A CMOS-based microelectrode array for interaction with neuronal cultures

S. Hafizovic\textsuperscript{a,*}, F. Heer\textsuperscript{a}, T. Ugniwenko\textsuperscript{b}, U. Frey\textsuperscript{a}, A. Blau\textsuperscript{b}, C. Ziegler\textsuperscript{b}, A. Hierlemann\textsuperscript{a}

\textsuperscript{a} Physical Electronics Laboratory, ETH Zurich, Wolfgang-Pauli Str. 16, 8093 Zurich, Switzerland
\textsuperscript{b} Department of Physics, University of Kaiserslautern, Kaiserslautern, Germany

Received 29 July 2006; received in revised form 2 April 2007; accepted 6 April 2007

Abstract

We report on the system integration of a CMOS chip that is capable of bidirectionally communicating (stimulation and recording) with electrogenic cells such as neurons or cardiomyocytes and that is targeted at investigating electrical signal propagation within cellular networks \textit{in vitro}. The overall system consists of three major subunits: first, the core component is a 6.5 mm $\times$ 6.5 mm CMOS chip, on top of which the cells are cultured. It features 128 bidirectional electrodes, each equipped with dedicated analog filters and amplification stages and a stimulation buffer. The electrodes are sampled at 20 kHz with 8-bit resolution. The measured input-referred circuitry noise is $5.9 \mu V_{\text{rms}}$ (10 Hz to 100 kHz), which allows to reliably detect the cell signals ranging from 1 mV to 40 $\mu V_{p-p}$. Additionally, temperature sensors, a digital-to-analog converter for stimulation, and a digital interface for data transmission are integrated. Second, there is a reconfigurable logic device, which provides chip control, event detection, data buffering and an USB interface, capable of processing the 2.56 million samples per second. The third element includes software that is running on a standard PC performing data capturing, processing, and visualization. Experiments involving the stimulation of neurons with two different spatio-temporal patterns and the recording of the triggered spiking activity have been carried out. The response patterns have been successfully classified (83% correct) with respect to the different stimulation patterns. The advantages over current microelectrode arrays, as has been demonstrated in the experiments, include the capability to stimulate (voltage stimulation, 8 bit, 60 kHz) spatio-temporal patterns on arbitrary sets of electrodes and the fast stimulation reset mechanism that allows to record neuronal signals on a stimulating electrode 5 ms after stimulation (instantaneously on all other electrodes). Other advantages of the overall system include the small number of needed electrical connections due to the digital interface and the short latency time that allows to initiate a stimulation less than 2 ms after the detection of an action potential in closed-loop configurations.

© 2007 Elsevier B.V. All rights reserved.

Keywords: CMOS; Microelectrode array; Extracellular recording; Event detection; Liquid state machine; Support vector machine

1. Introduction

Microelectrode arrays (MEA) (Morin et al., 2005; Potter, 2001) are used to extracellularly measure the induced voltage on an electrode underneath a cell upon the occurrence of an action potential (Hodgkin and Huxley, 1952). The obtained electrode signal amplitudes typically range between 100 $\mu$V and 1 mV, depending on the cell type. With increasing cell–electrode distance or upon less direct cell–electrode contact, signal amplitudes may become arbitrarily small (Claverol-Tinture and Pine, 2002).

MEAs have become an important tool for non-invasive recording in the fields of neuroscience and biosensing (Baumann et al., 2004; Marom and Shahaf, 2002; Morin et al., 2005). Artificial neural networks on MEAs can potentially be trained to perform, e.g., pattern recognition tasks (Marom and Shahaf, 2002; Wagenaar et al., 2006). Additionally, the ability to monitor the electrophysiological response of a cell, or cell culture to pharmacological agents facilitates applications in the area of pharmacoscreening (Stett et al., 2003). Previously described MEAs typically are either passive devices (Buitenweg et al., 2002b; Gross et al., 1977; James et al., 2004; Jimbo et al., 2003; Maher et al., 1999; MCS, 2006), or rely on integrated
multiplexing circuits combined with off-chip electronics for system control and signal conditioning (Baumann et al., 2004; Berdondini et al., 2005; Eversmann et al., 2003). The passive devices are usually robust and deliver good signal quality but are limited in the number of electrodes by the number of electrical interconnects. MEAs with integrated multiplexing but without on-chip filtering feature more electrodes but often suffer from lower signal quality, mainly due to the lack of anti-aliasing filters.

The CMOS-MEA presented here is highly integrated and features 128 bidirectional electrodes, each of which has dedicated signal conditioning circuitry and is sampled at 20 kHz and 8-bit resolution. Any subset of electrodes may be chosen for stimulation of the cell culture. After stimulation has been terminated, the read-out is immediately re-established due to an effective artifact suppression technique (see details in Section 2.2.2 and Fig. 7). The digital on-chip circuitry combined with external reconfigurable logic (field-programmable gate array, FPGA) for event detection entails a very short response time so that the system may be considered real-time with respect to neural cell signaling. A stimulation pattern can be triggered within 2 ms after detection of an event (latency <2 ms).

2. Materials and methods

The system includes three custom-engineered components: (1) the MEA ASIC, (2) the FPGA and (3) the software. Fig. 1 depicts the three components and their interfacing. Major challenges arise from the operating conditions (incubator: 95% rel. humidity, 37 °C), the requested biocompatibility of chip and package, the large data volumes (3.2 MB/s) that have to be handled, the need for low-noise analog amplifiers and filters, and the low-latency requirements (2 ms).

2.1. Chip design

The micrograph of the 6.5 mm × 6.5 mm chip in Fig. 2 (left) shows three main areas, which are clearly distinguishable: the 128-electrode array, the column of 16 A/D-converters, and the digital core. The design is modular in that the same electrode unit is repeated 128 times, and in that each row of eight electrode units includes an A/D converter, so that there are 16 analog-to-digital (A/D) converters in total. Considerable space between the electrodes and the bondpads is needed to allow for an effective sealing of the electrical interconnects off chip.

The raw electrode signal of 0.1–1 mVpp is conditioned in a circuitry block at each electrode, then multiplexed and further amplified before it undergoes the 20-kHz, 8-bit A/D-conversion. The total amplification is selectable, either 1000- or 3000-fold.

The combination of sophisticated analog signal processing with a digital interface renders the overall multi-electrode array system very compact. Fig. 1 (right) shows all needed components: a laptop computer, an USB-powered FPGA card, and a simple PCB to provide reference voltages and power supply stabilization.

Electrogenic cells are very sensitive to temperature, so that temperature changes may alter the cell activity or even lead to cell death. An on-chip temperature sensor is hence provided to monitor the cell-area temperature.
2.1.1. Electrode array/analog circuitry

Each of the $250 \mu m \times 250 \mu m$ electrode circuitry repeating units in the array includes a high-pass filter with 20 dB gain, a low-pass filter, a 30-dB amplifier, a stimulation buffer, and digital circuits for 2-bit mode storage as shown in Fig. 3. Electromagnetic interferences are reduced and an effective power supply rejection is realized by means of a fully differential architecture. The design goal for the input-equivalent noise of the circuitry is determined by the thermal noise of the platinum electrodes (Section 2.4) in physiological saline solution, which has been measured to be $3.19 \mu V$ root mean square (rms) for $10 \mu m \times 10 \mu m$ and $2.16 \mu V_{\text{rms}}$ for $40 \mu m \times 40 \mu m$ electrodes within a band from 10 Hz to 100 kHz.

In comparison to other CMOS electrode arrays (Eversmann et al., 2003; Kovacs, 2003), important advantages arise from the modular architecture with buffers and filters implemented at each electrode.

(i) The high-pass filter removes offset and slow drift of the biochemical signals and, thereby, allows for immediate signal amplification.

(ii) The low-pass filter limits the noise bandwidth and acts as an anti-aliasing filter for the subsequent A/D-conversion.

(iii) The signal is amplified and filtered in close vicinity of the electrodes, which makes the design less sensitive to noise and distortion picked up along connection lines.

(iv) A stimulation buffer makes the stimulation signal independent of the number of activated electrodes. Additionally, the high-pass filter has a reset in order to ensure operability of the sensitive recording electronics also immediately after stimulation.

A limitation in the number of electrodes arises from the chip area needed for the individual electrode circuitry units. A MEA based on the electrode circuitry unit presented here ($250 \mu m \times 250 \mu m$) with 1024 electrodes would consume a total

Fig. 2. Micrograph of the electrode array chip and close-up of one electrode repeating unit. Left: the chip features an 8-by-16 electrode array in the center, and 16 A/D-converters and the digital block at the right. Right: close-up of the 128-times-repeated circuitry unit.

Fig. 3. Schematic of the on-chip electronics architecture and its components. The stacked frames indicate that these subunits are repeated for each electrode or each row. The chip also includes a temperature sensor. HPF denotes high-pass filter, LPF low-pass filter. A concise description of the on-chip analog circuitry is provided in Heer et al. (2006).
chip area of 8 mm × 8 mm and could be readily implemented. However, electrode numbers as high as 16,384 are reported in (Eversmann et al., 2003) cannot be realized with this design.

2.1.1. Stimulation. Desired stimulation features include:

- Possibility to stimulate any subset of electrodes,
- Flexibility in the stimulation waveform,
- Current and voltage stimulation capability,
- Capability to do a fast switching from stimulation to recording even on the same electrode.

The real-time all-channel stimulator by Wagenaar and Potter (2004), based on discrete off-chip components, offers these features except for the last: the time, during which an electrode cannot be used for recording after stimulation was reported to be longer than 40 ms. The presented monolithic CMOS implementation meets all criteria except for the possibility to do current stimulation. Stimulation of neuronal cells can be performed by means of current or voltage pulses. Intracellular stimulation during patch clamp measurements is usually done in the current mode, since the stimulation of neurons in their natural environment also occurs through current inputs to their dendritic tree. For extracellular stimulation both, voltage stimulation (Bonifazi and Fromherz, 2002; Gross et al., 1993; Taketani and Baudry, 2006) and current stimulation (Buitenweg et al., 2002a; McIntyre and Grill, 2002; Wagenaar and Potter, 2004) methods have been widely used. A detailed description of stimulation methods for cell cultures with micro-electrode arrays is given in Wagenaar et al. (2004).

Biphasic voltage-controlled pulses with the first part of the pulse having a positive voltage sign constitute effective stimuli, since the sharp downward voltage transition between the positive and the negative phase corresponds to a strong negative current pulse. The pulse amplitude and its duration determine the stimulus efficacy (Buitenweg et al., 2002a; McIntyre and Grill, 2002; Wagenaar and Potter, 2004). Another important parameter is the electrode impedance, since it determines the current for a given voltage transient and the voltage on the electrode for a given current transient. Impedance values for the type of platinum electrodes used here have been measured and published in Heer et al. (2004).

An 8-bit flash D/A-converter has been implemented for stimulation. An arbitrary stimulation signal with a sampling frequency of 60 kHz can be applied to any subset of electrodes. Since stimulation voltages typically are in the range of 1 V, and recorded signals are below 1 mV, any stimulation will saturate the respective signal conditioning chain. To allow for immediate read-out after stimulation, the stimulating electrodes can be reset (“reset” in Fig. 3) to their operating points. All the on-chip circuitry has been described in Heer et al. (2006) in great detail.

2.1.2. Digital on-chip circuitry

The large data volumes and the desired low-latency time require high-performance communication and signal processing capabilities. Digital logic components distributed on the MEA ASIC and on the FPGA connect the A/D-converters to the PC.

The on-chip digital logic runs at 1.6 MHz and 5 V and serves two purposes: firstly, it performs control tasks like multiplexing, electrode selection for stimulation, reset of single electrodes, and it contains the successive-approximation registers of the A/D converters; secondly, it provides the chip interface to the FPGA. In the architecture shown here, this digital interface uses 16 lines for the data readout, 2 lines for stimulation and control data, and additionally, a clock, a reset and a frame sync line. All signals are synchronous to the 1.6 MHz clock chip.

2.2. FPGA

To manage the data output rate of 3.2 MB/s and the input rate of 0.4 MB/s, an FPGA running at 48 MHz in conjunction with an USB 2.0 interface chip and 1 MB SRAM (CESYS GmbH, 2005) has been utilized. I/O buffering and digital signal processing like low-pass filtering and event detection are implemented on the FPGA to reduce the data volume transmitted to the PC.

2.2.1. Event detection

Event detection is mandatory since data volumes are beyond the bandwidth of a human observer. In Lewicki (1998) a review of common approaches is given. The approach as presented in this paper is realized as a two-stage processing:

(1) During event detection, events are isolated from the continuous flow of the data. In this context an event is a segment of recorded data that potentially contains a single spike, multiple (possibly overlapping) spikes, a spike train or barrage.

(2) During spike discrimination, events are further processed to isolate spikes, to perform a unit separation and to resolve overlaps.

The benefits of performing event detection on the FPGA as opposed to a PC are two-fold. First, the PC is relieved from approximately 32 million arithmetic operations per second. Second, the load on the I/O system is decimated, which is a prerequisite for a future realization of larger arrays. The freed PC resources can be utilized for more complex tasks like on-line spike sorting, which are hard to perform in hardware. Additionally, the latency time of the whole system is reduced.

2.2.1.1. Algorithm. Two statistical variables per electrode are needed for event detection: the moving average of the signal, \( \bar{x}_i \), and the moving variance, \( \sigma_i^2 \), where \( i = 0, \ldots, 127 \) denotes the electrode number. As illustrated in Fig. 4, events are detected by comparing the absolute value of the highpass-filtered signal, \(|x_i - \bar{x}_i|\), with four times the standard deviation, \( \sigma_i \)

\[
(x_i - \bar{x}_i)^2 > 4^2 \sigma_i^2.
\]

(1)

The mean signal and the variance are continuously updated by numeric infinite-input-response (IIR) low-pass filters with:

\[
\bar{x}_i(k) = (1 - \varepsilon_{\bar{x}}) \times \bar{x}_i(k - 1) + \varepsilon_{\bar{x}} \times x_i(k)
\]

and

\[
\sigma_i^2(k) = (1 - \varepsilon_{\sigma}) \times \sigma_i^2(k - 1) + \varepsilon_{\sigma} \times (x_i(k) - \bar{x}_i(k))^2,
\]
The off-chip random-access memory is interfaced via a tri-state bus to two entities. Bus arbitration and address generation have been omitted for the sake of clarity.

![Image](image.png)

**Fig. 4.** Event detection exemplified using a chicken cortical-neuron spike. The central blue stripe indicates the 4σ-region. The onset of an event is registered and reported to the PC when the signal leaves the 4σ-region for longer than a defined minimum time (four samples = 0.2 ms). The end of an event is registered, when an overhang-long sequence of sub-threshold values is found, where overhang can be defined to, e.g., 16 samples = 0.8 ms. Then, a second report is sent to the PC, which also includes the event wave fragment.

respectively. Using \( \sigma^2 \) instead of \( \sigma \) circumvents square-root calculation. Sensible coefficients are \( \varepsilon \in \{2^{-n} | n \in \{6, 7, 8, 9\}\} \), which allows to use shift operations instead of multiplications and to keep the requirements concerning the fixed-point precision manageable. The required fixed-point accumulator width is determined by the input/output width (8 bit for \( \bar{x}_i \), 6 bit for \( \sigma_i^2 \)) plus \( n \) post-decimal-point digits. Hence, higher \( n \)-values require higher fixed-point precision and, consequently, more storage space. The corresponding time constants, \( \tau \), with \( \varepsilon = 2^{-n} \) are given by

\[
\tau = \frac{1}{f_{\text{sample}} \ln(1 - 2^{-n})},
\]

with \( f_{\text{sample}} \) being the sampling frequency.

For \( \varepsilon_2 = 2^{-6} \) and \( \varepsilon_8 = 2^{-8} \), the respective time constants are 3.17 and 12.78 ms. By using only every second value to update \( \bar{x}_i \) and \( \sigma_i^2 \), \( f_{\text{sample}} \) is effectively cut in half (10 kHz instead of 20 kHz). This way, the time constants are doubled to \( \tau_2 = 6.35 \) ms and \( \tau_8 = 25.55 \) ms. Good results have been obtained using these settings.

Additionally, the following special cases are accounted for:

- during stimulation; updating of \( \bar{x}_i \) and \( \sigma_i^2 \) is interrupted, otherwise events immediately following a stimulation cannot be detected, primarily due to increased \( \sigma_i^2 \)-values;
- \( \sigma_i^2 \geq 64 (= 234 \, \mu V^2) \); no events are generated because the high level of noise indicates a defective channel;
- \( \bar{x}_i \geq 252 (= 480 \, \mu V) \lor \bar{x}_i < 4 (= -480 \, \mu V) \); no events are generated because the mean is too close to the rails and, therefore, no reliable calculation of \( \sigma^2 \) is possible.

The voltage values in brackets are valid for a gain of 1000 and an ADC bias of 1 V.

2.2.1.2. Register-transfer-level implementation. **Fig. 5** shows how the event detection is embedded into the FPGA core architecture. Due to memory and bandwidth restrictions, the event length is fixed to a window of 256 samples, which corresponds to 256/20 kHz = 12.8 ms. One hundred and twenty-eight of such look-back windows are maintained in the form of circular buffers on the external memory amounting to 32 kB of memory, which clearly exceeds the Xilinx Spartan II XC2S200 on-chip block random-access memory (RAM) of 7 kB.

Since all circular buffers \( B_i, i = 0, \ldots, 127 \), are synchronously used, it is sufficient to maintain only one buffer pointer, \( \text{ptr} \), of 8 bit and to calculate the specific 15-bit buffer pointer using the relation \( \text{ptr}_i = \text{ptr} + i \times 256 \). The pointer, \( \text{ptr} \), is incremented after each completed frame, i.e., after 128 samples. In the rare case of an event length larger than 256 samples, the beginning of the event will be clipped.

The respective 128 tuples of \( (\bar{x}_i, \sigma_i^2) \), where \( \bar{x}_i \in [0, 255] \) and \( \sigma_i^2 \in [0, 63] \), are encoded using 14-bit fix-point numbers, where there are 6 and 8 binary digits to the right of the decimal point, respectively, and are then stored in the distributed on-chip RAM of the FPGA.

As can be seen in **Fig. 6**, event detection is performed on the fly; samples are processed in the order delivered by the neu-
Fig. 6. Finite state machine (FSM) as implemented on the FPGA to perform event detection. Annotations aside the state transition arrows have two meanings: if preceded by an asterisk, they represent a post condition, otherwise they indicate the premise for the respective state transition. The FSM may be structured into two parts: an outer FSM and a sub FSM. The outer FSM (blue ovals) manages the look-back ring memories, IIR filters for the floating average and variance calculation and retrieval of the per-channel sub-FSM state. The sub FSM (green ovals) performs the actual event detection per channel. The sub-FSM state (3 bit, one-hot encoded) and the counter (5 bit) need to be stored per electrode. Whenever a new sample $x_i$ arrives, the sub state is loaded into the sub FSM and executed into its next state.

In order to compensate for RAM latencies induced by I/O operations, a 32-value-deep first-in first-out (FIFO) buffer is placed upstream of the event detection. Upon availability of a new sample, $x_i$, from channel $i$, the respective ($\bar{x}_i$, $\sigma_i^2$) tuple is retrieved and updated with the new sample, $x_i$, and the $i$th sub-FSM state is retrieved from the memory, all within 2 clock cycles. Then the sub FSM is executed (1 clock cycle) and, potentially, an event is generated (10 clock cycles for a hot report, $\sim520$ clock cycles (depending on RAM bus availability) for the transmission of an event wave fragment$^1$), the new sub-FSM state is stored (one clock cycle), and the FSM then returns into the idle state and is ready for the next sample. Hence two kinds of event reports are sent to the PC: one is sent following 4 consecutive values (200 $\mu$s) above threshold and a second one, which includes the event wave fragment, is sent after a number of values below threshold, e.g., after 16 consecutive values (0.8 ms).

Benchmarking of the FPGA-design without the USB data transmission (which is not the limiting factor) in Modelsim (Mentor Graphics Corp., 2004) revealed a bandwidth sufficient to extract $50 \times 10^3$ events per second from 1024 channels. This spike rate corresponds to the order of magnitude that a neural burst on 1024 channels can reach during a network-wide burst. For 128 channels there is a margin of one order of magnitude. In the case of a buffer overflow, events are dumped, and an error flag is activated and reported to the PC to indicate a loss of events.

$^1$ The external RAM is used for the event detection and the USB 2.0 I/O data buffer. A bus-arbitration unit and an address-generation unit are implemented for the RAM bus assignment.
The major limiting factor is the bandwidth of the off-chip FPGA RAM, where the signal look-back windows are stored.

2.2.2. Stimulation

As described in Section 2.1, the architecture of the chip allows for selecting an arbitrary set of electrodes for stimulation with the same stimulus waveform.

Reconfiguration of the stimulation subset and reset of the filters proceed at the speed of the readout of the array, since the addressing scheme for read-out is also used for programming the electrodes; whenever an electrode is sampled for A/D conversion, either its “reset” state or its “stimulation” state may be reprogrammed. Consequently, the array may be reconfigured from the stimulation of a certain set of electrodes to the stimulation of any other set of electrodes within 50 μs. In case that simultaneous stimulation of electrodes with different stimulus waveforms is desired, two consecutive simulations with a pause of 50 μs need to be performed. When a stimulus of, e.g., 150 μs length is applied, the two stimuli are 200 μs apart, which, in most cases, can be considered to be simultaneous with respect to the time scale of neuronal events.

Requirements for the stimulation control include:

- accurate timing: one frame or 50 μs;
- different stimulation electrode sets and various stimulus waveforms;
- fast sequences of stimulations to facilitate time-coded stimulation.

Since the USB link latencies do not allow for a real-time issuing of commands, the FPGA needs to provide the infrastructure for accurate timing. A stimulation is specified by

- start_frame, relative or absolute (16 bit);
- electrode_selection (1 bit/electrode = 128 bits);
- reset_selection (1 bit/electrode = 128 bits);
- waveform (8 bit, 60 kHz, variable length).

A command providing the above-specified data is issued to the FPGA and enqued in the FPGA’s stimulation FIFO. The start frame triggers one of three possible timing modi immediate, absolute, or relative, the first two of which are self-explanatory, whereas relative means “stimulate n × 50 μs after the last stimulation.” The stimulation state machine walks through the following states:

Wait: wait for data. start_frame triggers one of three possible timing modi:

- start_frame = 0 ⇒ immediate stimulation
- 0 < start_frame < 2^15 ⇒ relative stimulation: “continue start_frame frames after last stimulation”
- 2^15 ≤ start_frame ⇒ absolute stimulation: “continue when the lower 15 bits of the global frame counter (50-μs clock) and start_frame match.”

Program reset: at each electrode, program the filter reset switches according to reset_selection

Program stimulation buffers: at each electrode, program the stimulation buffers’ on/off and stimulation switch according to electrode_selection

Replay waveform: stimulate according to waveform

Program stimulation buffers: switch all stimulation buffers off and disconnect them from electrodes

Program reset: switch off all reset switches, back to Wait

2.2.3. Stimulation artifact suppression

Fig. 7 shows the efficiency of the on-chip reset function for stimulation artifact suppression. In Fig. 7, left pane, bipolar pulses (220 μs overall duration) of different amplitudes (±0.15 and ±1 V) in saline solution have been applied, while the reset has been operational (red trace ±0.15 V; green trace ±1 V) or not (blue trace ±0.15 V, black trace ±1 V). The recordings from the stimulation electrode are shown. The readout filter for the red and green trace was reset 50 μs before the stimulation pulse was applied and kept until 50 μs after the stimulation pulse was finished. Without reset, it takes 20–100 ms (depending on the stimulation amplitude) for the recording circuitry to return to the measurement range. The inset shows a close-up of the initial 20 ms at better temporal resolution. With reset, it takes less than 5 ms for the recording circuitry to return to the measurement range. It is worth mentioning that a stimulation sequence always ends with a stimulation of a value close to the equilibrium potential of the electrode so that the electrode quickly returns to its equilibrium potential before stimulation.

A similar experiment on a chip with rat cortical neurons is documented in Fig. 7, right pane. The stimulus is a biphasic rectangular pulse of 1.1 Vpp and 110 μs duration. The measurement shows action potentials recorded on the stimulation electrode 9 ms after stimulation. Action potentials have been recorded less than 5 ms after stimulations on the same electrode. The remaining artifact signal shows very slow temporal characteristics and can readily be removed by digital highpass filtering. The other electrodes that do not provide a stimulation signal record the stimulation signal as a spike: the signal is sufficiently attenuated in the culture bath so that the signal is too weak and too short to impact the respective highpass filters. Fig. 8 shows how a direct stimulation response is reliably recorded 2 ms after stimulation on a neighboring electrode at 250 μm distance.

2.3. Software

The software is a crucial component of the system. It organizes the USB data handling, the spike sorting, the data visualization, the stimulus generation and the automated feedback generation. Low latency time and high data throughput are key requirements. It is implemented in C++ using the Qt library (Trolltech Inc., 2007) for the graphical user interface and runs on Linux.
Fig. 7. Left: efficiency of the on-chip reset function for stimulation. Bipolar pulses (220 μs overall duration) of different amplitudes (±0.15 and ±1 V) in saline solution have been applied, and the reset has been operational (red trace ±0.15 V; green trace ±1 V) or not (blue trace ±0.15 V, black trace ±1 V). Without reset, it takes 20–100 ms for the recording circuitry to return to the measurement range. The inset shows a close-up of the initial 20 ms at better temporal resolution. Right: stimulated signals from cultured rat cortical neurons on the stimulation electrode. The stimulus is a biphasic rectangular pulse of 1.1 V pp and 110 μs duration. The remaining slow drift can readily be eliminated by digital high-pass filtering. Both measurements have been performed with 20-μm-diameter platinum electrodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Fig. 8. Stimulated response of cultured rat cortical neurons to a single bipolar stimulation pulse of ±800 mV amplitude and 50 μs duration. Left: post-stimulus time histogram for 142 stimulation responses on one electrode. The peak around 2 ms represents the direct response with near-to 100% probability. Right: stimulated action potentials recorded from an electrode at 250 μm distance from the stimulated electrode (raw data, as delivered by the chip). The first action potential at around 2 ms after stimulation occurs with 98% reliability and causes the peak around 2 ms in the post-stimulus time histogram on the left hand side. The stimulation does not impact the functioning of the highpass filter (see Fig. 7).

There are three concurrent, widely independent tasks: the I/O-handling, the event processing, and the user interface. Therefore, the program named Neurotalker runs in three main threads as sketched in Fig. 9.

2.3.1. I/O

The I/O thread interfaces the USB subsystem to the main loops of the program. Since USB is a pure master-slave bus (USB Implementers Forum, 2005), and since it relies on block transfers, a further protocol layer is required to have the capability to transfer data packages of arbitrary size at any defined time.

A combined software and VHDL package, named ce-usb-kit that provides this additional protocol layer has been developed and encapsulated into an independent library containing both VHDL and C code. It is available from the sourceforge project (Hafizovic, 2005). The ce-usb-kit provides a VHDL code that includes

- an external-RAM I/O buffer,
- a receiver finite-state machine that decodes data packages sent by the PC,
- a sender finite-state machine that encodes data, performs padding and calculates an 8-bit cyclic redundancy checksum (CRC).

The 8-bit CRC is appended to each transferred data block and is very useful for testing purposes, especially with respect to the communication between FPGA and FX-2 chip. The software part of the ce-usb-kit runs exclusively in the user-space and provides routines to handle the USB device. The communication between PC and FX-2 relies on exchanging data with the FX-2.
firmware as provided by CESYS via I/O-controls. Data are sent in packages of multiples of 512 bytes, i.e., after a bulk transfer request for 4096 bytes is issued, the FX-2 has to deliver the requested amount of data. In the case that less data are available than requested, the excess volume is padded. The polling interval and the package size are dynamically regulated to ensure an efficient data transport. The output buffer on the FPGA is dimensioned to 32 kB.

2.3.2. Modules
A number of universal modules provides the infrastructure to perform standard tasks. The list of modules includes:

- **Signal plotter**: plots data of arbitrarily selectable channels
- **Raster plot**: plots events in a raster plot
- **Stimulator**: stimulates via an arbitrary set of electrodes
- **Recorder**: stores the channel data and the events into files
- **Feedback**: triggers actions, i.e., issues stimulation signals or a beep upon events matching certain criteria
- **LUT programmer**: programs the Look-up table (LUT) that controls the multiplexing of the A/D converters to the electrodes.

Sampling rates of up to 160 kHz are possible, when only one electrode per row is read out.

I/O modules include the USB-IIo and the replayer. The replayer streams in data that have been previously recorded with the recorder module.

It is possible to seamlessly add further modules owing software design, which includes techniques like instantiation by name (Taschini et al., 1999). Dedicated modules for special experiments can be generated and added.

2.4. Fabrication

The chip has been fabricated in industrial 0.6-µm CMOS-technology (XFAB, 2007). The platinum electrodes and the additional passivation layers have been fabricated in house using a 2-mask postprocessing as described in (Heer et al., 2004). Thereby, a variety of CMOS MEAs differing in electrode diameters (10–40 µm) and locations (pitch 50–500 µm) can be fabricated on the same CMOS chip (Fig. 2).

The chip operating conditions are the same as for conventional cell incubation, i.e., 95% rel. humidity, saline solution as cultivation medium and a temperature of 37°C over extended time (chips have been successfully used for more than 9 months), so that a biocompatible package and inert electrical contacts are required. The array chip has been mounted and wire-bonded to either a ceramic package or onto a PCB, both featuring gold contacts. Afterwards, a glass ring has been mounted, and the bond wires have been encapsulated with epoxy resin (EPOTEK 302-3M (Epotek, 2007)). Fig. 10 shows a ready-to-use packaged chip. The chip has been glued to a PCB and wire-bonded. The glass ring has been mounted on top, and the bond wires have been covered with a biocompatible epoxy resin, EPOTEK 302-3M (Epotek, 2007).

Synplify Synplicity (2005), and, finally, the layout with Xilinx ISE Xilinx (2005).

3. Results

Electrical testing of the on-chip bandpass filter revealed a tuning range for the lower corner frequency between 1 Hz and 1 kHz and for the upper corner frequency between 1 and 30 kHz. The midband gain of the bandpass was measured to be 19.12 dB. The total-equivalent-input noise was found to be 5.9 µVrms (10 Hz–100 kHz) at a gain of 1000. This is comparable to the thermal noise of the platinum electrodes in physiological saline solution, which was measured to be 3.19 µVrms for 10 µm × 10 µm and 2.16 µVrms for 40 µm × 40 µm electrodes in a band between 10 Hz and 100 kHz. The total noise in the measurements is dominated by the background activity of the neuronal culture, which is on the order of 10–30 µVrms in a dense culture (see Fig. 8).

The measured power consumption was 120 mW at 5 V supply, 20 mW of which are dissipated within the electrode array. The increase of the temperature within the electrode array was measured to be less than 1°C so that cooling is not required.

3.1. Closed-loop stimulation latency

For closed-loop experiments involving stimulations triggered upon the detection of spike events, a quick reaction of the system is desirable. We have performed 1000 measurements of the time between the detection of an event on a specific electrode and a stimulation initiated on another electrode. The time measurement is started after recording four consecutive samples (200 µs) above threshold and includes the steps of

- transmission of the event report, which includes the electrode number and a time stamp, from the FPGA to the PC via USB,
- filtering of the reports (for electrode number and hold-off time) in the software,
- generation of the stimulation wave form,
- USB transmission of the stimulation wave and instructions from the PC to the FPGA,
- configuration of the electrode array for stimulation.
Fig. 11. Histogram of closed-loop stimulation latency times for executing spike-triggered stimulations. The stimulations are triggered by the onset of a spike (after 0.2 ms above threshold) and the trigger does not include any spike sorting.

Fig. 11 shows the result of 1000 latency-time measurements. The mean latency time is 1.2 ms, 98% of the stimulations were initiated in less than 1.35 ms with the shortest and longest latency times amounting to 0.8 and 2 ms.

The configuration used here allows for issuing and executing a stimulation of several 100 μs duration even before the action potential that has triggered the stimulation has terminated. In order to include spike sorting in the feedback loop, it is necessary to wait for the termination of the detected spikes or activity, which adds the overall duration of the spike as well as the time needed for spike sorting to the latency of ~1.2 ms.

3.2. Stimulated activity in rat cortical-neuron cultures

As a demonstration of precise spatio-temporal stimulation patterns and event detection, we present the following experiment inspired by the liquid-state-machine concept as presented in (Maass et al., 2002, 2003). The underlying idea is to use the stimulation responses of arbitrary excitable media for information processing. As shown in Fig. 12 the model consists of a stimulator, a neuronal network on a neurochip, an event detector, a leaky integrator and a classifier.

3.2.1. Methods

3.2.1.1. Stimulation. Two stimulation spots a and b have been defined, where a represents a stimulation on electrode 52 and b a stimulation on electrodes 96–104, both with one biphasic rectangular pulse of 1.1 Vpp and 110 μs duration. In Fig. 2 location a corresponds to the fifth electrode in the third row from the bottom and location b to the sixth row of electrodes from the top. For this experiment, the stimulation pulse amplitude of 1.1 Vpp was found to be the lowest voltage that reliably evoked action potentials.

Based on these two stimuli, the two spatio-temporal patterns aba and bab have been stimulated with an inter-stimulus period of 40 ms. The inter-stimulus period of 40 ms was chosen to be on the order of the neuronal time constant, which is in the range of 20–100 ms (Kandel et al., 2000, p. 222), to allow for temporal summation to occur. The period in between aba and bab stimulations, the inter-pattern period, was chosen to be 2.3 s so that it is short enough to suppress spontaneous electrical activity of the culture. Without stimulation interference, spontaneous activity occurred at intervals of 3–5 s.

As described in Section 2.2.2, the stimulation patterns are generated on two levels: the inter-stimulus period of 40 ms is generated on the FPGA with 50 μs accuracy, and the inter-pattern period of 2.3 s is generated on the PC with millisecond accuracy.

With regard to potential stimulation-induced damage to the neurons, we noticed that after more than 200 stimulations on the

Fig. 12. Schematic of a “liquid state machine” implementation. Inputs, either aba or bab, are stimulated in the neuronal culture. From the neuronal culture response, liquid states are calculated by leaky integration of the time stamps; 200 ms after stimulation, the liquid state is sampled and classified in a support vector machine.
same electrode, the cells could still be stimulated without any observable response signal deterioration.

3.2.1.2. Neuronal culture. Hippocampal tissue was extracted from new-born Sprague–Dawley rats and dissociated by trypsinization (0.25% trypsin/EDTA) and gentle trituration. Cells were plated on laminin (10 μg/ml) and poly-L-lysine (10 (g/ml) coated CMOS chips at a density of 15,000 cells/mm² and held in Neurobasal medium (Gibco) containing B27 supplement (2%, Gibco), fatty acid supplement (0.1%, Sigma), lipid mixture (0.1%, Sigma), alanyl-glutamine (2 mM, Gibco) and sodium pyruvate (1 mM, Sigma) (Fig. 13). Measurements were carried out after 37 days in vitro. For all biological measurements, the packaged chip, as shown in Fig. 10, (20-

![Fig. 14. Raster plot and network-wide firing rate of the aba stimulation pattern repeated every 2.3 s.](image1)

![Fig. 15. Plot of the outputs of the leaky integrators, x(t), also referred to as liquid states, upon a stimulation. The left plot shows an aba stimulation, the right plot shows a bab stimulation. The gray background indicates the time span, during which the three stimulations on the a and b sites at 0, 40 and 80 ms take place; 200 ms after the first stimulation, the 128 leaky integrators are sampled, and the obtained vector is then classified.](image2)
μm-diameter electrodes, pitch 250 μm) was mounted onto a heated PCB (37 °C, manually controlled heating power delivered by an adjustable power supply to a constantan-wire meander with an aluminum heat spreader below the PCB) in ambient atmosphere. In order to obtain better defined and more physiological measurement conditions, it is planned to perform future measurements in a dry incubator with controlled CO2 level and temperature as reported on by Potter and DeMarse (2001).

3.2.1.3. Readout. From the detected events (described in Section 2.2.1, example shown in Fig. 14), 128 sets of timestamps, \( x_i \), have been generated, where \( i = 0, \ldots, 127 \) designates the channel. From each of these sets, \( x_i \), a liquid state function

\[
x_i(t) = \sum_{t_i / t_{\text{inc}} < t < t_{i+1}} e^{-(t-t_i)/\tau} = \sum_{t_i < t < t_{i+1}} e^{-(t-t_i)/\tau}
\]

has been synthesized, where \( \tau = 20 \text{ ms} \) is the leak-time constant and \( t_{\text{cutoff}} = 40 \text{ ms} \). \( t_{\text{cutoff}} \) explicitly limits the “memory” of the readout function. For each stimulation experiment, \( j \), a time snapshot of the liquid state, \( x_j = x(t_{\text{stimulus},j} + 200 \text{ ms}) \), has been calculated (Fig. 15). These 128-dimensional response vectors represent the neuronal network response to the stimulation pattern that was initiated 200 ms before.

A direct information transfer from the stimulator to the readout that might bypass the neuronal network has been excluded by choosing the \( x_j \)-sample time as \( t_{\text{stimulus},j} + 200 \text{ ms} \), with \( t_{\text{cutoff}} = 40 \text{ ms} \), and by having the last stimulation at \( t_{\text{stimulus},j} + 80 \text{ ms} \). With these settings, there is a time span of \( 200 \text{ ms} - t_{\text{cutoff}} - 80 \text{ ms} = 80 \text{ ms} \), where the neuronal network is the exclusive mediator of any information. Thereby, the function of the neuronal network in the liquid state machine is defined.

Classification of the responses, \( x_j \), with respect to the preceding stimulation pattern \( \text{aba} \) or \( \text{bab} \) has been performed off-line using Mathematica (Wolfram Research, 2005) and a support vector machine package (Nilson et al., 2006). For the presented experiment, a total of 44 stimulation–response pairs, 18 \( \text{aba} \) stimulations succeeded by 26 \( \text{bab} \) stimulations, have been recorded. From each of the two classes the first 30% (6 and 9) of the pairs were used to generate a weight vector, \( w \in \mathbb{R}^{128} \), and a bias \( \text{bias} \in \mathbb{R} \). With a linear kernel being used, \( w \) can be seen as the normal vector of a 128-dimensional hyperplane given by \( x \in \mathbb{R}^{128}: w^T \cdot x + \text{bias} = 0 \), where \( T \) denotes the transposition that separates the \( \text{aba} \) and \( \text{bab} \) responses. Hence, the sign \( (w^T \cdot x + \text{bias}) \) represents the classification of \( x_j \).

3.2.2. Results

As can be seen in Fig. 16, the respective first 30% of the stimulation-response pairs that have been used as training set can be correctly classified and are separable. The remaining two thirds of each data set, in total 29 responses, constitute the test set and have been classified according to the weight vector \( w \) and bias as obtained from the training based on the first 30% stimulation–response pairs. From the test set, 83% (24 of 29) have been correctly classified. 75% (9 of 12) of the \( \text{aba} \) test set and 88% (15 of 17) of the \( \text{bab} \) test set have been correctly classified. Fig. 16 shows the respective results.

3.2.3. Discussion

This first experiments indicate that information processing with dissociated neuronal networks using the liquid state machine model might be possible. There are still many issues to be addressed, e.g., the sequence of the \( \text{aba} \) and \( \text{bab} \) patterns should be random so that the effects of other transient changes in the neuronal culture can be eliminated. Also, the measurement time needs to be extended by performing the measurements in an incubator to be able to better characterize each of the stimulation sites \( a \) and \( b \). These issues and other possible shortcomings will be addressed in future work.

4. Conclusion

The presented signal processing chain reaches from the cell culture, which lives on top of a CMOS-based microelectrode chip, to the C++ control and analysis software. The latter enables a bidirectional communication with electrogenic cell cultures...
in vitro. The combination of the CMOS chip with an FPGA-based device allows to perform online event detection and, upon demand, full-depth transmission of all 128 20-kHz signals. The event detection is implemented on the FPGA running at 48 MHz FPGA-clock speed, is fully self-adjusting and requires no user inputs. The data volume transferred to the PC is effectively decimated by relocating the event detection from the PC to the FPGA. The system is capable of identifying an event and triggering a stimulation within less than 2 ms after the event has been detected, so that it may be considered real-time with respect to neuronal networks cultured on the chip.

Acknowledgements

The authors thank Prof. Henry Baltes for sharing laboratory resources and for his ongoing stimulating interest in their work. The authors are indebted to Frauke Greve, ETH Zurich, for help and advice on neuronal cell cultures and Beat Hofstetter for his contribution to the Neurontalker software. Prof. Wolfgang Maass, University of Graz, Austria, is acknowledged for stimulating discussions on liquid state machines and liquid state classification. Funding has been generously provided by the European “Information Society Technologies” “Future and Emerging Technologies” program, and the Swiss Bundesamt fuer Bildung und Wissenschaft (BBW), contract number IST-2000-26463.

Appendix A. Supplementary data


References

Ashenden PJ. The designer’s guide to VHDL. 2nd ed. Morgan Kaufmann; 2002.


Wolfram Research, I., 2005. Mathematica 5.2. Champaign, USA.

XFAB, 2007. 0.6 µm process, triple-metal, double-polysilicon CMOS process. Dresden, Germany.

Xilinx, 2005. Xilinx ISE 7.1. San Jose, California, USA.