connected to one synapse driver. All synapses of the same column are connected to one neuron circuit. Middle right: Left half of the merger tree. Neuron input from the top gets routed to the buses on the bottom. Several inputs can be merged on the same bus. Background-generators are used to inject additional signals generated on-chip. Right bottom: sketch of the bus system. Buses are connected by a sparse switch matrix. Repeaters, used to regenerate the signals, connect buses of neighboring HICANNs.

With 110 KiB per HICANN, the configuration registers of the synapses make up the largest part of the tested memory. They are split into two synapse arrays per HICANN, each of which is programmed by a custom on-chip SRAM controller described in [27]. In the tests, on 1.97% of the synapse arrays, unstable behavior is observed. This means, consecutive write/read operations with fixed values on a single synapse register show varying results. Since problems in individual synapse registers are very unlikely and could also derive e.g. from the control chain, a special stability test is introduced. There, each register is tested several times with the same value. If a single register shows unstable behavior, the whole synapse array is excluded. Thereby, at the expense of functional components, only stable programmable synapses are used during experiments.

A test with ten write/reads of random data per component and a stability test with ten repetitions takes approximately 70 s per HICANN. Since the tests can be executed in parallel for each HICANN group, a full wafer test takes approximately 10 min and can be executed periodically to track the state of the systems.

4.3. Effective exclusion of components

In special cases, it is not enough to skip malfunctioning components during an experiment, but it is also important to be aware of hardware specific dependencies that can be linked with these components. This is achieved through an additional step, the effective exclusion of components, where functional but dependent components are excluded. Several dependencies lead to an effective exclusion. Some of them are visualized in figure 8.

- Unstable repeater controller: to enhance the signal integrity of spike events that have to be routed across several HICANNs, the signal is regenerated between dies by repeaters. These repeaters are organized in blocks where each block has a custom on-chip controller used to program its repeaters. Since failures in the digital memory of the repeaters are very unlikely, more than one failing repeater per block indicates that there could also be a problem in the control chain. To ensure no unstable components are used, all repeaters connected to the corresponding repeater block are removed from the availability database in such cases.
- Buses connected to malfunctioning repeaters: buses are used to route spike events between neuron circuits. On boundaries between two HICANNs, the buses are connected to repeaters that regenerate the signal. Each repeater is connected to a bus on its own HICANN as well as on a neighboring one. If a repeater is failing the memory test, there is no possibility to test if it sends wrong signals to its connected buses. To circumvent this, all buses connected to such a repeater are excluded and thus not used during an experiment. The same holds for repeaters on HICANNs without JTAG connection. As the repeaters cannot be initialized correctly, all neighboring buses connected to repeaters on the problematic HICANN are excluded.
- Malfunctioning FG controller: the FGs are not only used to configure the neurons but also to supply bias voltages to the spike event routing. If an error in the controller programming the FGs is found, the whole HICANN is excluded from the availability database and, in the following, treated as if there would be no JTAG connection. Such a HICANN is not used at all in experiments.
- Without high-speed: HICANNs that have no high-speed connection are, due to the higher bandwidth requirements, not used to emulate neurons or external inputs but only used to route spike events. This is achieved by removing all neurons and external input mergers from the availability database.
- No routing options: to improve the placement and prevent lost connections, the algorithm checks that all the components required to establish a route from each neuron and external input merger are available. If not, the neuron or the external input merger is excluded and therefore skipped in the process of building a network.
- Handling hardware versioning: in an earlier version of the post-processing, connections were established to HICANNs on the edges of the wafer that must not be connected. To prevent leakage currents from these dies, the connected buses are excluded. Therefore, it is unnecessary to distinguish wafer versions in all the following steps.

An overview of removed components before and after the effective exclusion of components can be seen in table 1. The availability database, used to handle the excluded components, allows for storing different states on disk, so malfunctioning components and effective components can be differentiated afterward. This is for example important during the initialization of the HICANNs, where only malfunctioning components have to be handled specifically.

4.4. Analog readout tests

Before usage, the analog recording system gets verified for correct connectivity and configuration by running a series of tests. Each HICANN is set in sequence to generate two different voltage levels, which the AnaRM measures. The voltage levels originate from the configuration of one of the FGs. A recording that agrees with the settings and whose noise levels are within a tolerance threshold indicates that the system is ready for experiments or calibration runs.

4.5. Calibration

VLSI transistors are subject to manufacturing variations translating into differences in signal response. This problem and the potential impacts have been noted since the first approaches to neuromorphic computing using VLSI [1]. Consequently, the HICANN's microelectronic analog circuits require correction mechanisms to deliver homogeneous responses. As the manufacturing variability is stationary within the components' operating ranges, thus termed fixed-pattern noise, it can be reduced by suitable calibration. In addition, to allow the emulation of networks defined in the biological parameter space without requiring knowledge of hardware implementation details, an automatic translation of the neuron model parameters to a set of hardware parameters is in place and transparent for the users when running an experiment.

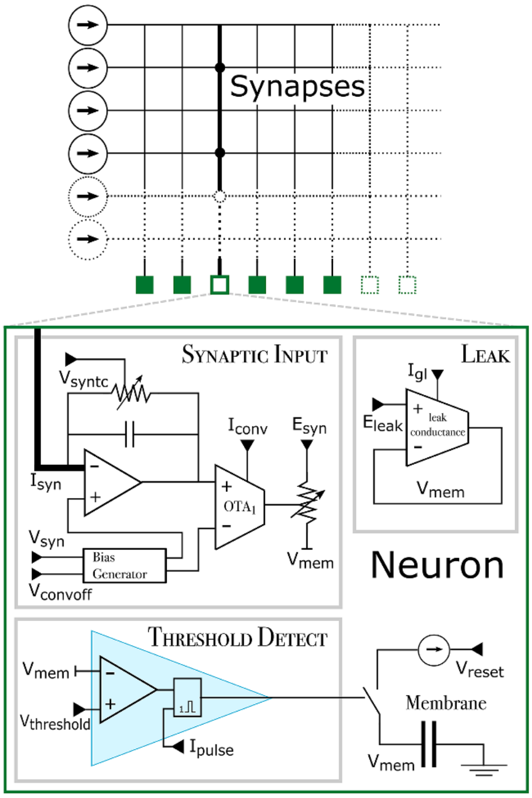

Figure 9. Simplified neuron circuit schematic, displayed on the bottom, with the most relevant calibrated parameters in the BrainScaleS-1 system. The leak conductance controlled by I_{gl} is constantly driving the membrane potential V_{mem} towards the rest voltage E_{leak}. A spike is elicited when the membrane potential reaches the threshold voltage $V_{\mathrm{threshold}}$. After a neuron spikes, its membrane's potential is connected to the reset voltage V_{reset} for a period controlled by the parameter I_{pulse}. For simplicity, one synaptic input is displayed out of two through which a neuron integrates excitatory and inhibitory input currents I_{syn}; this controls a conductance between the reversal potential E_{syn} and the membrane with a synaptic time constant controlled by V_{syntc}. Each input receives currents from all the synapses connected to one column in the synaptic array, displayed on top. Additional parameters V_{syn}, V_{convoff}, and I_{conv} provide further control over the synaptic input, as further discussed in the supplementary material.

We considered different calibration methodologies, like those introduced in [28, 29]. However, they either mandate a time-consuming recalibration before experiments or do not allow for flexible network topologies. Therefore, we opted for a one-time circuit characterization that runs sequences of experiments that sweep neuron parameters, measure the effect in the observable, and perform appropriate fits on suitable models. Thereby, we allow for the emulation of different networks with flexible parameterization and avoid a long time overhead before experiments.

The process creates a database that holds the calibration results and is stored to disk. Upon usage, the user selects the desired database for each experiment by providing the path to the stored files; these are then loaded for each wafer module, identified by a unique number assigned to the attached MaCU. See [26] for details.

The calibration procedure configures all the neuron circuits at once and then processes the individual measurements to allow for parallel rewriting of the FGs. Although the FGs represent a low-power solution to store analog operation settings and their size suits the high number of parameters in the HICANN, their reprogramming is involved, as it is done via an onboard DAC through incremental loops with feedback, which results in write-cycle to write-cycle variability. Parameters that are more sensitive to such variability benefit from an increase in the number of programming cycles, while also an increase in the number of calibration steps could improve the quality of the fits. Consequently, calibration time and precision of the results require balancing. Parallelizing the analysis algorithms further optimizes the time required for calibration.

4.5.1. Calibration methodology

In the BrainScaleS-1 system, the only analog neuron property that can be directly recorded is the membrane voltage. Accordingly, all parameter calibrations are based on membrane recordings under different parameter configurations. In general, the calibration of one parameter sweeps over its operating range while maintaining the rest of the parameters constant. The execution order is relevant, as some calibration routines

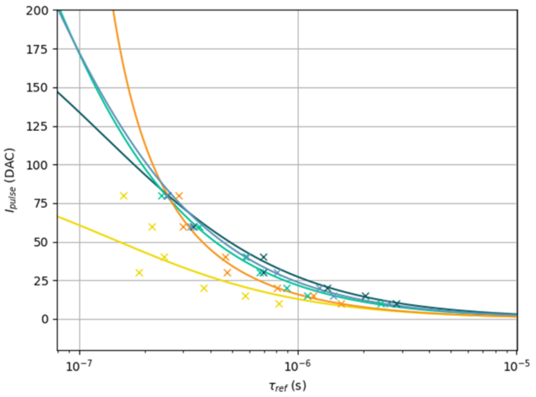

Figure 10. Exemplified calibration procedure for the refractory period. Sample fits obtained for a set of neurons, relating the I_{pulse} parameter configured with the floating gates with the measured τ_{ref}. Seven values within the dynamic range of I_{pulse} were used for the fits.

require an already calibrated subset of parameters. Furthermore, the calibration accounts for analog readout noise, and measurements can be repeated to factor in FG parameter variability.

The main neuron calibration parameters are summarized in figure 9. In the following, the calibration procedure is exemplarily shown for the parameter I_{pulse}, which controls the refractory period τ_{ref}, i.e. the time after the emission of a neuron's action potential during which its membrane is clamped to the reset potential and the neuron can elicit no further spike. The higher I_{pulse} is, the shorter the achieved τ_{ref}. Each I_{pulse} calibration step sets the resting potential E_{leak} above the level at which a spike event is elicited, i.e. $V_{\text{threshold}}$, which causes the neurons to spike continuously. The inter spike interval (ISI) is the measurable result.

In the first step, I_{pulse} is set to maximum, and the corresponding ISI is regarded as ISI_0, the minimum attainable interval under the current settings. Larger refractory periods are referenced to ISI_0 by using

$$\tau_{\text{ref}}\left(I_{\text{pulse}}\right) = \text{ISI}\left(I_{\text{pulse}}\right) - \text{ISI}_0, \tag{1}$$

making the minimum τ_{ref} zero seconds by definition. Afterward, each step's distinct target FG values of I_{pulse} are programmed, causing changes observable in the ISI and thus in τ_{ref}. The obtained set of configured parameters and their achieved refractory periods is then fit to a model, which in the case of τ_{ref} corresponds to

$$I_{\text{pulse}} = \frac{1}{(c_0 + c_1 \cdot \tau_{\text{ref}})}. \tag{2}$$

Such a model derives from transistor-level simulations described in [30]. The resulting fits for five neurons are shown in figure 10.

The pair of constants c_0 and c_1 corresponding to model equation (2) is stored in the calibration database for each neuron, which is then used for translation from τ_{ref} in seconds to I_{pulse} in digital value. Further details for each parameter calibration are provided in the supplementary material.

Depending on each parameter's sensitivity to the programmed FG values, some calibrations enable a more precise setting of parameters than others. An increased sensitivity due to non-linear hardware dependencies is found where small changes in FG values cause large changes in the observables. Furthermore, for some FGs only a limited range of their available parameter space is used, reducing the ability to set their corresponding parameters precisely. As can be seen from the measured values in figure 10, such is the case for I_{pulse}. For comparison, figure 11 shows how the leak potential E_{leak}, which is easier to control, obtains a more precise calibration than I_{pulse}. For this reason, the control precision of several parameters was improved in the second-generation BrainScaleS-2 chip [31] partly by enabling digital value storage.

4.5.2. Synapse weight calibration

The calibration of the synaptic input differs from the other calibrations due to its additional dependency on the synapse drivers. The strength of a synapse is configured by three hardware parameters. The 4-bit digital weight *w* stored per synapse, a scaling factor *gmax_div* stored per synapse row, and the FG-stored reference parameter V_{gmax}. This last parameter is set per synapse row and selects one of four possible values shared by blocks of 128 neurons. Calibrating this large parameter space for each of the 512 neurons with 110 connected synapse drivers using the analog readout system, which allows for measuring 12 membrane traces in parallel,

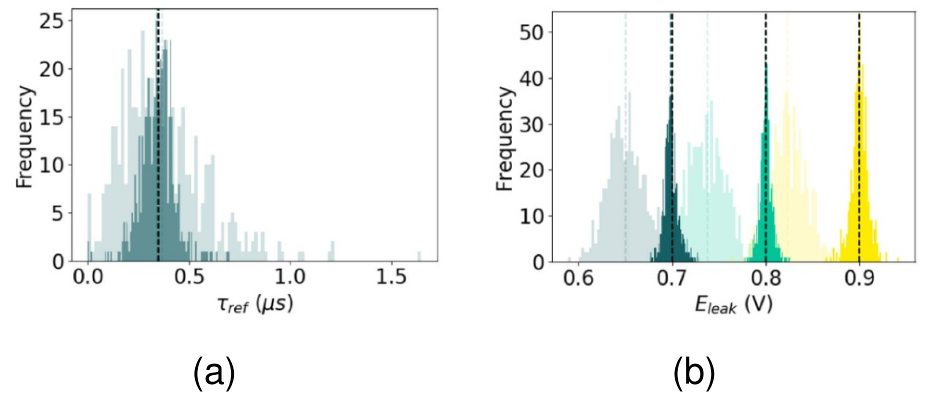

Figure 11. Histograms of achieved parameter settings on all neurons of one HICANN for (a) the refractory time constant controlled by the parameter I_{pulse} and (b) the leak potential controlled by the parameter E_{leak}. Pale and intense colors correspond to the hardware achieved time constants and voltages for different target values (shown in black-dashed lines), before and after the calibration is applied, respectively. The narrowing and centering of the achieved value distributions is better for E_{leak} than for τ_{ref}.

is not possible in a reasonable time frame. Therefore, a per wafer translation is performed, where only some of the components are taken into account to find the average circuit behavior. The measurement requires the results of all previous calibrations. Neurons on different HICANNs are stimulated by a single spike for different combinations of the three hardware parameters to cover the whole parameter range. Subsequently, a fit of the conductance based neuron model is applied to the recorded membrane traces to extract the ratio between biological weight and membrane capacitance $\frac{w_{\text{bio}}}{C_{\text{HW}}}$. Since the membrane capacitance is fixed during experiments, it is unnecessary to separately determine both values. During the fit, the model parameter of the already calibrated reversal potential is fixed. The reduced χ^2 value of the fit is used to identify and exclude saturation effects of the involved operational transconductance amplifier$_1$, cf figure 9, which might occur for large weight values. Finally, the weight translation is found by fitting the expected hardware behavior

$$A\left(\frac{w \cdot V_{\text{gmax}}}{gmax\_div} + i_0 + i_1 \cdot w_1 + i_2 \cdot w_2 + i_4 \cdot w_4 + i_8 \cdot w_8\right), \tag{3}$$

adapted from [32], to the results of the first fits. The fit parameters $i_{0\text{–}8}$ characterize the effect of parasitic capacitances found in the synaptic circuit for each enabled bit of the 4-bit weight value w. Figure 12(a) demonstrates the large parameter space of the synapse weight calibration. It shows the measurement of a single neuron, stimulated by a single synapse driver for a single V_{gmax} value without rewriting the FGs. The performance of the fit applied on the whole measured parameter space is shown for fixed values of *gmax_div* in figure 12(b) and for fixed digital weight values w in figure 12(c). Although the whole neuron circuit and consequently the expected noise of each individual component is involved, the error of each measurement does not exceed the variations observed in other calibrations. However, additional deviations arise from rewriting the FGs, which is demonstrated in figure 12(d); this renders the search for a more precise fit function unbeneficial. In addition, the per wafer calibration opted over a per neuron circuit calibration introduces a dominant error due to the deviations between neuron circuits, shown in figure 12(e). A precise weight calibration within a reasonable runtime would be achievable via a parallel measurement of each neuron circuit. This would also allow to exclude neurons showing unintended behavior. However, this is not possible with the currently used analog readout system. Nonetheless, the lack of a perfect weight calibration can be circumvented via in-the-loop training on the BrainScaleS-1 system, as shown for inference tasks in previous results [14].

4.5.3. Calibration based exclusion of components

The operation of the HICANNs during the calibration is similar to the operation during experiments. All components have to work correctly for the calibration to succeed. Failing calibrations indicate unintended behavior. This allows for testing the whole die, especially the analog circuits that cannot be tested directly. Additionally, thresholds can be defined to exclude outliers. Consequently, neurons that do not pass all calibration steps are excluded from the availability database. Numbers of calibration based excluded neurons on a typical wafer are given in table 2.

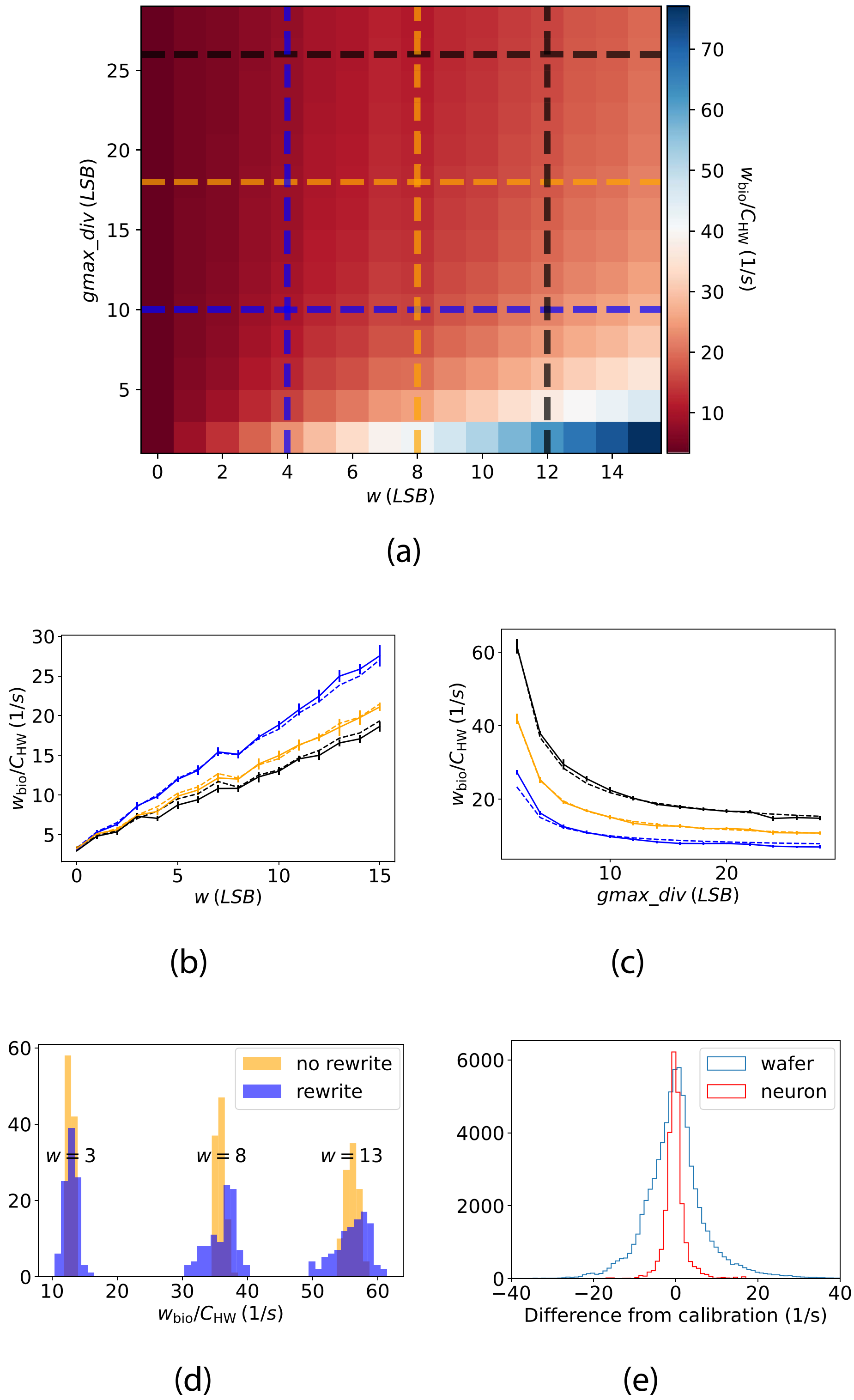

Figure 12. Results of the synapse weight calibration. (a) Weight measurement for a fixed neuron circuit for different settings of the digital weight *w* and hardware parameter *gmax_div* with $V_{\mathrm{gmax}} = 700$ LSB. Horizontal dashed lines indicate cuts with fixed values of the hardware parameter *gmax_div* shown in (b), vertical dashed lines indicate cuts with fixed digital weight values *w* shown in (c). In (b) and (c), solid lines represent measured values, dashed lines the results of the fit of equation (3) applied on the whole measured parameter space. (d) Variations of weight measurements with and without rewriting of the floating gates. Values are extracted for three digital weight parameters *w* from a fixed neuron with fixed hardware parameters ($V_{\mathrm{gmax}} = 700$ LSB, *gmax_div* = 2 LSB). (e) Comparison of a per wafer and a per neuron weight calibration. Measurements for the entire parameter space are performed on a subset of neurons. The calibration is then performed for the whole subset or per individual neuron. The histogram shows the difference between the measured and expected values using the obtained calibrations.

Table 2. Overview of calibration based excluded neurons of a fully assembled wafer module. In the column labeled 'Neurons' the first entry shows the number of neurons taken into account for the calibration, the second entry the number of neurons taken into account for the effective exclusion.

Neurons	Excluded by calibration	Effective exclusion
182 272/190 976	10.25%	14.59%

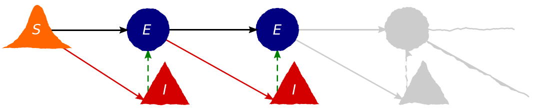

Figure 13. Structure of the synfire chains presented in this section. The synfire chain is made up of several groups of excitatory (blue) and inhibitory (red) populations. The inhibitory population connects to the excitatory population within the same group and aims to improve the chain's filtering for synchronous input [39, 42]. Each excitatory population is connected to the excitatory and inhibitory population of the next group. By repeating this construction schema (grey), chains of arbitrary length can be realized. The network is excited by a stimulus population (orange) which projects to the excitatory and inhibitory population of the first group.

Table 3. Parameters used for the synfire chains presented in section 5 .

Parameter	Short chain	Long chain
Chain length	6	190
Stimulus neurons	100	80
Excitatory neurons per group	100	80
Inhibitory neurons per group	25	20
Total number of neurons	750	19 000
Total number of neuron circuits	3000	76 000
Total number of synapses	≈73 000	$\approx 1.4 \times 10^6$
Used HICANNs	48	230

5. Experiment showcase—synchronous firing chain

Previous experiments on the BrainScaleS-1 system relied on a small subset of the available neurons [14, 33, 34]. In this section, we use a synchronous firing chain (synfire chain) to utilize a large number of the available wafer module resources. We start with a relatively short chain to illustrate the behavior of the network and finally present a longer one that utilizes a large part of a single wafer module.

Synfire chains can filter for synchronous activity and propagate the activity along a chain of neuron groups [35, 36]. We choose synfire chains since they can easily be scaled up to arbitrary sizes by increasing the chain length as well as the number of neurons in a single group and have been studied extensively in previous publications [37–39]. Furthermore, synfire chains were used to showcase the functionality of the predecessor of BrainScaleS-1 [40] and to characterize the behavior of the current system in software simulations [41].

Figure 13 displays a synfire chain with feed-forward inhibition. Each chain link consists of an excitatory and inhibitory population. The inhibitory populations are connected to the excitatory population within the same group. This feed-forward inhibition can enhance the filtering properties of the chain [39, 42]. The excitatory population forwards its outputs to both populations within the next group. External stimulus is injected in the form of Gaussian pulse packages [37]. The strength a denotes the number of input spikes per stimulus neuron and σ the standard deviation of the Gaussian from which the spike times are drawn. We will use (a, σ) to refer to specific packages.

5.1. Network behavior

In a first step we will look at a relatively short chain with six chain links, shown in figure 14, to illustrate how the filtering properties of the chain can be tuned. Table 3 summarizes some of the key properties of the network. We used the manual placement described in [26] to place the different populations on the wafer. Specifically, we distribute the external stimulus over several HICANNs in order to minimize spike loss due to limited bandwidth.

As mentioned previously, synfire chains are able to filter for synchronous input and to synchronize less-synchronous input as it travels along the chain [36, 37]. Figure 14(a) shows the propagation of three different input stimuli along the chain. In case of a relatively weak and synchronous input $(1, 1)$ a single, narrow package travels along the chain. If the input is stronger and more asynchronous, we observe a broader response in the first groups of the chain which is synchronized as the signal propagates along the

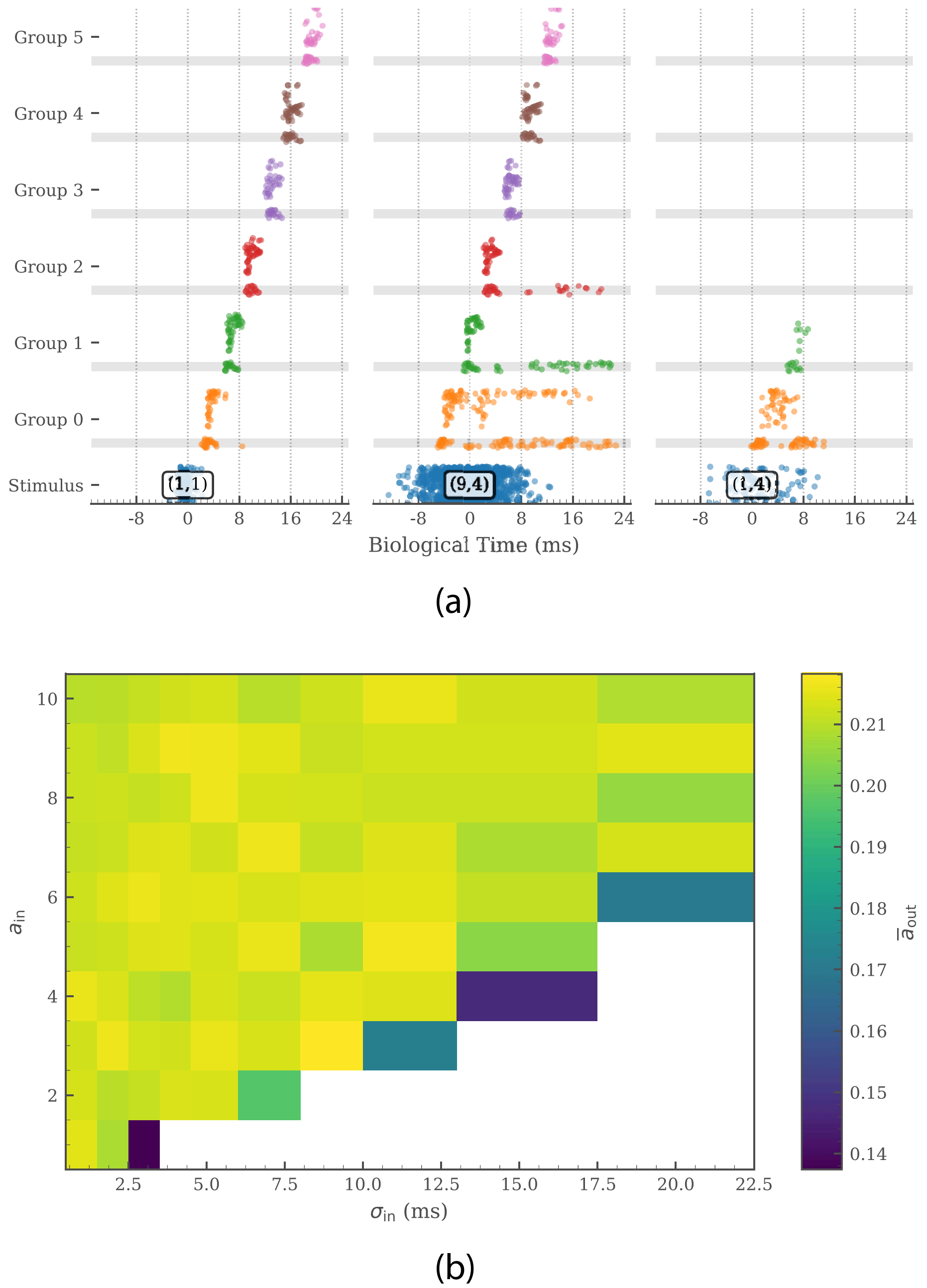

Figure 14. Hardware emulation of a chain with six chain links. (a) Propagation of pulse packages along the chain. Successful propagation depends both on the strength a as well as the synchronicity σ of the initial stimulus, represented by (a, σ). Broad input stimuli synchronize along the chain or do not reach the end of the chain. (b) Average number of spikes per neuron in the final group $\bar{a}_{\text{out}}$ of the chain in dependency on the initial strength a and synchronicity σ. Each input package was presented 40 times and the results are averaged over all presentations. The pulse packages propagate if the initial input is strong and synchronous enough. In the region of stable propagation the output strength is almost constant, near the separation of the two regimes the average strength of the final pulse package decreases. This separation line between successful propagation and failure of transmission can be controlled by several parameters such as the synaptic weights.

chain such that the responses in the final group are comparable. Too weak and asynchronous input, here $(1, 4)$ as an example, dies out and does not cause a response in the final group. This is in agreement with previous results [37, 40–42].

Figure 14(b) shows in more detail for which input stimuli the propagation along the chain is successful. In agreement with the previous observations, weak and asynchronous input is not transmitted to the final group. The response in the final group is almost uniform. This indicates that the packages are synchronized as they travel along the chain. Setting appropriate parameters which reproduce the expected results from simulations relies on the calibration routines, introduced in section 4.5. The calibration allows to set model parameters in the biological domain and reduces the inherent mismatch between the physical components.

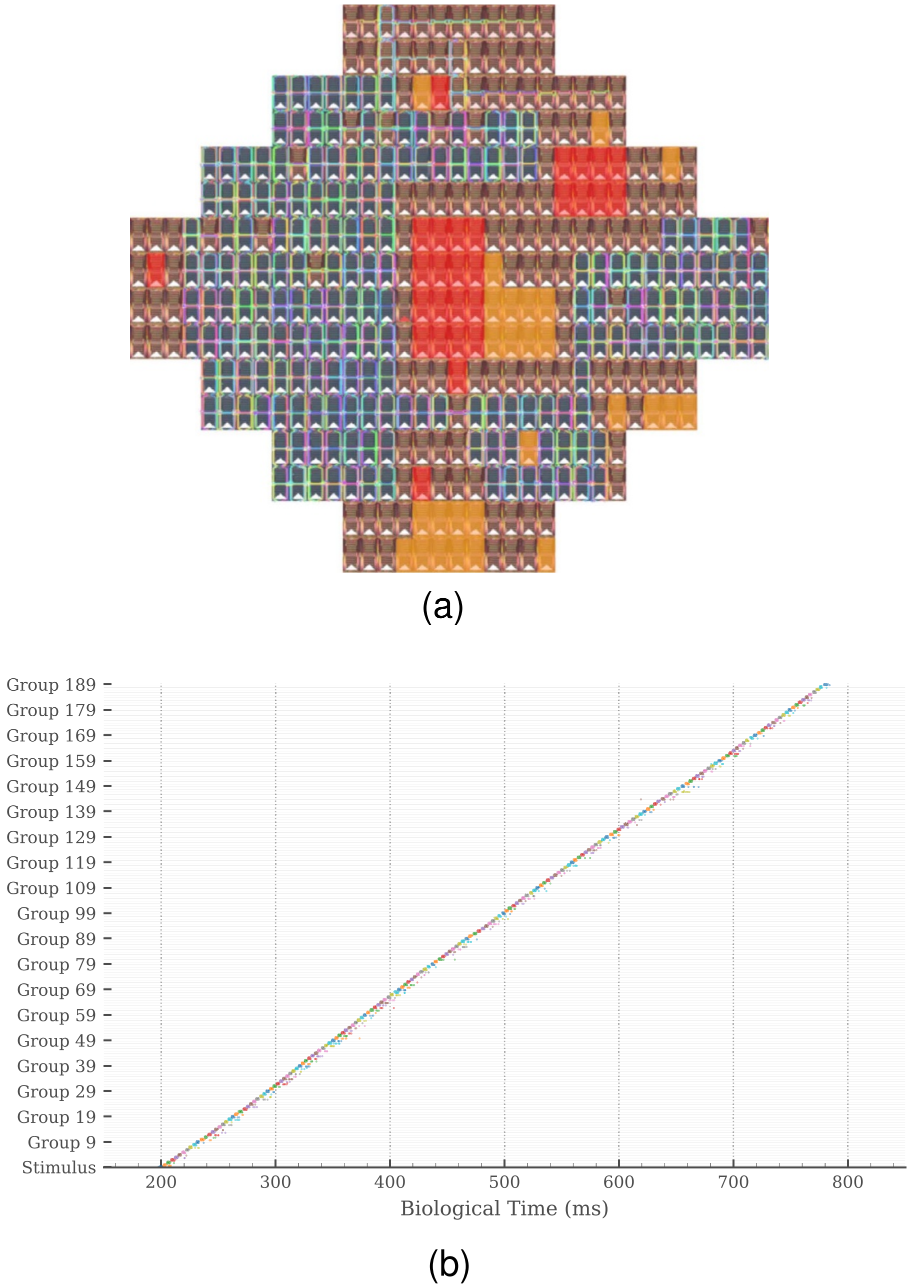

Figure 15. Hardware emulation of a chain with 19 000 neurons. Further parameters of the network can be found in table 3. (a) Mapping of the network to a BrainScaleS-1 wafer. HICANNs excluded from the availability database are marked in red, cf section 4.2. HICANNs which cannot host an entire group are marked in orange and are not used in the experiment. On each HICANN colored in blue an entire group of neurons is placed. Colored lines indicate synaptic connections. (b) Response of the chain to an input packet of strength $a = 1$ and spread $\sigma = 1$.

5.2. Wafer-scale network

The previous section demonstrates the implementation and control of a short synfire chain on the BrainScaleS-1 system. This section shows that the commissioning efforts described in section 4 also facilitate the implementation of wafer-scale networks. The properties of this synfire chain are summarized in table 3.

The complexity of the emulation increases with the size of the model. While for a relatively short chain it is possible to investigate the behavior of individual neurons and manually detect malfunctioning and bad calibrated entities, this is not feasible for larger experiments. Therefore, digital tests described in section 4.2 are essential to automatically avoid these components during the experiment.

To simplify the automatic routing of the abstract network description to physical entities on the wafer, we once again employ manual mapping, see figure 15(a). We place the different groups in a zig–zag pattern starting from the top-left side towards the bottom of the wafer and then back up towards the top-right side. This placement schema allows the BrainScaleS-1 operating system [26] to find appropriate connections between the different populations and minimizes synapse loss, i.e. synaptic connections that could not be mapped to the hardware.

We were able to successfully emulate a synfire chain with 190 chain links on the BrainScaleS-1 system. Figure 15(b) shows an example of a pulse package that travels along the full length of the chain. The activity of the individual groups still depends on the exact neuron and synapse properties, but the calibration ensures that the pulse package remains compact. A synchronous pulse reaches the final group after a signal propagation time of about 600 ms in the biological regime, which corresponds to 60 μs wall-clock time.

6. Discussion

Starting its development more than ten years ago, the first-generation BrainScaleS wafer-scale neuromorphic system represents a milestone toward a large-scale analog neural network emulation platform. Over years during which several modules have been commissioned and experiments run, we have learned important lessons on building and handling such a complex system. We discovered drawbacks in our first implementation; some of them could successfully be circumvented via our commissioning software. Our second-generation neuromorphic BrainScaleS-2 chip [31] addresses BrainScaleS-1's design weaknesses. Moreover, it enables the application of advanced learning mechanisms by introducing a digital plasticity processor, neuron multi-compartment capabilities, as well as extended analog to digital conversion capacities.

In this paper, we described the individual components of a BrainScaleS-1 wafer module and showed the necessary steps to assemble it. A wafer-scale analog system is complex and requires many hardware components working concurrently. Once a wafer module is assembled, it is often not possible to pinpoint defects in individual components. To alleviate this, the first lesson we would like to provide is the importance of tests, which are good engineering practice in any case but become critical with more components in a system. Here, each component must get tested on its own; malfunctioning ones must be repaired or replaced before they are added to the system. Also, tests during the assembly are crucial to find and solve errors that arise during that process. Furthermore, we demonstrated that still remaining problems can be handled by an appropriate resource management, in our case by the availability database, which ensures the correct operation of the system.

The importance of the tests and monitoring remains after the wafer module gets placed in the rack. For example, tight monitoring during system operation is necessary to uncover the wear out of system components. Automated alerts warn in case of values deviating over time. Furthermore, the tests executed nightly help keep track of the wafer modules' state. We learned that these measures are fundamental to maintain systems of this complexity over a long period of time.

Concerning the wafer in the core of the BrainScaleS-1 system, the probability of fabrication defects in microelectronics is proportional to the circuit area [43]. Thus, it is unfeasible to build such a large analog system without malfunctioning components. This will further intensify in the future by utilizing novel materials. With this in mind, the digital tests introduced are executed nightly to identify such malfunctioning components. Special about the tests is that each component is tested individually, and all hardware-specific interdependencies are considered, achieving a minimal but sufficient exclusion of components from the availability database; this is only possible due to the redundant system design. Moreover, this database contains only a sparse resource representation, resulting in reduced memory consumption and loading time during operation. Different database states can be stored to disk and flexibly adapted, even enabling the usage of experiment-specific availability data. Here, we emphasize the importance of testability of individual functional units in the system, ideally using, e.g. on-chip direct analog readout of relevant parameters or quantities, as well as means of direct software access for testing. Without such direct tests, failing sub-units could only be detected by malfunctions in parts with a higher hierarchical level on the system, potentially including working sub-units. Consequently, we compensate for missing test capabilities by excluding additional usable components, which is only feasible due to the large component count in our system.

An additional challenge using analog hardware is the fixed-pattern noise introduced by unavoidable manufacturing process variations. In the BrainScaleS-1 system, this is worsened by the design decision to use FGs to store the neuron configuration. These cells allow for long-term storage of analog parameters without storing digital values onboard. However, the current implementation introduces write-cycle to write-cycle variability. Though small, these variations lead to noticeable errors if they are further enlarged by non-linear dependencies between control signal and observable. To minimize these effects, we presented our calibration framework, which also allows non-expert users to configure experiments in the biological domain without specific knowledge of the hardware. We demonstrated the narrowing and centering of the achieved value distribution for exemplary parameters after the calibration was applied, limited by thermal noise and the variations caused by the FGs, nonetheless. There, we want to advise that, in order to reliably operate a mixed-signal system of the size of BrainScaleS-1, especially analog circuits should rather be designed a bit more conservatively than close to the limit of what is possible with the used manufacturing technology (e.g. regarding operating speed and component size). Else, even small adjustments of the operating point could lead to non-linear circuit behavior. Furthermore, in our experience the FG cells were not suited for precise parameter storage since they are non-standard devices and not supported by the manufacturer. Therefore, the second-generation BrainScaleS-2 chip reverts to a digital storage scheme for parameter values as employed in a previous neuromorphic architecture [44], thereby vastly improving analog parameter accuracy. Since the second generation uses a manufacturing process with much smaller geometry, namely 65 nm vs. 180 nm, the area penalty for the digital parameter storage is manageable. A further advantage of the novel parameter storage is the reduced programming time [45]. In the presented wafer-scale implementation, the FG parameter storage was the only feasible solution to achieve the required number of analog parameters for the neuron circuits.

On top of explaining the calibration methodology, we demonstrated the necessity for parallel execution of the calibrations. The large parameter space of the synapse weight calibration exceeds reasonable runtimes using the current readout system. In order to circumvent this, we introduced a per wafer calibration which, compared to a per circuit calibration, shows larger errors but can be generated in a reasonable time frame. To improve this, we developed a new readout system, which will replace the external set of ADCs with on-wafer-module boards, increasing the parallel readout capabilities from 12 to 96 channels [46]. It is directly mounted on the wafer module to reduce the connection length and thereby improving the signal quality. Moreover, in the BrainScaleS-2 chip, we introduce a per neuron-circuit ADC system, which allows for a massive parallel calibration [31]. A per-circuit calibration before each experiment becomes feasible with such a solution. Therefore, the takeaway is to provide enough parallel measurement capabilities for important observables.

Finally, we demonstrated the operation of a fully commissioned BrainScaleS-1 wafer module implementing synfire chains. While small chains portray the capability to fine-tune the network parameters, extending to a long chain of 190 links illustrates the possibility to scale up networks. Successfully mapped to an inherently imperfect substrate, it consists of the largest spiking network emulation run with analog components and individual synapses to date.

Our endeavor in developing and maintaining the BrainScaleS-1 system has demonstrated, while illustrating the field's challenges, that building wafer-scale analog neuromorphic hardware is feasible. Furthermore, the BrainScaleS-1 wafer module with its operating system laid the foundation for the next-generation systems; all lessons learned from the first generation contribute to the success of future large-scale neuromorphic systems.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The authors wish to thank all present and former members of the Electronic Vision(s) research group contributing to the BrainScaleS-1 platform, development and operation methodologies, as well as software development. We thank S Schiefer and S Hartmann from the group Hochparallele VLSI-Systeme und Neuromikroelektronik at TU-Dresden for the development and before-assembly routine testing of the communication boards and the WIOs, as well as for providing test details and images for this writing. We thank M Yaziki and O Ceylan from the group of Yasar Gürbüz at Sabanci University, Istanbul for the development of test boards that are being used during assembly of the wafer modules. We thank Würth Elektronik, Germany for dedicated and detailed support during the development of the main PCB. This work has received funding from the EU ([FP7/2007-2013], [H2020/2014-2020]) under Grant Agreements 604102 (HBP), 269921 (BrainScaleS), 243914 (Brain-i-Nets), 720270 (HBP SGA1), 785907 (HBP SGA2) and 945539 (HBP SGA3), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster), the Helmholtz Association Initiative and Networking Fund (ACA, Advanced Computing Architectures) under Project SO-092, as well as from the Manfred Stärk Foundation.

ORCID iDs

Hartmut Schmidt https://orcid.org/0009-0006-1630-3614
José Montes https://orcid.org/0009-0002-4687-5315
Andreas Grübl https://orcid.org/0000-0002-3955-4815

Jakob Kaiser https://orcid.org/0000-0002-3586-2634
Christian Mauch https://orcid.org/0000-0001-6566-7170
Eric Müller https://orcid.org/0000-0001-5880-2012

References

[1] Mead C A 1989 *Analog Vlsi and Neural Systems* (Addison Wesley)
[2] Mead C A 1990 Neuromorphic electronic systems *Proc. IEEE* **78** 1629–36
[3] Indiveri G *et al* 2011 Neuromorphic silicon neuron circuits *Front. Neurosci.* **5** 73
[4] Furber S B, Galluppi F, Temple S and Plana L A 2014 The SpiNNaker project *Proc. IEEE* **102** 652–65
[5] Davies M *et al* 2018 Loihi: a neuromorphic manycore processor with on-chip learning *IEEE Micro* **38** 82–99
[6] Merolla P A *et al* 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345** 668–73
[7] Moradi S, Qiao N, Stefanini F and Indiveri G 2018 A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs) *IEEE Trans. Biomed. Circuits Syst.* **12** 106–22
[8] Furber S 2016 Large-scale neuromorphic computing systems *J. Neural Eng.* **13** 051001
[9] Schuman C D, Potok T E, Patton R M, Birdwell J D, Dean M E, Rose G S, and Plank J S 2017 A survey of neuromorphic computing and neural networks in hardware (arXiv:1705.06963)
[10] Schemmel J, Fieres J and Meier K 2008 Wafer-scale integration of analog neural networks *Proc. 2008 Int. Joint Conf. on Neural Networks (IJCNN)*
[11] Fieres J, Schemmel J and Meier K 2008 Realizing biological spiking network models in a configurable wafer-scale hardware system *Proc. 2008 Int. Joint Conf. on Neural Networks (IJCNN)*
[12] Schemmel J, Brüderle D, Grübl A, Hock M, Meier K and Millner S 2010 A wafer-scale neuromorphic hardware system for large-scale neural modeling *Proc. 2010 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp 1947–50
[13] Millner S, Grübl A, Meier K, Schemmel J and Schwartz M-O 2010 A VLSI implementation of the adaptive exponential integrate-and-fire neuron model *Advances in Neural Information Processing Systems* vol 23, ed J Lafferty, C K I Williams, J Shawe-Taylor, R Zemel and A Culotta pp 1642–50
[14] Schmitt S *et al* 2017 Neuromorphic hardware in the loop: training a deep spiking network on the BrainScaleS wafer-scale system *Proc. 2017 IEEE Int. Joint Conf. on Neural Networks (IJCNN)* pp 2227–34
[15] Gerstner W and Brette R 2009 Adaptive exponential integrate-and-fire model *Scholarpedia* **4** 8427
[16] Millner S 2012 Development of a multi-compartment neuron model emulation *PhD Dissertation* Ruprecht-Karls-University Heidelberg
[17] Lande T, Ranjbar H, Ismail M and Berg Y 1996 An analog floating-gate memory in a standard digital technology *Proc. 5th Int. Conf. on Microelectronics for Neural Networks (Lausanne, Switzerland, 12–14 1996)* (IEEE Computer Society Press) pp 271–6
[18] Zoschke K, Guettler M, Boettcher L, Gruebl A, Husmann D, Schemmel J, Meier K and Ehrmann O 2017 Full wafer redistribution and wafer embedding as key technologies for a multi-scale neuromorphic hardware cluster *2017 IEEE 19th Electronics Packaging Technology Conference (EPTC 2017)*
[19] Thanasoulis V, Partzsch J, Hartmann S, Mayr C and Schüffny R 2012 Dedicated fpga communication architecture and design for a large-scale neuromorphic system *2012 19th IEEE Int. Conf. on Electronics, Circuits and Systems (ICECS 2012)* pp 877–80
[20] Hartmann S, Schiefer S, Scholze S, Partzsch J, Mayr C, Henker S and Schuffny R 2010 Highly integrated packet-based aer communication infrastructure with 3Gevent/S throughput *2010 17th IEEE Int. Conf. on Electronics, Circuits and Systems (ICECS)* pp 950–3
[21] Scholze S, Schiefer S, Partzsch J, Hartmann S, Mayr C, Höppner S, Eisenreich H, Henker S, Vogginger B and Schüffny R 2011 Vlsi implementation of a 2.8 Gevent/s packet-based aer interface with routing and event sorting functionality *Front. Neurosci.* **5** 117
[22] Sterzenbach L 2014 Entwicklung einer selbstüberwachenden Spannungsversorgung für ein auf Wafer-Ebene integriertes neuromorphes Hardware-System *Bachelor Thesis* (German) Hochschule Mannheim University of Applied Sciences
[23] Davis C 2006 Graphite-project Carbon (available at: https://github.com/graphite-project/carbon)
[24] Grafana Labs 2018 Grafana: the open observability platform (available at: https://grafana.com)
[25] Elastic 2015 Elasticsearch: the official distributed search & analytics engine (available at: www.elastic.co/elasticsearch)
[26] Müller E *et al* 2022 The operating system of the neuromorphic BrainScaleS-1 system *Neurocomputing* **501** 790–810
[27] Friedmann S 2013 A new approach to learning in neuromorphic hardware *PhD Dissertation* Ruprecht-Karls-Universität Heidelberg
[28] Buhry L, Grassia F, Giremus A, Grivel E, Renaud S and Saïghi S 2011 Automated parameter estimation of the Hodgkin-Huxley model using the differential evolution algorithm: application to neuromimetic analog integrated circuits *Neural Comput.* **23** 2599–625
[29] Neftci E O, Toth B, Indiveri G and Abarbanel H 2012 Dynamic state and parameter estimation applied to neuromorphic systems *Neural Comput.* **24** 1669–94
[30] Schwartz M-O 2013 Reproducing biologically realistic regimes on a highly-accelerated neuromorphic hardware system *PhD Dissertation* Universität Heidelberg
[31] Schemmel J, Billaudelle S, Dauer P and Weis J 2022 *Accelerated Analog Neuromorphic Computing* (Springer International Publishing) pp 83–102
[32] Koke C 2017 Device variability in synapses of neuromorphic circuits *PhD Dissertation* Ruprecht-Karls-University Heidelberg
[33] Kungl A F *et al* 2019 Accelerated physical emulation of Bayesian inference in spiking neural networks *Front. Neurosci.* **13** 1201
[34] Göltz J *et al* 2021 Fast and energy-efficient neuromorphic deep learning with first-spike times *Nat. Mach. Intell.* **3** 823–35
[35] Aertsen A, Diesmann M and Gewaltig M O 1996 Propagation of synchronous spiking activity in feedforward neural networks *J. Phys.* **90** 243–7
[36] Gewaltig M O, Diesmann M and Aertsen A 2001 Propagation of cortical synfire activity: survival probability in single trials and stability in the mean *Neural Netw.* **14** 657–73
[37] Diesmann M, Gewaltig M-O and Aertsen A 1999 Stable propagation of synchronous spiking in cortical neural networks *Nature* **402** 529–33
[38] Diesmann M 2002 Conditions for stable propagation of synchronous spiking in cortical neural networks: single neuron dynamics and network properties *PhD Dissertation* Ruhr-Universität Bochum

[39] Kumar A, Rotter S and Aertsen A 2008 Conditions for propagating synchronous spiking and asynchronous firing rates in a cortical network model *J. Neurosci.* **28** 5268–80

[40] Pfeil T, Grübl A, Jeltsch S, Müller E, Müller P, Petrovici M A, Schmuker M, Brüderle D, Schemmel J and Meier K 2013 Six networks on a universal neuromorphic computing substrate *Frontiers Neurosci.* **7** 11

[41] Petrovici M A *et al* 2014 Characterization and compensation of network-level anomalies in mixed-signal neuromorphic modeling platforms *PLoS One* **9** e108590

[42] Kremkow J, Perrinet L, Masson G and Aertsen A 2010 Functional consequences of correlated excitatory and inhibitory conductances in cortical networks *J. Comput. Neurosci.* **28** 579–94

[43] Werner S, Navaridas J and Luján M 2016 A survey on design approaches to circumvent permanent faults in networks-on-chip *ACM Comput. Surv.* **48** 1–36

[44] Schemmel J, Meier K and Muller E 2004 A new VLSI model of neural microcircuits including spike time dependent plasticity *Proc. 2004 Int. Joint Conf. on Neural Networks* (*IJCNN 2004*) (IEEE Press) pp 1711–6

[45] Hock M, Hartel A, Schemmel J and Meier K 2013 An analog dynamic memory array for neuromorphic hardware *2013 European Conf. on Circuit Theory and Design* (*ECCTD*) pp 1–4

[46] Ilmberger J 2017 Development of a digitizer for the brainscales neuromorphic hardware system *Master Thesis* Ruprecht-Karls-Universität Heidelberg