College of Engineering
TEMPLE UNIVERSITY®

Architecture Design for a Neural Spike-Based Data Reduction Platform Processing Thousands of Recording Channels

Ph.D. Dissertation Proposal

Department of Electrical and Computer Engineering

Submitted by:

Nashwa Elaraby

Advisor:

Dr. Iyad Obeid

Committee Members:

Dr. Dennis Silage

Dr. Joseph Picone

May 1st 2013

Contents

1. Introduction:

What beauty is shown in the preparations obtained by the precipitation of silver dichromate deposited exclusively onto the nervous elements! But, on the other hand, what dense forests are revealed, in which it is difficult to discover the terminal endings of its intricate branching… Given that the adult jungle is impenetrable and indefinable, why not study the young forest, as we would say in its nursery stage.

Santiago Ramón y Cajal (1852-1934)

It is commonly accepted that the information processing in the brain is carried out by large groups of interconnected neurons. Neurons are the cells responsible for encoding, transmitting, and integrating signals originating inside or outside the nervous system. The transmission of information within and between neurons involves changes in the resting membrane potential, when compared to the extracellular space. The inputs one neuron receives at the synapses from other neurons cause transient changes in its resting membrane potential, called postsynaptic potentials. These changes in potential are mediated by the flux of ions between the intracellular and extracellular space. The flux of ions is made possible through ion channels present in the membrane. The ion channels open or close depending on the membrane potential and on substances released by the neurons, namely neurotransmitters, which bind to receptors on the cell's membrane and hyperpolarize or depolarize the cell. When the postsynaptic potential reaches a threshold, the neuron produces an impulse. The impulses or spikes, called action potentials, are characterized by a certain amplitude and duration and are the units of information transmission at the interneuronal level [1]. The discovery of the neuron was a milestone in brain research and paved the way for modern neuroscience, but the brain is yet to yield the vast majority of its secrets.

The current neuroscience research operates at two disconnected levels: The macro- and microscopic levels. The macroscopic level uses imaging techniques like functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) to measure regional changes in metabolism and blood flow associated with changes in brain activity. It captures whole brain activity patterns that allow the mapping of brain regions associated with a particular behavior or task. These techniques lack single-cell details and the requisite temporal resolution to permit detection of neuronal firing patterns. The microscopic level is concerned with investigating how individual nerve cells work, studying their response to stimulation and monitoring the firing rates associated with a certain behavioral output, mental state or motor activity. This can be done using implanted electrodes to record the rates and timing of action potentials. The sparse sampling of neuronal activity monitoring tens to few hundreds of neurons does not give the global view of signaling in neural circuits that can involve millions of neurons.

There is a gap between the two levels, that is believed to entail an answer to the question of how neuron cells collaborate to process information. To fill in the gap, we need the static anatomical map of the brain circuitry describing the synaptic connections within any given brain area, as well as the dynamic map revealing the patterns and sequences of neuronal firing by all neurons over time scales on which behavioral outputs or mental states occur. Hence the aspiration is not only to map the "impenetrable jungle" that Cajal referred to but also to map the dynamical traffic within the jungle and analyze it. Research efforts are conducted to approach that ultimate goal, and along the hard path to achieve it, technological breakthroughs evolved and more are bound to arise. New technologies may include new optical techniques to image in 3D, new capabilities for storage and manipulation of massive data sets, new clinically viable Brain Machine Interfaces to help paralyzed patients and development of biologically inspired computational devices.[2]

Focusing on the Microscopic level, the following are two of the research fields concerned with recordings of the spiking activity of neurons using microelectrode arrays.

(a) Brain-Machine Interface:

Extracting motor control signals from the firing patterns of populations of neurons and using these control signals to reproduce motor behaviors in artificial actuators are the key operations of Brain-Machine Interface (BMI) [3]. The typical neural signal processing pathway as shown in fig.1 is designed to measure the instantaneous frequency of neural action potentials, or spikes. Since any given electrode may sense spikes from multiple neurons, it is typically necessary to sort all detected spikes by wave shape (i.e. by neuron). Firing rates of sorted spikes are typically measured by moving average; these rates can then be used by "decoding" algorithms which use statistical models to correlate spiking activity with behavioral or motor activity in the subject.

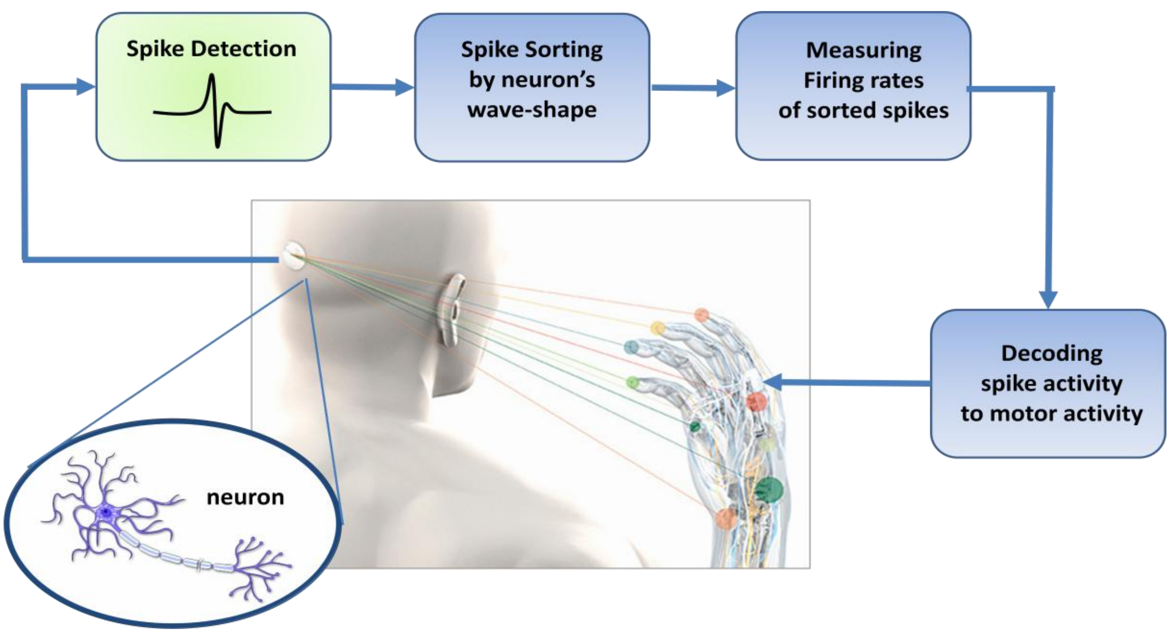


Fig.1.1 Block diagram of the typical pathway of brain machine interface

Hence invasive BMIs rely on the physiological property of individual cortical neurons to modulate their spiking activity in association with movements [3]. These modulations are found to be highly variable from neuron to neuron and from trial to trial. Yet averaging across many trials reveals fairly consistent firing patterns. Based on the hypothesis that the function of neural circuits is an emergent property that arises from the coordinated activity of large numbers of neurons, this phenomenon can be explained. Individual neurons generally form synaptic connections with thousands of other neurons. In distributed circuits, the larger the connectivity matrix the greater the redundancy within the network. Given their distributed connections and their plasticity, neurons are likely to be subject to continuous dynamic rearrangement, participating at different times in different active ensembles [2]. Accordingly both accuracy and reliability of predictions of motor activity improve considerably with increasing the number of simultaneously recorded neurons and decreasing the errors due to individual neuron firing variability. Pursuing this motivation, the number of simultaneously recorded neurons has been approximately doubling every 7 years since 1950's [4]. Standard recording techniques using 704 implantable micro wire arrays have been reported in literature [5]. Recently Nicolelis lab at Duke University announced their success to simultaneously record the electrical activity produced by a population of 1,874 interconnected single neurons at work in a primate.

(b) Brain in a Dish:

At present, the prime methodology for studying neuronal extracellular activity under in vitro conditions is by using substrate-integrated microelectrode arrays (MEAs). This methodology permits simultaneous, long-term recordings (i.e. of up to several weeks) of extracellular field potentials. Correlating MEA recordings with microscopic imaging and stimulations is widely used to study the circuit-connectivity, dynamics and propagation effects in neuron assemblies. It is also used to investigate population coding, activity patterns, plasticity and pharmacological testing on either dissociated neuronal cultures or brain slices of embryonic rats, i.e. the young forests as Cajal described them.

Commercially available MEA systems integrate typically 60–120 microelectrodes of 10–30 μm in diameter with pitches on the order of hundreds of micrometers. Typical neuron soma dimension in vertebrates is few micrometers long and the typical neuronal networks have 10000–50000 neurons, the limited number of electrodes and their rather large pitch results in a substantial spatial undersampling of the overall network activity [6] as shown in fig2.

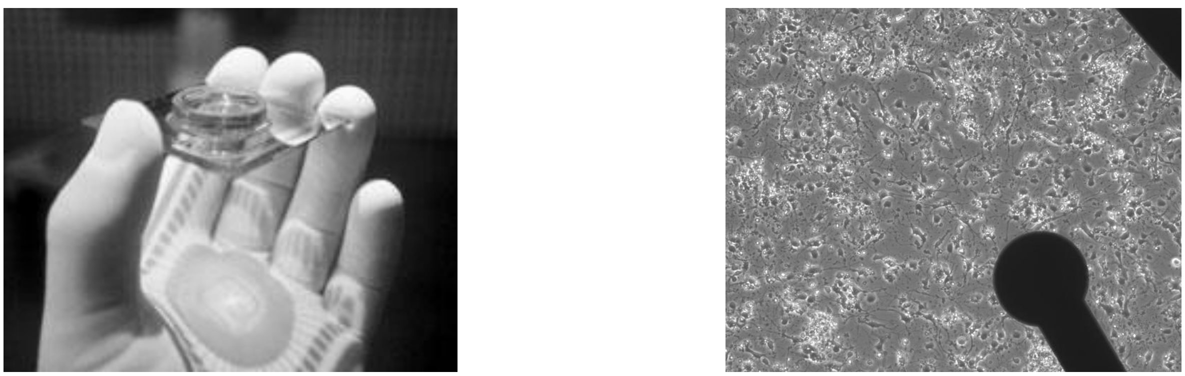


Fig. 2: Substrate-integrated MEA dish. The microscopic image of the electrode (black) and neurons

The development of higher spatial and temporal resolution at low noise levels are prerequisites for opening the perspective to access the network electrical activity at the global and cell levels. Recently, CMOS-based high-density MEAs were developed featuring switching techniques to manage a large number of electrode channels interconnections, multiplexing, amplification, and filtering. Active Pixel Sensor based MEA platform providing 4096 microelectrodes at 21µm inter-electrode separation and 7.7KHz sampling rate has been documented [6].

Considering the ultimate goal of Brain Activity Map [2], the current neuroscience in vivo and in vitro research states and the advancement of high density microelectrode arrays, the migration to monitoring thousands of recording channels at high temporal resolution is achievable.

1.1 Increasing the number of Recording Channels:

More is Different - The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead at each level of complexity entirely new properties appear.

Philip Warren Anderson

The augmentation of the number of recording channels carries different challenges to the neural signal processing system. The primary challenge is the massive increase in recorded data that needs proactive strategies for data transfer, reduction, management and analysis. The implementation of real-time signal processing becomes essential to alleviate huge data storage requirements. The access to a more detailed view of neuronal networks might reveal new properties and challenges pushing for the development of new analyzing tools.

With the continuous advancement of data acquisition systems featuring high-count recording channels, there exists a clear need for a test bed to develop and investigate a more suitable new generation of Neural Signal Processing (NSP) algorithms and computational tools. The platform has to offer programmable flexibility to allow the trial of different new strategies and novel computational techniques as well as rigorous testing for evaluation.

A plausible NSP platform that can handle thousands of recording channels has to provide means of high data transfer. As a numerical example, a NSP platform handling 2560 channels sampled at 31.25 KH at a sample precision of 16-bits must be capable of managing an input data stream of 1.28Gbps. The data transfer interface has to be compatible with high-density neural data acquisition systems [7].

Data reduction based on the sparse nature of the neural signal with respect to time and the redundancy perceived across multiple electrode recordings becomes essential. Spike detection is the essential first step building block that allows the system to deliver only the action potential waveforms, their respective occurrence times and channel ID instead of the entire raw signal. The AP waveforms are then used by an autonomous spike sorter to first distinguish true spikes from false detections, then, to associate each spike to its generating neuron in case of multi-unit recordings. Depending on the performance and inter-electrode spacing, the AP waveforms might be necessary to identify redundancy over multiple recording channels.

The spike detection settings for each channel is independent from the settings of other channels, and hence spike detection over different sites can run in parallel. Applying parallel processing whenever possible limits the overall latency and assists in achieving real time implementation.

The NSP platform has to be fully autonomous and functional under expected Signal-to-Noise Ratios delivered by the data acquisition system. The system must be adaptive to varying noise levels over different channels and over time.

The main objective of the proposed project is to design an experimental test bed that can facilitate dealing with a large number of recorded neurons in real time. It also presents an architecture that performs spike-based data reduction.

1.2. Why consider FPGA?

Ross Freeman (1944-1989) established the leading FPGA developer Xilinx in 1984 and invented a year later the first Field Programmable Gate Array (FPGA). **FPGAs** are programmable semiconductor devices that are based around a matrix of Configurable Logic Blocks (CLBs) connected through programmable interconnects. FPGAs can be configured to implement custom hardware applications and functionalities. Since their invention, FPGAs have evolved far beyond the basic capabilities present in their predecessors, and incorporate hard

Application Specific Integrated Blocks of commonly used functionality such as RAM, clock management, and DSP.

FPGAs are parallel in nature, so different processing operations do not have to compete for the same resources. Each independent processing task is assigned to a dedicated section of the chip and can function autonomously without any influence from other logic blocks.

As integrated circuits grew smaller and maximum toggle rates increased the need for input/output bandwidth exploded. With more hardware resources and faster clock speeds, conventional I/O resources became the bottleneck to FPGA performance. In 2002, Xilinx embedded high-speed serial Multi-gigabit transceivers (MGTs) on their FPGAs and introduced them commercially under the name Rocket I/O. MGTs are Serializers/Deserilizers (SERDES) that allow serial data transmission over differential pairs at speeds of up to 28.05Gbps per lane (see Fig. 1.3). Alternatively, multiple MGTs can be bonded together to form a higher bandwidth interface. Multiple MGTs are integrated above and below the Block RAM columns providing close availability for ingress and egress FIFOs. Rocket IO serial transceivers (see Fig 1.4) are compliant with standard gigabit communication protocols.

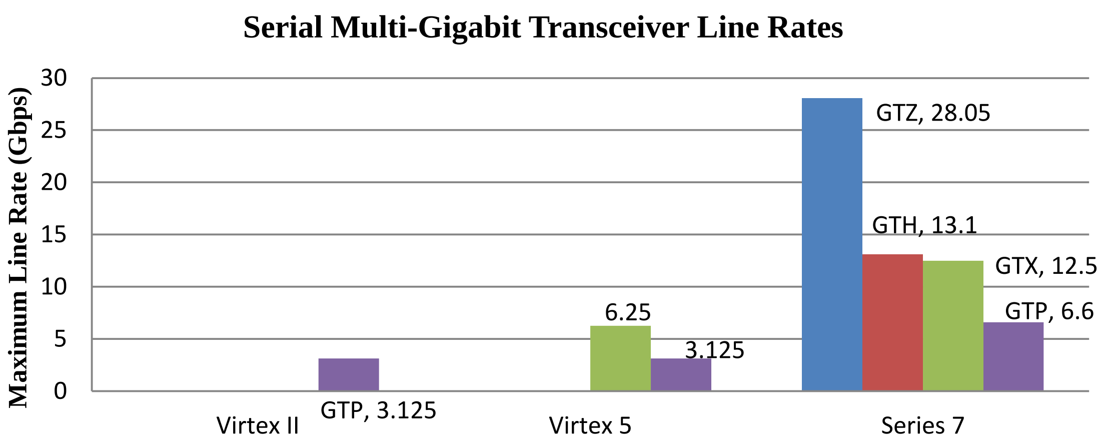


Fig. 1.3. Bar graph presenting the available serial Multi-Gigabit Transceiver line rates.

FPGAs offer massive parallel processing performance, reconfigurable flexibility and superior capabilty of streaming data, and therefore present an appealing hardware implementation solution for a NSP testbed that can handle a large number of similarly strutured parallel channels in real time.

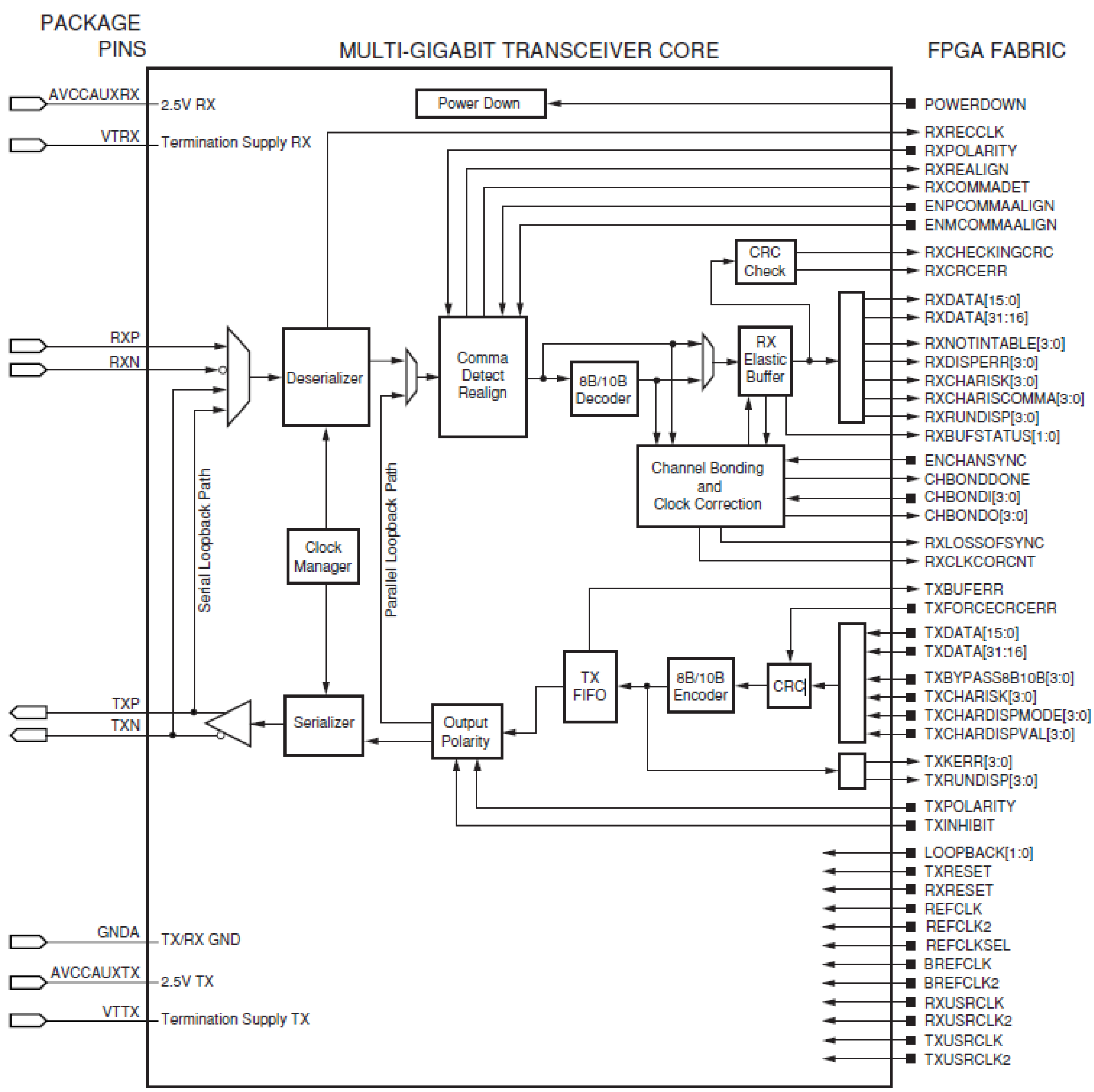


Fig. 1.4: Functional Block diagram of Xilinx® GTP MGT.

2. Background:

Monitoring the interplay of neuronal ensembles in the brain is important for understanding mechanisms underlying memory, learning and behavior. Recently a group of neuroscientists have proposed launching a large-scale, international public effort called "Brain Activity Map" (BAM) Project, aimed at reconstructing the full record of neural activity across complete neural circuits [2]. They describe the neural circuit function as being emergent, meaning that it arises from complex interactions among millions of neurons and that the circuit state is not predictable from responses of individual sparsely sampled cells. They propose the dynamical mapping of the "functional connectome", the patterns and sequences of neuronal firing by all neurons. Correlating this firing activity with both the connectivity of the circuit and its functional or behavioral output could enable the understanding of neuronal codes and the regulation of behavior and mental states. Some of the mental illnesses that could not be understood using single-level analysis, such as autism and schizophrenia, may be possible to explain on an emergent level analysis. Clearly, the benefits of getting the full dynamical picture of the brain will be invaluable to address many questions in neuroscience, but to achieve this vision there is a clear need to develop novel technologies and significant innovations in systems engineering.

At present, population coding is studied either by monitoring the spiking activity of a few hundreds of individual neurons working with intact, living animals or by studying the basics of distributed information processing using cultured neuronal networks. Cultured neuronal networks lack many features of real brain, but they retain others such as developing synaptic connections and exhibiting different patterns of electrical activity [8]. The neural activity cannot be correlated to a behavioral or mental output as *in vivo*, but it can be correlated to a structural connectome and to stimulation patterns. Advancement in micro-electrode array technology and multi-photon microscopy, has made it possible that every cell in a cultured monolayer network of dissociated neurons can be observed , monitored, stimulated and manipulated with temporal resolution in the submillisecond range, and spatial resolution in the submicron range, in a non-destructive manner [8]. Currently, such detailed analysis is not feasible in living animals, or even brain slices, but it remains an open question however, whether any of the processing done by cultured neurons is relevant to that carried out by intact brain.

This chapter serves to present efforts from a number of research groups to upgrade the recording capabilities of neuronal activity to higher spatial and temporal resolution across a large-scale neuronal ensemble to approach the model of *in vivo* brain. It will review some of these efforts reported on the data acquisition level. With the increasing number of recording sites, the chapter also discusses architecture design considerations at the spike detection level.

2.1. Multi-electrode Arrays:

Multielectrode arrays or microelectrode arrays are data acquisition devices that contain multiple plates or shanks through which neural signals are acquired, basically serving as neural interfaces that connect neurons to electronic circuitry. The signal then passes through amplification and filtering to remove some of the background noise. MEAs can be classified into two groups: implantable MEAs, used in vivo, and non-implantable MEAs, used in vitro. Using advances in multisite microelectrode array fabrication techniques varying shape and recording capacity of the electrodes, it is possible to record the activity of tens to hundreds of neurons in parallel [9]. Integrated microelectronic circuits were applied to enable the transition to even higher recording capacities [10]. Development of *in vivo* and *in vitro* multi-electrode probes share many of the same hardware and data analysis problems and mutually contribute to the advancement of the state of the art.

2.1.1 *In Vitro* Micro-Electrode Arrays:

Multi-electrode array culture dishes allow simultaneous recoding from and stimulation of neurons. These wired Petri dishes are also called planar electrode arrays [2]. Early microelectrode developments by Gross [11], Wise, Meister and others paved the way for enabling chronic multi-single-cell recording. They were able to record neural spike potentials with good fidelity from a few tens of neurons.

MEA's have become commercially available just within the last decade. MEA systems capable of recording at least 60 electrodes are produced by MultiChannel Systems of Germany, and Panasonic of Japan. Guenter Gross supplies MEAs that can be used with multi-electrode processing hardware and software made by Plexon Inc [8]. MEAs typically consist of less than 100 planar metal electrodes on an insulating glass substrate with a diameter > 30μm and a pitch >100μm. For commercially available MEAs, amplification and filtering are realized by discrete off-chip components [6].

Considering the dimensions of neurons, which range from below 10μm for vertebrates up to 100μm for invertebrates, the development of high-density arrays was needed to acquire more details from cell-based biological experiments on brain slices and to elucidate the contribution of individual cells to collective network. An advanced multi-electrode array system has been developed to study how the retina processes and encodes visual images. This system can simultaneously record the extracellular electrical activity from hundreds of retinal output neurons and consists of 512 planar microelectrodes with a sensitive area of 1.7 mm^2 and a noise level of a few μV [13]. However, some brain structures, such as hippocampus or cerebral cortex, extend over distances of many millimeters [14]. To record from these larger structures, an increased density of electrodes and a larger array would be required in order to fully analyze all the neurons of interest.

CMOS-based devices presents several advantages for managing a large number of electrode channels' interconnections, multiplexing, amplification and filtering. They have been initially implemented for *in vivo* neural probe recordings [15]. Later they have been used for in vitro devices at a larger scale to overcome the connectivity limitation by making use of on-chip signal multiplexing [12]. A number of voltage recording microelectrode array devices have been developed with significantly higher electrode densities and larger areas. Due to hardware bandwidth limitations, these devices all make some compromise between speed, electrode count, multiplexed sampling, and noise [14].

A high-density 128x128 biosensor array CMOS chip was designed featuring a frame rate of 2K frames per second, and a pitch of 7.8µmx7.8µm over 1mm^2 extent [12]. The device has a very high spatial resolution recording of small areas of tissue, but was reported to have noise levels in the range of 250µVrms, which could make recording smaller extracellular spike signals (20-100µV) a challenge [14]. The simultaneous recording from all electrodes required the front-end amplifiers being placed in each recording site, which, due to area constraints, entailed the high noise levels.

A switch-matrix-based high-density microelectrode array [16] was developed as a hybrid between low electrode count and high resolution arrays. The device has only 126 output channels but these could be digitally selected from among 11,000 electrodes, separated by a pitch of 18µm, using a reconfigurable electrode/readout-channel routing. The device has very low noise levels of 7-9µV, since the front-end circuitry were placed outside the array, where sufficient area for low-noise circuit implementation is available.

Imfeld and coworkers developed an electrode multiplexing , 4096 pixel recording array with a 42 µm pitch and a 2.7mmx2.7mm extent that can record the full frame at a rate of 8KHz. The device has high spatial resolution, a relatively good temporal resolution and a wide extent of ~7mm². The data recording has a hardware implementation inspired by image/video processing concepts. It implements an Active Pixel Sensor (APS) concept CMOS design, acquiring the data as a time sequence of images [17]. Basic amplification was performed underneath each electrode, and a tradeoff between spatial resolution and noise dictated the inter-electrode spacing. The noise level is in the range of ~26µV rms. The complete architecture of the acquisition system is shown in Fig. 2.1. Control and timing of the APS-MEA as well as the bank of the Analog to Digital Converters (ADC) is performed by an FPGA. Filtering the 4096 channels in real time is also carried out on the same FPGA.

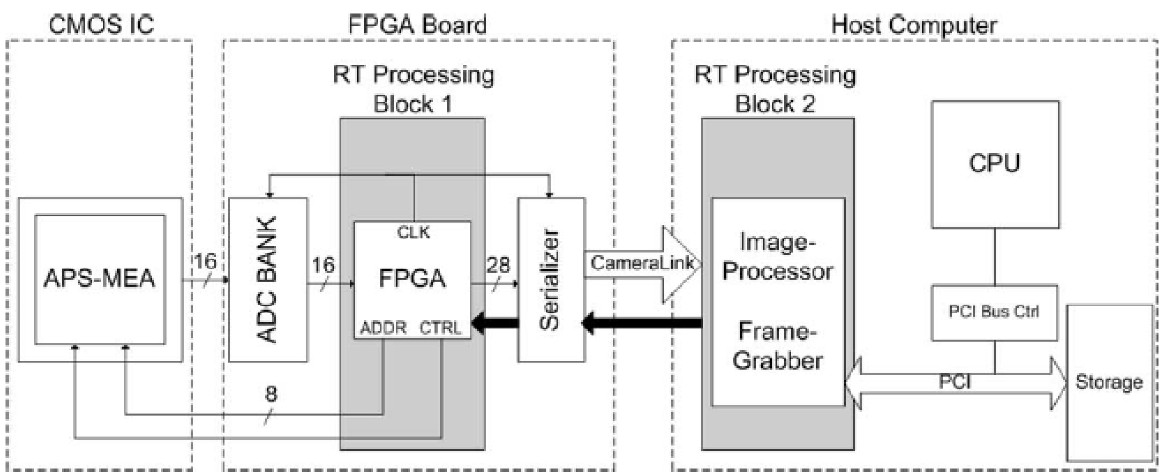


Fig 2.1 Block diagram of the acquisition platform. [17]

Recently a high-electrode count Pico-current Imaging Array (PIA), based on an 81,920 pixel readout integrated circuit camera chip was developed. While originally designed for interfacing to infrared photo-detector arrays, it was adapted for neuron recording by bonding it to microwire glass. The full frame of an area of 9.6mm by 7.7mm can be recorded at 100Hz. [14]

2.1.2 *In Vivo* Micro-Electrode Arrays:

Implantable MEA research considers more requirements and restrictions for acute and chronic implantation. Some research areas focus on the fabrication process, insertion techniques, chronic response of tissue on the implant, wireless implant design and power issues. In this section the main focus will be only on presenting a few of the research efforts on increasing the number of recording sites of neural signals. Some Labs are mainly interested in monitoring more neurons in different cortical areas of the brain [18], while others are interested in changing the microstructure of the neural probes to increase the spatial resolution [19-21].

Researchers at the Duke University lab published a paradigm for recording the activity of single cortical neurons from awake, behaving monkeys [5]. They implanted high-density microwire arrays, developed at Duke University, totaling up to 704 microwires per subject in five cortical areas. Early this year the lab announced that they were able to simultaneously record the firing patterns of close to 2000 neurons. Four multielectrode arrays with 448 electrodes were inserted in rhesus monkey motor and sensory cortices of both hemispheres. There are no publications yet explaining the detailed instrumentation used.

The microwire and similarly structured silicon-based arrays feature one recording site per wire, which limits the capability of the array to capture dense neuronal activity in 3-dimensional setting. Alternatively in 1985 the planar microelectrode array was introduced, using multiple electrodes arranged on implantable silicon shafts [20]. The planar microelectrodes increased the recording spatial precision. It was later modified by proposing double-sided electrodes [22]. These devices contain electrodes on two parallel planes separated by the thickness of the implantable shaft, presenting a building block for a 3-dimensional recording geometry.

Du and coworkers at the California Institute of Technology have fabricated a dual-side electrode array by patterning recording sites at the front and back of an implantable microstructure. They proposed stacking several two-dimensional multishank arrays into three dimensional probe arrays, to access 3-D neuronal structures as shown in Fig. 2.2.

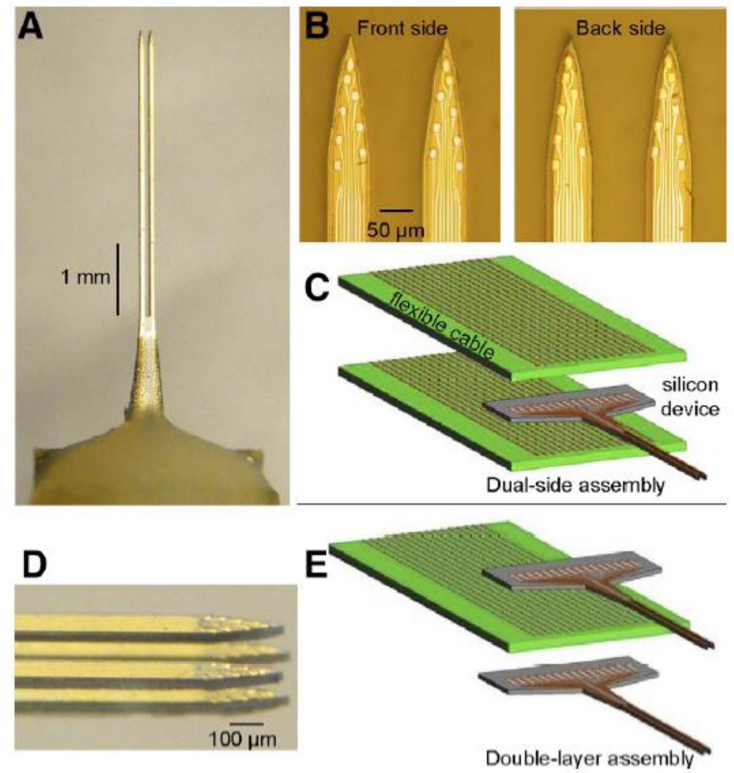


Fig. 2.2. Dual-side and double-layer microelectrode arrays were built on thin silicon shafts. *A*: front view of the device. The shaft dimensions are 4 mm x70 µm x50 µm (*l* x *w* x *t*). *B*: expanded view of the front and back sides of the dual-side array. The recording sites have a geometric area of 100µm². *C*: layers involved in connecting dual-side arrays to flexible printed circuit boards (PCBs, green), one board for each side. Electrical connections were made via low-profile flip-chip bonds. *D*: view of the tip of a 2 x 2 shaft, double-layer array. *E*: a modular assembly scheme used to make the multilayer structure. Note that the PCB contained conducting leads on both sides and thus the same board connected to the upper recording sites on the bottom layer and the lower sites on the top layer. [21]

The nano-probe design presents a potential for hundreds or thousands of recording sites, but it holds a high risk of brain tissue damage. To minimize the disruptive interface between the silicon electrodes and the brain, the nano-probes will pass through more testing and evaluation to determine the optimal shaft size and shaft spacing.

It is evident that there are several efforts aiming to increase the number of recording channels *in vivo* as well as *in vitro* and *in situ*, which leads us to the next section of presenting the available signal processing tools and their capability of handling the resulting high amount of recorded data.

2.2. Neural Signal Processing systems:

Recordings of extracellular neural activity are used in many research studies and clinical applications. Usually, these signals are analyzed as a point process, and spike detection is used to estimate the times at which action potentials from one or more neurons occurred. Recordings from high-density MEAs and low-impedance microelectrodes often have a low signal-to-noise ratio (SNR < 10) and contain action potentials from more than one neuron. Hence, spike detection is often followed by spike sorting, that involves clustering, to assign each event to separate neurons based on AP waveforms.

2.2.1 Spike Detection Algorithms:

The main challenge in detecting spikes is the interference due to background noise. Various spike detection algorithms with different levels of complexity and performance have been presented [23]. The absolute threshold method is widely used as it requires the least computations, but it is highly sensitive to background noise. Various techniques have been proposed for autonomously selecting the threshold based on the statistical characteristics of the recorded signal, while others set the threshold based on a visual inspection of the detected spikes. A different type of algorithms is based on template matching. These algorithms scan the recorded signal for instances, where segments of the signal are similar to templates of spike waveforms. In this case a priori knowledge of the spike waveforms is required and the user should supply a threshold for similarity measures. A different approach suggests using a preprocessors, such as the Nonlinear Energy Operator NEO to give emphasis to the spikes relative to the noise before applying the absolute threshold, consequently improving the spike detection performance.

2.2.2 Neural Signal Processing Systems:

Existing commercial recording systems are limited to a few hundred channels and rely on multiple sequential logic processors connected in parallel. While functional, such systems are difficult to manage, and do not scale well to larger channel counts. The paradigm described by researchers at Duke University [5] for acquiring neural signals from monkeys incorporated the multichannel acquisition processor MAP by Plexon. The MAP recorded all the events that crossed the voltage threshold, set by the user, for subsequent offline spike sorting analysis. Each MAP processor can handle up to 128 channels. For their experiments, they used a custom made MAP cluster, formed by three 128-channel MAPs connected in parallel and synchronized by a common 2MHz clock signal. The initial step in all recording sessions required the experimenter to manually set the voltage threshold for each of the MAP channels connected to an implanted microwire [5]. The threshold was set based on visual inspection of the original analog signals displayed in an oscilloscope as well as the digital signal displayed on the screen of the computer controlling the MAP. With the increasing number of recording channels, it becomes impractical to require the user to tune the spike detection algorithm to the signal properties visualized on each channel. Currently, Plexon is offering an upgraded version of the MAP called OmniPlex® D Neural Data Acquisition System. The system can handle up to 256 channels sampled at 40KHz with a sample precision of 16 bits.

With the rising demand to process a large number of similarly structured parallel channels in real time, there has been an emerging interest in hardware implementation over sequential processors. FPGAs offer massive parallel processing performance and reconfigurable flexibility, which makes them an attractive alternative for real-time signal processing.

The data acquisition systems integrated with the high-density MEAs presented in section 2.1. perform signal conditioning in terms of amplification and filtering, and then send the complete signal to a host PC for storage, off-line spike detection and clustering. [17]. As high-density MEA platform produce data streams in the range of hundreds or thousands of Megabits, the amount of data storage required increases drastically with longer recording times. Real-time spike detection and data compression become vital to limit the amount of data storage.

2.3. Spike-based data reduction:

The idea of data reduction has been addressed mainly in wireless implantable devices for Brain-Machine-Interfaces. Several efforts have been proposed to implement on-line hardware spike detection and send only the spike waveforms while disregarding the interspike samples. The spike waveforms are the only information needed for successive spike sorting. With a limited telemetry bandwidth, it was essential to consider spike-based data compression algorithms to reduce the amount of sent data. With power restrictions of implantable devices, there was also a need to avoid high power consumption associated with the continuous transmission of raw data. The proposed schemes aimed at providing an efficient use of the available transmission bandwidth and an increase of the device throughput. Based on the sparse nature of the neural signal with respect to time, and the average neuron firing rates, the amount of sent data can be reduced to approximately ~2.25% of the total amount of raw data [24].

With a focus on telemetry transmission, Bossetti et al [24] raised an important design consideration for spike-based data reduction in real-time. They demonstrated that although the spike-based compression might be very appealing from the point of view of average bandwidth, it is subject to telemetry bottlenecks during periods of multichannel neuron bursting causing queuing-based transmission delays at the output buffer. They drew the attention to the relation between the ratio of the output to average input bandwidth and transmission latency, the number of samples per spike waveform, the mean firing rate MFR, and the needed queue depth of the output buffer memory. Bottlenecks and latencies are mainly a consequence of accumulating the input data samples over short periods of time before their transmission at the output, waiting for the AP waveform to complete at the output queue. The basic common building block of a spike-based data reduction is shown in fig.2.3. The research paper has concentrated mainly on the transmission delay. The hardware implementation delay is the time between the arrival of the spike waveform at the input buffer and its appearance on the output buffer. The method of spike detection employed will dictate the size and temporal pattern of spike data arriving at the output buffer. These patterns could impact the timing significantly.

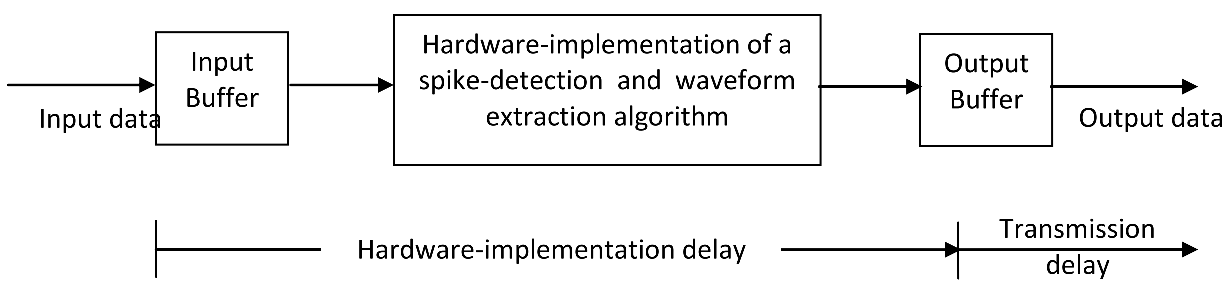


Fig. 2.3 Fundamental Block diagram of a spike detection based compression scheme.

Aside from the delay depending on the scheme control and data handling between the input and output buffers, there are other delays related to the computational overhead and memory read/write times, that depend on the system clock. The performance of the spike detector will also affect the required output bandwidth. A high false detection rate will increase the overall MFR and change the design consideration of the system [24].

2.4. Spike Detector Design Schemes:

The design of the data flow in the spike detection hardware-implementation defines the system latency and memory requirements. With the increasing demand to monitor thousands of recording channels, the efficient use of hardware resources, especially memory blocks on the FPGA becomes vital. Only a few literature have presented detailed patterns and sequences of the data flow on their spike-based compression architectures. This section presents two examples of spike detection architectures with different data flow sequences, and discusses their possible application on high channel-counts. The first example is a spike detection scheme designed for an implantable data acquisition system for BMI application [25]. The second example is is an architecture of a Neural Spike Detection platform NSP [26].

2.4.1. Spike Detection architecture for Implantable Application: [25]

The spike detection based data reduction scheme shown in fig. 2.4. handles the time division multiplexed data recorded from 16 channels. In this design the 64 most recent samples from each channel are stored in the input data storage buffer memory. Once a spike has been detected on a channel, the hardware waits until an additional 34 samples, representing the spike waveform refractory period, from the same channel have been acquired. After the 45 samples of AP waveform is completed in the buffer memory, it waits for its turn in the queue for detected spikes to be written out to the FIFO buffer, where it is held until the embedded PC and wireless card transmit them to the host station.

(a) Queuing based delay: The AP waveform passes through queuing-based transmission twice. Once to be copied from the input buffer to the output FIFO, and again for transmission to the host. The delay increases in case of neuron bursting across the channels.

(b) Memory consumption: In this design, each channel was assigned 64 sample words on the input buffer, as in case of spike detection, the system waited for the full 45 sample AP waveform to be completed in the input buffer before it was copied to the output FIFO. With the increase of the number of recording channels handled by the system we might consider a different design sequence to lower the memory usage on the hardware. For example, copying the AP waveform in single samples, as they arrive at the input buffer, or small groups of samples to the output buffer, can decrease the memory space assigned for each channel in the input buffer.

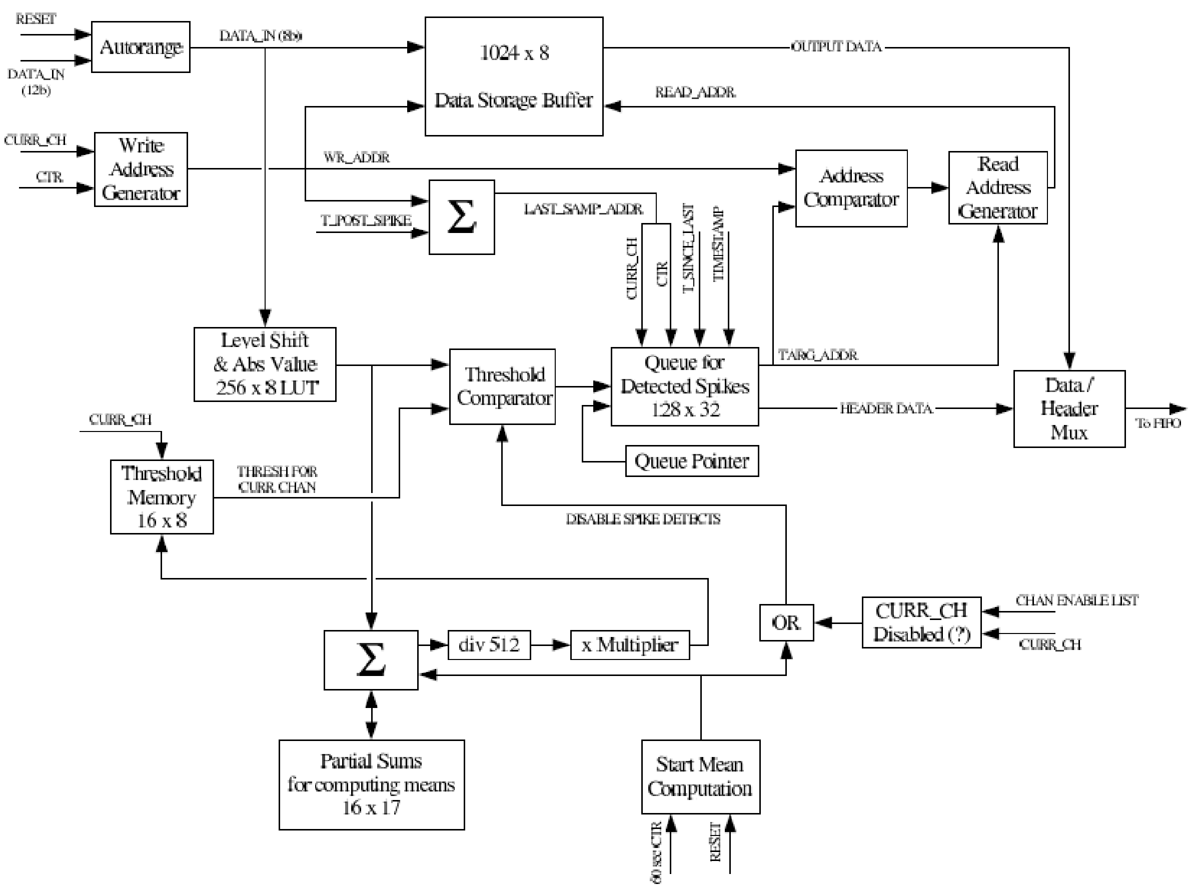


Fig. 2.4: Block diagram describing the spike detector's functionality. [25]

2.4.2. Spike Detection architecture on NSP platform: [26]

The NSP processor incorporates a spike detection processor core (p-core) and a spike sorting p-core controlled by two Microblaze processors as shown in fig 2.5.

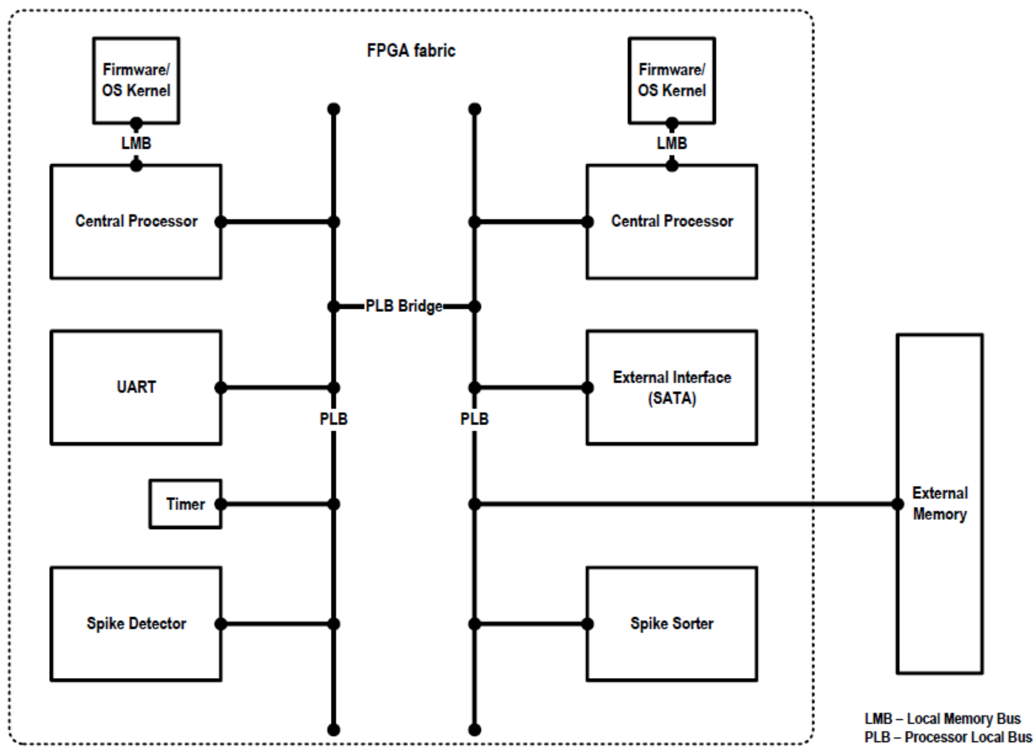


Fig. 2.5. Schematic of the embedded system architecture representation of the NSP platform. The central processors are connected to the firmware layer via the LMBs. Communication between the processors and other subsequent layers are channelized through the PLBs. Two PLBs are interconnected using a PLB-bridge. Each processor behaves as masters in their respective PLBs, while all the other peripherals and p-cores are connected as slaves. [26]

Focusing on the scope of the proposed research, only the spike detection p-core design is investigated. The PLB interface connects the Microblaze processor to the p-core. The central processor manages two tasks: DETI The transfer of input data to the p-core as well as DETO monitoring the spike detection process.

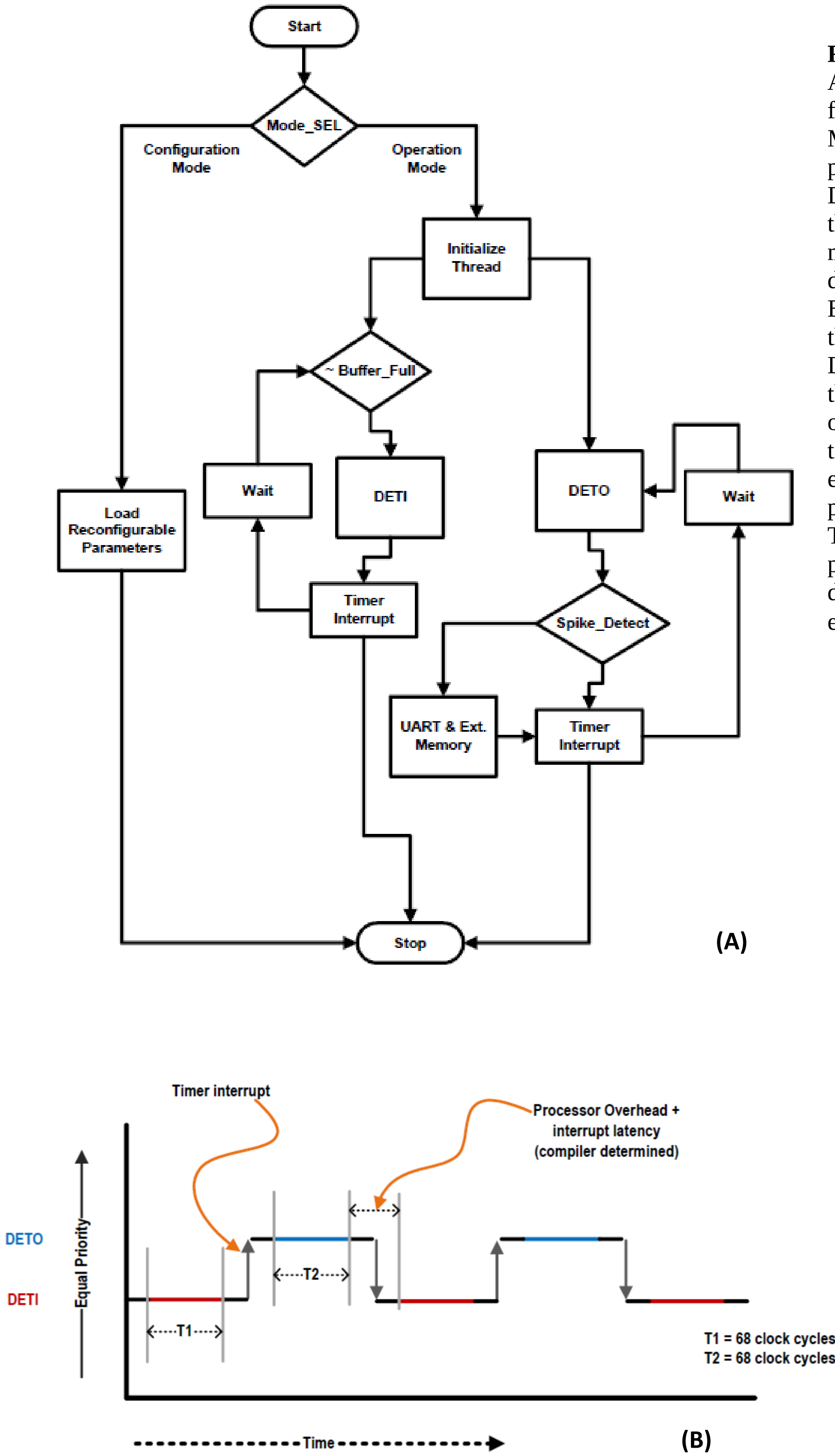


Figure 2.6:
A. Flow chart of the real-time firmware implemented in the MicroBlaze processor-1. This processor manages two tasks, DETI (transfer of data between the acquisition and the detector module) and DETO (spike detection monitoring task)
B. The timing diagram shows the sequencing of the threads DETI and DETO by the Processor-1. Each thread operates for a pre-determined time and the control transfer is exchanged repeatedly till the processor is reset or turned off. The threads are added with processor overhead time determined by the compiler to enable proper scheduling.

(a) MicroBlaze Interrrupt Latency:

The processor operates at 100 MHz. Each instance of the data sampling process occupies 68 clock cycles of the spike detector and another 68 clock cycles were allocated to monitor the spike detection. The scheduler issues interrupts at the end of each task and based on testing results, an average interrupt latency of 5.2 μsec was needed by the real-time firmware. The interrupt latency occupies a significant share of the processor cycles limiting the maximum operational frequency of the p-core. Hence the p-core was set to work at 10MHz, a ten times lower speed than its maximum possible operating frequency defined by the routing critical path. Assuming that the neural signal data is pipelined through the spike detector, the maximum number of channels that can be handled by the platform is limited to ~300 channels. [26]

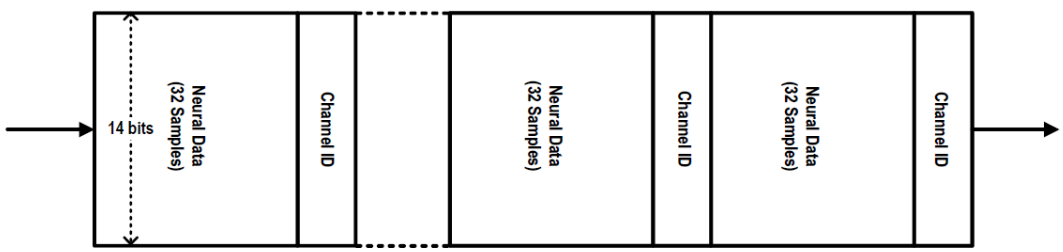
$$\text{Number of Channels} = \frac{\text{Effective Spike Detector Operating Frequency}}{\text{Neural Data Sampling Frequency}} = \frac{10\text{ MHz}}{3.125\text{ KHz}} = 320\text{ channels}$$

The hardware implementation advantages were restrained by the dependency on the MicroBlaze processor to control the operation sequence. If the p-core was to be implemented as standalone module, it can operate at around its maximum operating frequency, defined by the critical routing path. The proposed alternative design solution features a standalone implementation of a spike detector using Finite State Machines (FSMs) to control the interface between the data acquisition and the spike detection core as well as the interface between the spike detector and the output.

The use of a higher processor speed of ~1 GHz will allow the p-cores to run at 100MHz speed. In that case, it is quite possible that the interrupt latency can be reduced by an order of the magnitude (~0.52 μsec) and that p-core could perform at 100 MHz. [26]

(b) Input Data format:

The data processing architecture was based on receiving the neural data as a stream of frames of 32 successive samples recorded from one channel and preceded by their channel ID. Fig.2.7 Simultaneous MEA data acquisition systems incorporate a Time Division Multiplexer (TDM). The rearrangement of the data in the flow scheme required adds control and storage burdens as well as latency to the interface between the data acquisition and the platform. As the system is required to extract valid spike waveforms, the platform has to deal with action potentials split between two data frames, again imposing avoidable complexity to the design.



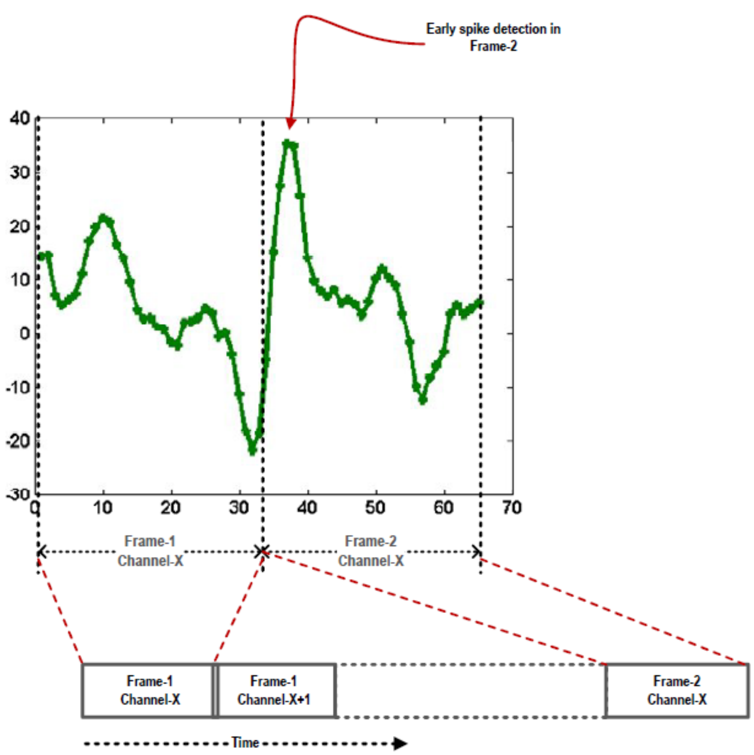


Figure 2.7: Data flow through the input buffer is shown (top diagram). Each set of 32 samples are preceded by a channel ID data. Of the 14 bits, the 13th bit is used to differentiate between channel ID data and neural data (bottom). The 14th bit is used as a validity bit used by the output section of the module

(c) The Threshold comparator and Threshold selection:

Fig 2.8 shows the block diagram of the spike detector core. The threshold comparator compares the neural data from the preprocessor, based on the nonlinear energy operator, to a user-defined threshold to detect spikes. The threshold here was the same fixed threshold for all the channels. The signals recorded by different electrodes may vary markedly in their SNR, and on the same channel SNR may fluctuate over time. With different SNRs the threshold has to be set adequetly for each channel. Dealing with massive number of recording channels threshold selection has to run autonomously without user interference, as manual channel settings become impractical.

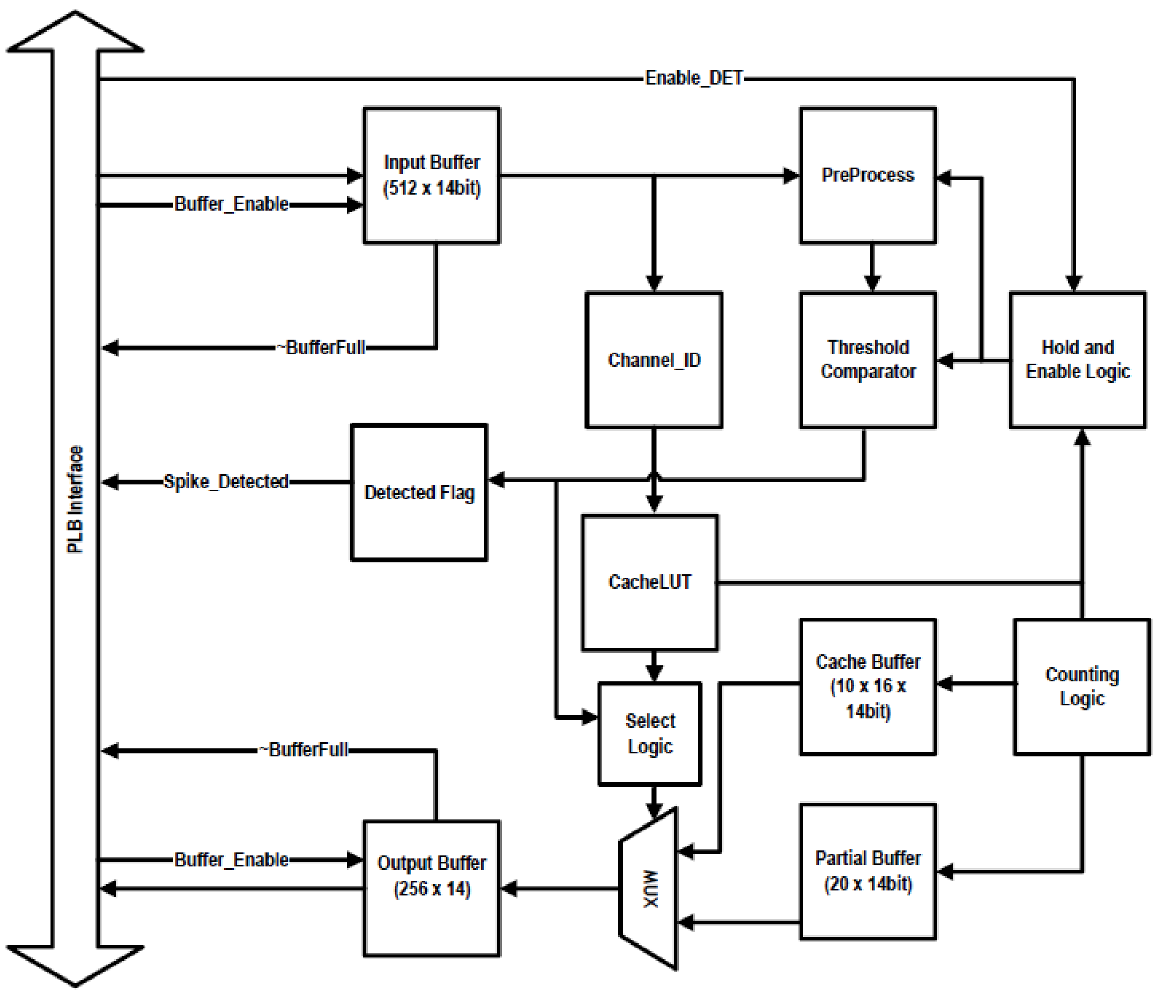


Figure 2.7: Schematic of the spike detector p-core. The PLB interface connects the p-core to the processor. Neural data is input to the input buffer and valid spike waveforms are read out of the output buffer[26].

3. Proposed Design Architecture:

3.1. System Overview:

The Neural Spike Detection platform receives time division multiplexed serial samples from a high number of neural recording channels at the multi gigabit receiver port of the FPGA. The receiver performs deserialization of the data and ensures correct sample-word alignment. The system affiliates each sample to its source channel and performs spike detection. If a spike is detected the spike waveform along with its time stamp and channel ID are passed to an output buffer for further spike sorting or data analysis. Fig. 3.1 presents the integration of the spike detection platform in a typical neural signal processing system.

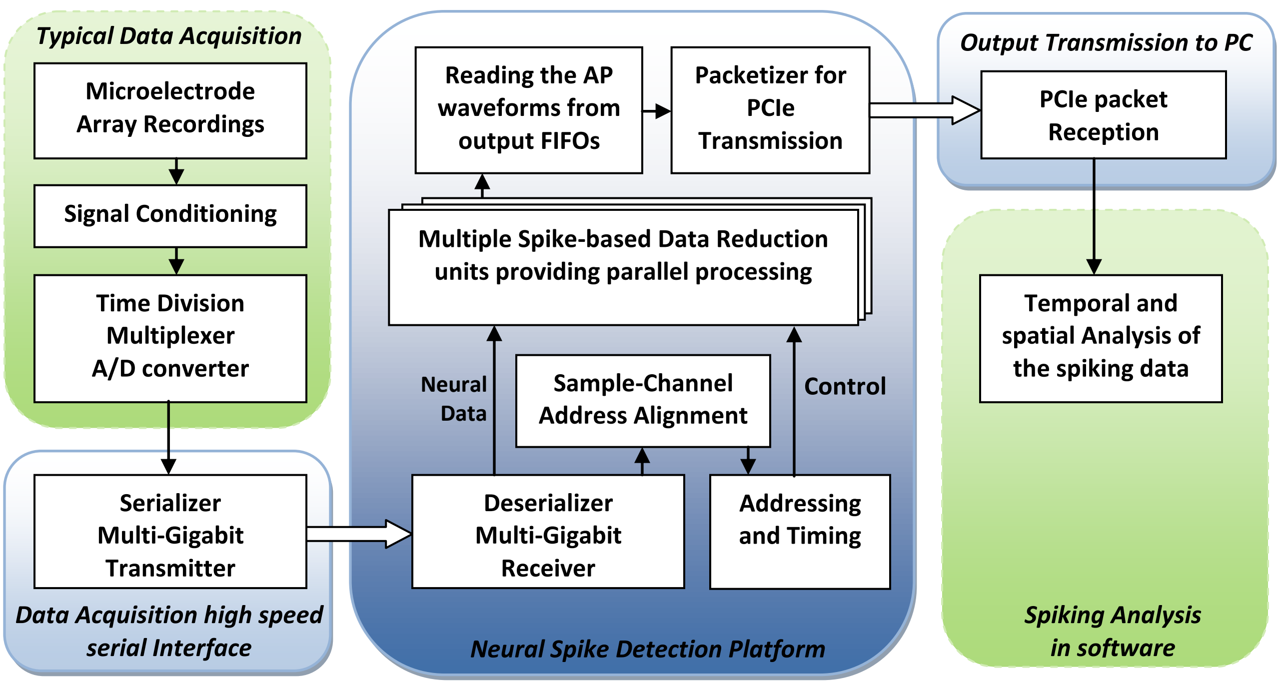


Fig 3.1: A block diagram of the proposed Neural Spike Detection platform and its integration in a Neural Signal Processing system. The center block (dark blue) presents the Neural Spike Detection (NSD) platform performing spike-based data reduction. The blocks (light blue) connected to the NSD platform on the left and right sides present the interface required to embed the platform into a NSP system. The upper left and bottom right (green) building blocks present typical neural data acquisition and spiking analysis on a host PC, respectively. These are not part of the design.

The spike detection platform performs spike-based data reduction. The average reduction ratio:

$$\text{Average Reduction Ratio} = \frac{\text{MFR/electrode} \cdot \text{Number of samples per AP waveform}}{\text{Neural Signal Sampling frequency}}$$

where MFR = Mean Firing Rate. For a MFR of 18 spikes/s/electrode, 50 samples per AP waveform, and a sampling frequency of 40KHz, [24] The reduction ratio = 0.025.

The typical neural signal processing pathway starts with a data acquisition system that records extracellular potentials from an MEA. The data acquisition provides amplification, filtering, time division multiplexing and A/D conversion of data read from the different electrodes. Then the signal passes through spike detection followed by spike sorting, spike binning and analysis. The work proposed focuses on the spike-based data reduction module and is thus concerned with the interface between the ADC of the data acquisition system and the interface with the spike sorting on FPGA or sending the data to a host PC for further analysis.

As the system is designed to handle thousands of recording channels, it has to offer enough bandwidth to receive the massive amount of neural data from the data acquisition system in real time. For example for a 2560 channels sampled at 31.25 KSps, and a precision of 16-bits per sample, the data rate has to be 1.28 Gbps. Consequently, the platform architecture integrates the application of high-speed serial transceivers to allow for the required input data transmissions.

Although, the amount of data is significantly reduced, the system needs to integrate a high-speed communication link to transfer the AP waveforms to the host PC, accounting for transmission bottlenecks during periods of multi-channel neuron bursting [24]. A PCI express link is integrated to minimize queuing-based transmission latencies and performance degradation when the output data overwhelms the transmission bandwidth of the device.

3.2 Spike-based Data Reduction Unit:

The main building block of the proposed architecture is a spike-based data reduction unit that handles 128 channels. This unit can be replicated to process a higher number of recording sites. A block diagram of the spike detection module is shown in Fig.3.2. The spike detection unit receives time division multiplexed 16-bit sample data from 128 channels; it tests the samples for possible spikes, and then sends the complete Action Potential (AP) waveform of a detected spike preceded by the time stamp and the channel ID to the output buffer memory. This section presents the main building blocks of the unit and indicates how the design parameters were selected based on the spike detection algorithm to be applied on the platform.

3.2.1 Spike Detector:

The Spike detector block holds the hardware implementation of the spike detection algorithm. Various spike detection algorithms with different levels of complexity and performance have been presented [23, 27] and can be applied on the proposed platform with proper modifications of the system design parameters. As an example, the design model applies spike detection based on the absolute threshold after passing the signal through a Nonlinear Energy Operator (NEO) preprocessor eq.3.1 in order to give emphasis to the spikes relative to the noise and consequently, improve the spike detection performance.

$$NEO\ [n] = x^2[n] - x[n-\delta]\ x[n+\delta] \quad where \quad 1 \le \delta \le 4 \tag{3.1}$$

where *x[n]* is the neural data sample at any instance *n* .

The threshold for a given channel is set to a multiple of an estimate of the noise level on that channel. The detailed Threshold selection method and block diagram is presented in 3.2.3.

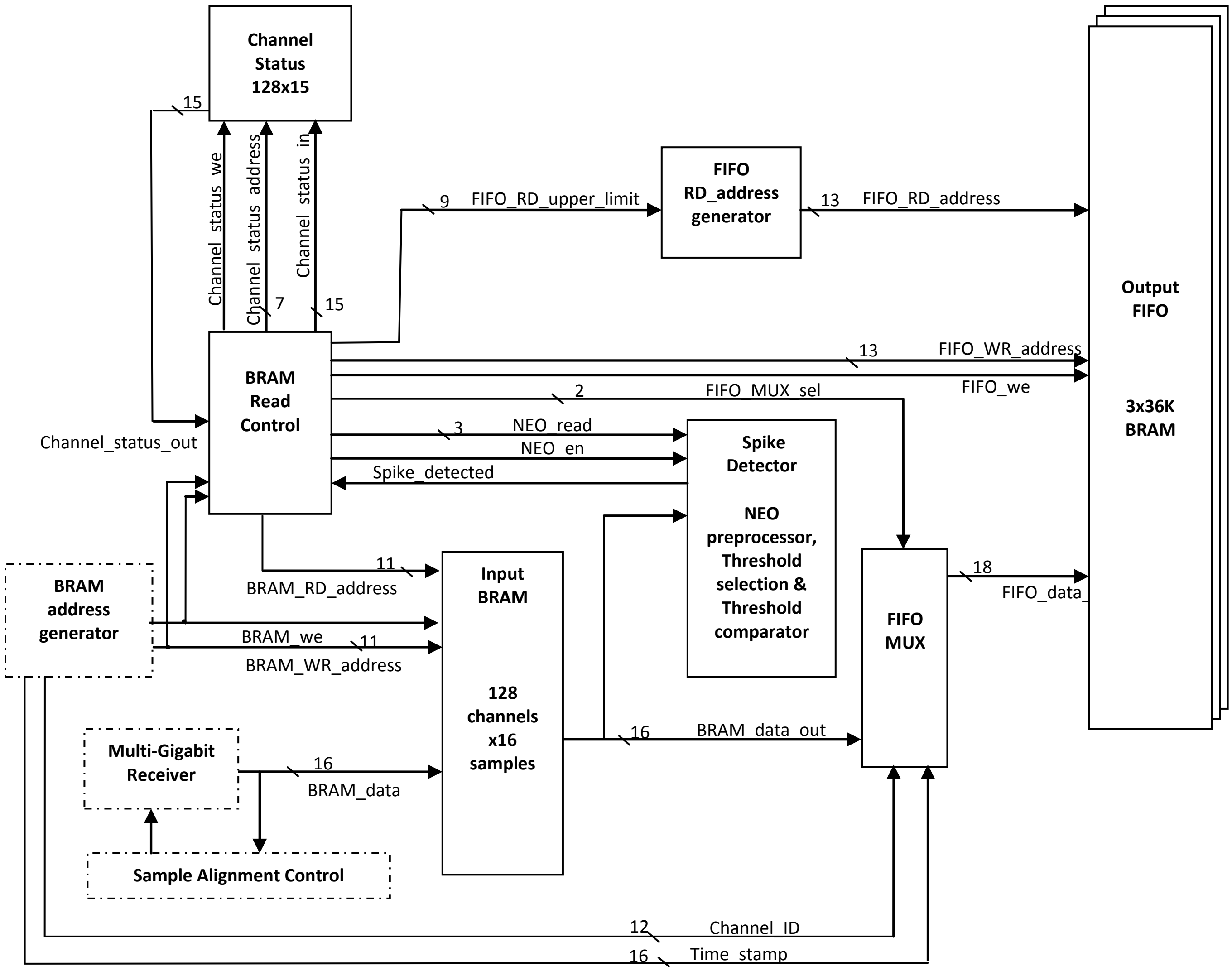


Figure 3.2: A block diagram describing the spike detection process. The spike detection unit is designed to detect neural spikes over 128 channels.

3.2.2 Output Buffer

A neural AP has duration of ~ 1.5ms on the average. Considering sampling rates in the range of 30 KHz and based on the wave-shape, a full AP waveform was assumed to have 10 pre-spike samples, 1 spike sample and 35 samples representing the spike refractory period. This assumption was optimum for organizing the FIFO memory and address assignment. The output FIFO memory 3x36K can hold up to 128 spike waveforms at a time, counting for the worst case scenario if firing neurons are detected on all channels at the same time. When the unit receives a sample from one of the channels it is written in the input memory.

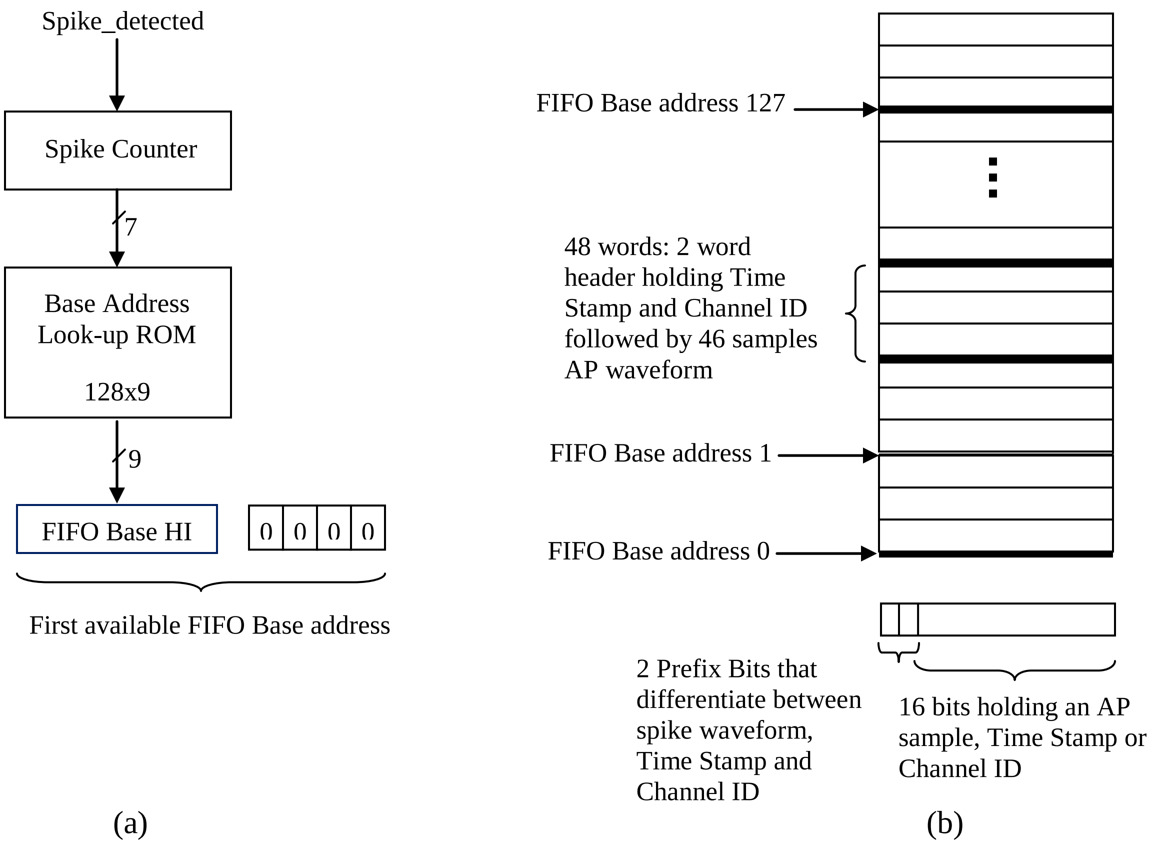
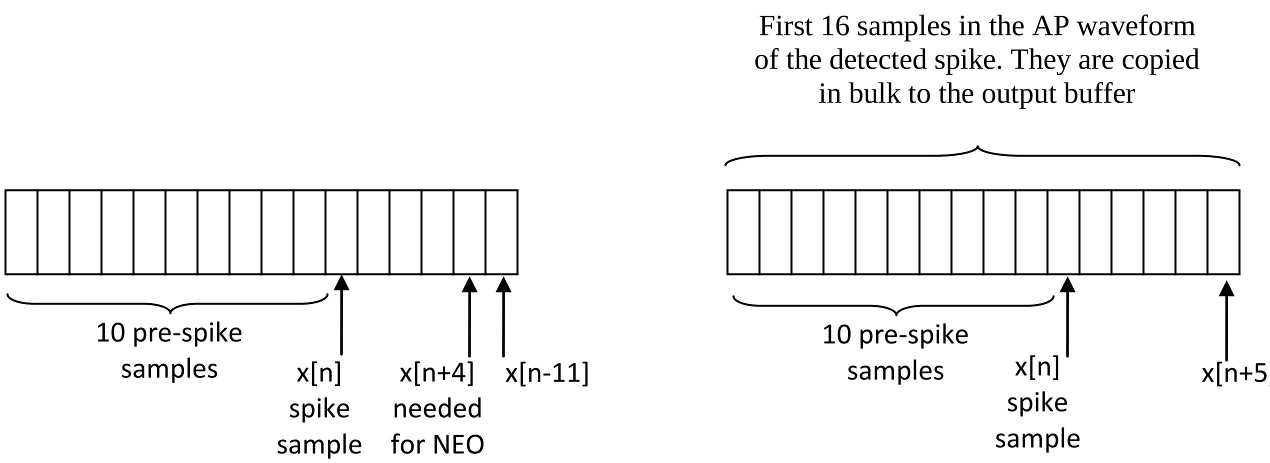


Fig 3.3: (a) Organization of the output FIFO, (b) Spike counter and Base address look-up ROM used to determine the first available memory space in the output buffer to store a detected spike AP.

3.2.3 Input BRAM:

For spike detection consecutive samples are needed to identify a spike. Each channel is assigned a memory space on the input BRAM to hold the most recent 16 samples. The depth of the memory space assigned to each channel was chosen to hold enough sample history to acquire the ten pre-spike samples, the spike sample x[n] and five post-spike samples. Four of the post-spike samples are the "future" samples held to reach x[n+4] needed for the NEO computation, and x[n+5] is added for timing control, as would be explained in the operation management section. The design does not copy the AP waveform as a bulk to the output buffer, instead it copies the first 16 samples, and then sends the refractory period sample by sample as they arrive at the input BRAM. This scheme has minimized the memory space depth needed for each channel, saving on total memory usage.



(a) BRAM memory space assigned to a channel *k* at instance (n+4), at which the spike is detected

(b) BRAM memory space assigned to a channel *k* at instance (n+5), at which the first 16-channels of AP are copied to output buffer

Fig 3.4. : An example of the arrangement of samples on the input BRAM space assigned to a channel k, when a spike is detected and when the initial part of corresponding AP waveform is copied to the output buffer.

3.2.4 Channel Status:

Switching between multiple time multiplexed channels with different statuses requires holding the status of each channel to determine the operation to be applied on the respective incoming input sample. The channel_status memory holds 128 words describing the status of each channel handled by the spike detection unit.

Each word has fifteen bits. Two bits describe the state of the channel, and 13 bits hold the FIFO address needed to copy the AP samples at the right location and space assigned for it on the output buffer in case a spike was detected.

Channel-status	Channel-status description
00	The channel has no detected spikes
01	The channel has a detected spike, time-stamp and channel ID were saved on output buffer. The first 16 samples need to be copied as a complete portion to the output buffer
10	AP samples 17 to 30 are being read sample by sample upon their arrival at the input BRAM
11	AP samples 31 to 46 are being read sample by sample upon their arrival at the input BRAM

Table 3.1: Channel-status-bits and the corresponding status description

3.2.5. BRAM Read Control:

When the unit receives a sample from one of the channels it is written in the input memory. The BRAM read control checks the status of the channel being updated and plans the reading procedure accordingly. The channel_status word can indicate 3 possible cases:

(1) The channel has currently no detected spikes: (channel-status = 00)

In this case the incoming sample is sent to the NEO module and threshold comparator for testing. If a spike is detected, a memory block space of 48 words is saved in the FIFO to hold the corresponding AP waveform. The spike detector unit has a spike counter that is used along with a look up ROM to determine the first FIFO output memory space available for AP waveform storage as shown in Fig. 3.3. If a spike is detected, the counter is incremented, and the time stamp and channel ID of the detected spike are copied in the lower first available FIFO address indicated by the look up ROM. The channel_status word is updated to save block base address that saves a space on the output buffer to hold the AP waveform.

(2) The channel has a detected spike and saved memory space in the FIFO: (Channel-status = 01)

In this case the reading control copies the first 16 samples of the AP waveform available in the input BRAM to the output FIFO memory. (10 pre-spike samples, 1 spike sample, 4 post-spike samples required for the NEO and the incoming sample) This is the longest cycle of the AP waveform transfer to the output FIFO.

(3) The refractory period of the AP waveform is being completed: (Channel-status = 10 or 11)

The incoming sample is copied directly to the output FIFO. The 35 samples of the refractory period are each copied upon arrival at the input BRAM to the output FIFO. This step is repeated 35 times to complete the refractory period. At each cycle the channel_status is updated with the FIFO address that will hold the next incoming sample in the refractory period. Once a spike waveform is completely copied to the output FIFO, the BRAM reading control updates the upper-limit for the FIFO emptying process. The two states (10 and 11) were split into two states to apply an address counter for the lower 4 bits of the FIFO address only, instead of applying an address counter for the whole 13 address bits. The 9 most significant address bits are updated the when the channel moves from state 10 to state 11.

The AP refractory period arrives in single samples at the output buffer. Once the last sample arrives at the input buffer, it is directly transmitted to the FIFO and the complete waveform becomes available for further processing or transmission to a host PC. The design avoids queuing-based transmission that arise from copying the AP waveforms as a whole to the output buffer. The memory space assigned for each channel on the input buffer memory is also reduced. The spike detection module and output FIFO have access to read data samples from input BRAM.

3.2.6 Operation Management :

To control the sequence and timing of operations, a controller employing a finite state machine is used. Figure 5.4 presents an overview of the BRAM read control state diagram. The channel status word has two bits describing the spike copying stage. They are used to decide whether input stream should be passed through the NEO detection module or copied directly to the output FIFO.

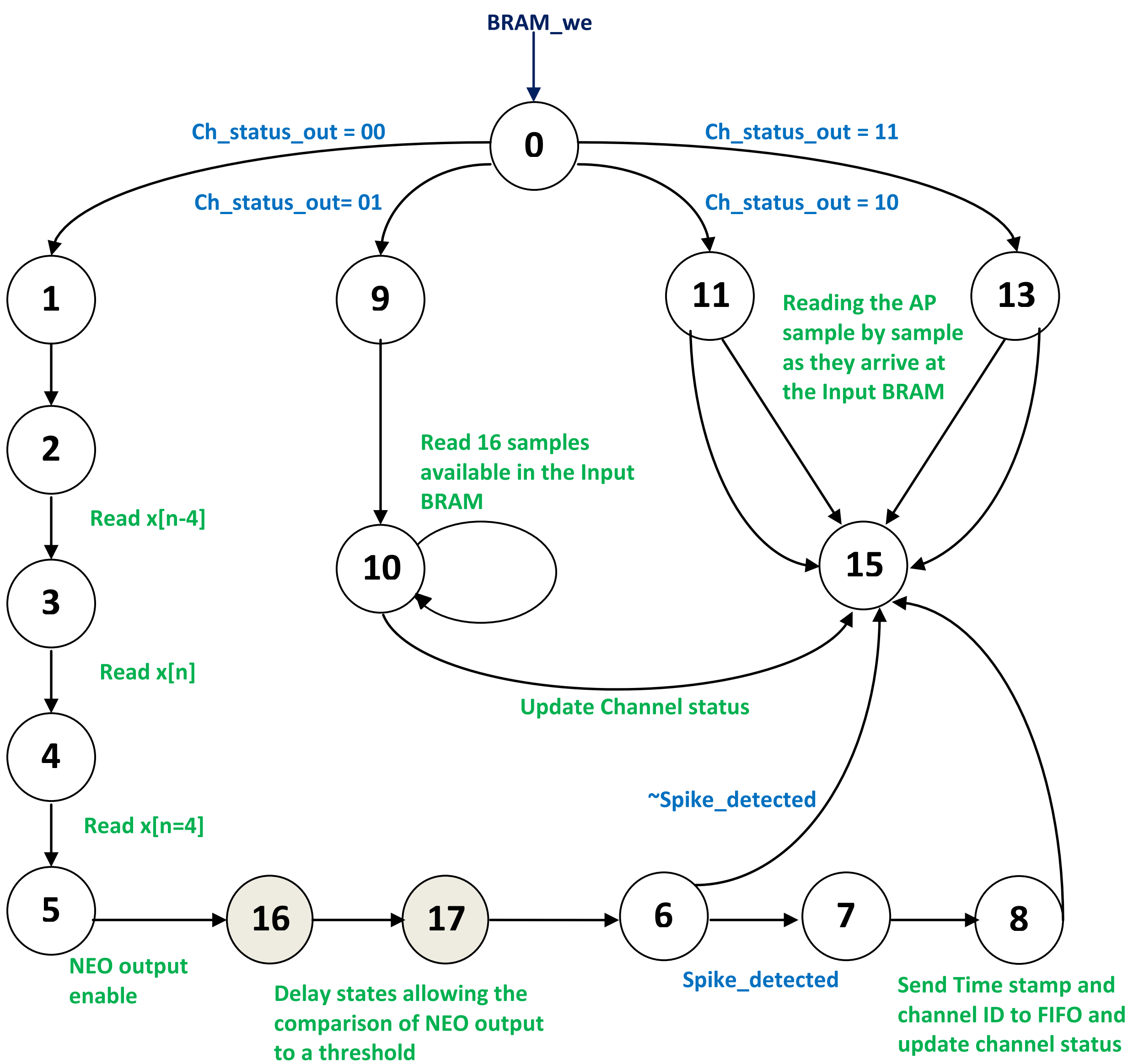


Figure 3.5: Overview of the state diagram describing the controller operation

3.2.7 Autonomous Threshold selection:

With the high channel count automatic threshold selection for each channel is vital. After reset, the system starts computing the threshold for each channel as a multiple of the Mean Deviation MD of a window of its incoming data. The channels are disabled until their thresholds are calculated, and saved on a threshold RAM. Fig.3.6 describes the details of the NEO preprocessing, threshold comparator operation and threshold computation.

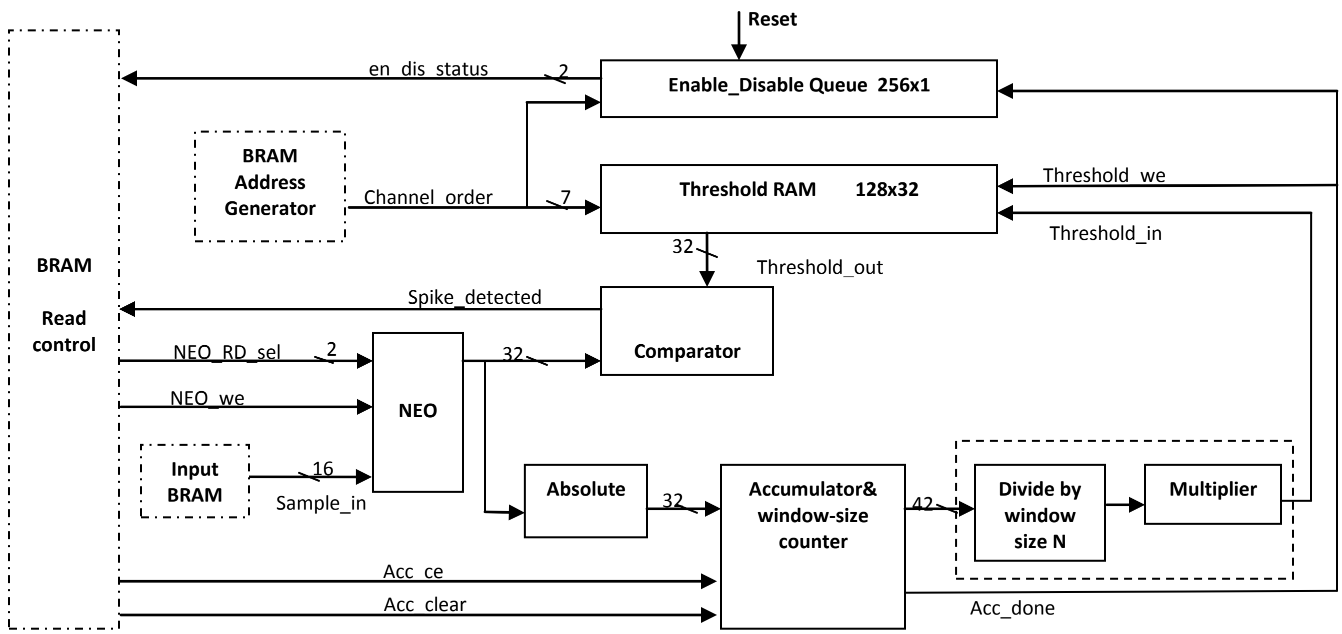


Fig.3.6: Block diagram describing the NEO preprocessing, threshold comparator and threshold

In the normal operation, the samples are passed through the NEO module, the computed output is compared to the threshold of the corresponding channel. In the case of threshold computation, the output of the NEO is passed to a MD computation (3.2),

$$MD = \frac{1}{N}\sum_{N} \left|NEO[n]\right| \tag{3.2}$$

where N is the window size of the data being used to measure the background noise.

N is chosen to be a power of 2, so that the division by N can be performed by right shifting of the dividend. Based on the threshold selection guidance provided in literature [Rizk] the multiplier is chosen to be 16.

3.3. Integration of Several Spike Detection Units:

The total number of channels to be processed is reconfigurable. According to the neural signal processing algorithms used, the longest procedure applied after sample reading was to copy the first 16 samples of an AP in case of spike detection. This procedure required 19 clock cycles. To have an optimum hardware usage, 20 spike-based reduction units were integrated, so that channels on other units can be updated with their respective sample inputs while this longest procedure is being completed, and before that same unit receives a new incoming sample.

Fig.3.7 presents the integration of 20 spike detection units to handle a total of 2560 channels.

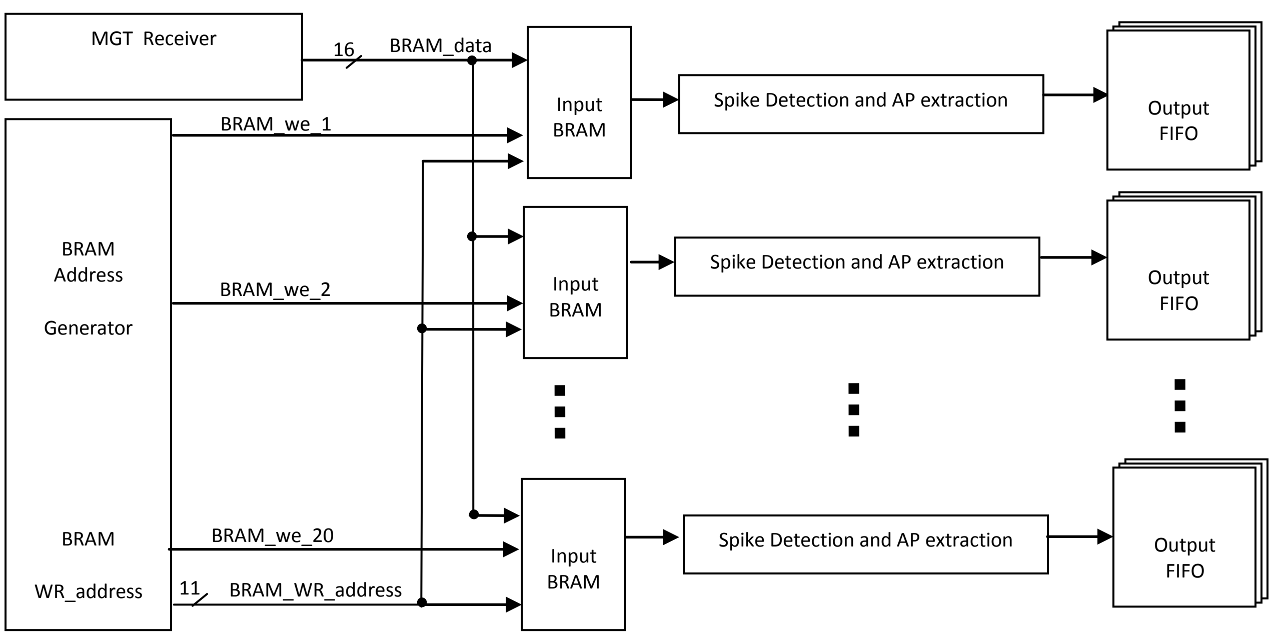


Fig. 3.7: The integration of 20 spike detection units to handle a total of 2560 channels. The BRAM Address Generator generates a 16 bit Time_Stamp signal as well as a 12-bit Channel_ID signal that are connected to a multiplexer at the ingress of each Output FIFO.

3.4. Addressing and Timing:

The BRAM assignment has been chosen so that the BRAM_address can provide direct information on the channel order on the input BRAM and the sample number as shown in Fig.3.8 The write address generator constructs the BRAM write address to rearrange the sample data in preparation for a structured processing. It concatenates the output of three counters to write each sample data in the corresponding channel location.

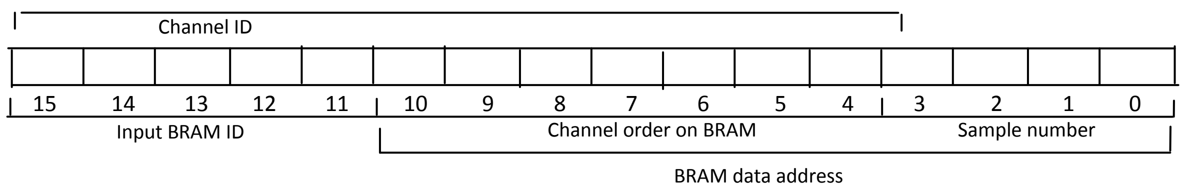


Fig.5.7 BRAM write address structure generated by the Write Address

The BRAM address generator operates at a frequency f, where

$$\begin{aligned} f &= \text{Sampling frequency per channel} \cdot \text{Number of channels} \\ &= 31.25\,\text{KHz/channel} \cdot 2560\,\text{channels} \\ &= 80\,\text{MHz} \end{aligned}$$

The BRAM address concatenates the output of three counters:

(a) a 5-bit counter presenting the Input BRAM ID (20 input BRAMs)

(b) a 7-bit counter presenting the channel order on the BRAM (128 channels per BRAM)

(c) a 4-bit counter presenting the sample number. (16-sample space per channel)

Counter (a) is the fastest changing at every clock cycle. Counter (b) is incremented after (a) reaches a full count cycle of 20 and then is reset. Counter (c) is the slowest counter, that only increments at the full count of counter (b).

3.5 Transmitting the APs from the Output Buffers to host PC :

The design can be extended to integrate spike sorting blocks. In this case the spike sorter will be reading the AP waveforms from the output buffer in their complete format. The proposed design does not include a spike sorter, instead the AP waveforms will be sent to a host PC for system evaluation. The data will be transmitted using the PCI express (Peripheral Component Interconnect express) protocol. The transmission will rely on using the MGTs again for fast performance and low latency. Using Bus Mastering Direct Memory Access (BMD) the output data will be written into the host PC kernel memory for further evaluation or processing.

This section is part of the ongoing work towards completing the proposed project.

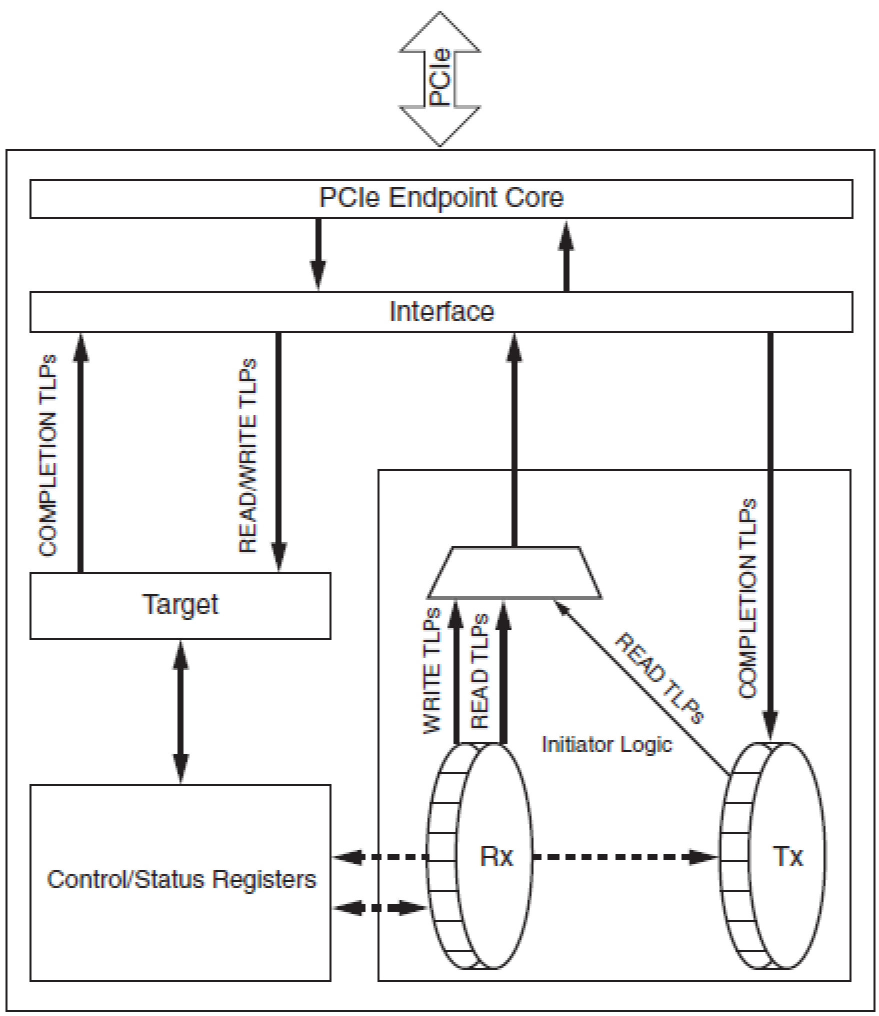


Fig 3.8: Bus master validation design architecture [28]

So far the example design provided by Xilinx was applied successfully in hardware. The design reads one double word repeatedly and sends it to fill the required memory space. The design is being modified in order to send the data on the output buffer instead.

3.6. Data Acquisition High Speed Serial Interface:

The Multi-Gigabit transceiver offers useful features to support a wide variety of interface applications. It has built in Physical Code Sub-layer (PCS) features, such as 8B/10B encoding, comma alignment and clock correction. The comma detection and alignment circuit was activated to properly align the 16-bit input data during the initialization process.

It is worth mentioning that some of the recently developed ADCs have integrated SerDes high speed serial differential that can interface with the FPGA receivers. They offer sampling speeds that include the operating frequency used in the proposed design, i.e. 80MSps. (AD9644 by Analog Devices ®).

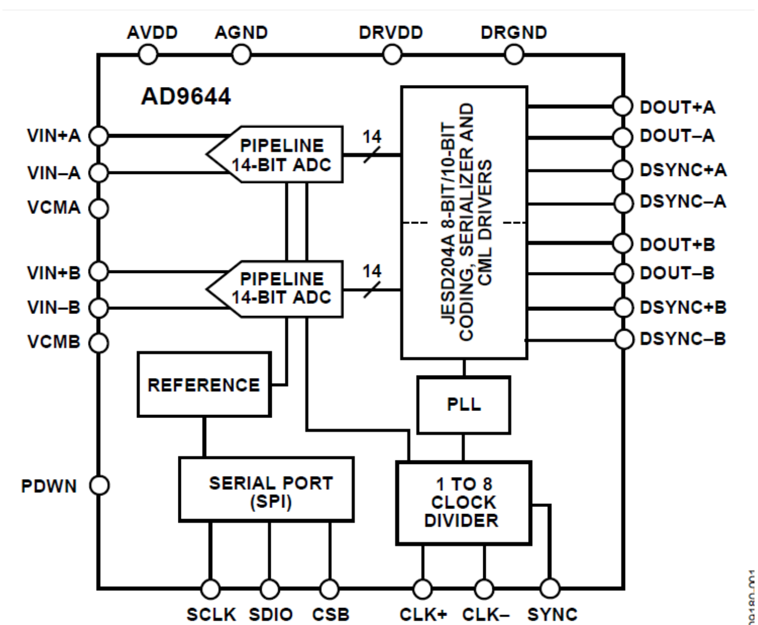


Fig. 3.9 Functional Block diagram of Analog Devices AD9644 [Data Sheet]

3.7. Preliminary Results:

The spike detection processing modules were designed using Verilog HDL code. They were simulated using Xilinx® ISim and tested for functional verification. The Xilinx® Core generator was used to create a Verilog wrapper in order to configure the high speed Rocket IO transceiver. The modules were synthesized and implemented using ISE Design Suite 13.1. For design verification in hardware and as a proof of concept and functionality the proposed design architecture was implemented on a Xilinx® Virtex-5 LX110T FPGA evaluation board. The design model was tested using ISE Chipscope.

3.7.1 Hardware Implementation Setting:

In lieu of interfacing the FPGA to a high speed multichannel analog to digital acquisition system, sample data used as test vectors have been stored on BRAMs on the FPGA as shown in Fig.5. The channel samples are Time Division Multiplexed (TDM). Based on the sampling frequency of the neural signal on the recording channels and the number of channels monitored, the operating frequency of the TDM can be determined. Assuming that each channel was sampled at 31.25 KHz and 2560 channels are monitored, the TDM operates at 80MHz. The 16-bit wide TDM output is serialized by MGT transmitter connected to an SMA connector on the Xilinx platform board, then sent via differential copper cables to the MGT transceiver. The Rocket IO is offers useful features to support a wide variety of interface applications. It has built in Physical Code Sublayer (PCS) features such as 8B/10B encoding, comma alignment and clock correction. The comma detection and alignment circuit was activated to properly align the 16-bit input data during the initialization process.

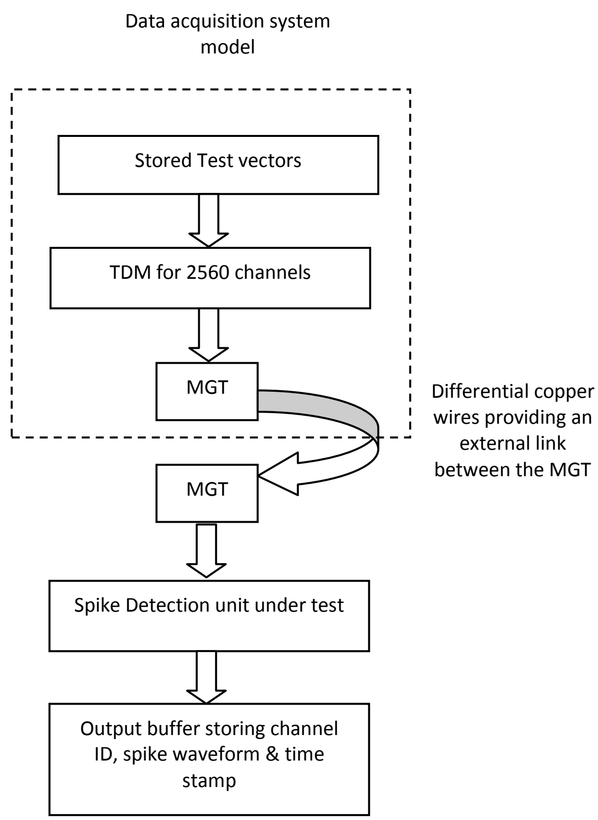


Fig. 3.10: Hardware implementation setting

3.7.2. Testing the spike-based data reduction procedure:

Different testing methods have been conducted to evaluate the performance of the presented design architecture. The aim of this test was to make sure that the spikes have been detected and that their AP waveforms are copied to the output FIFO with the correct alignment required, correct time-stamp and channel ID.

For this test short windows of neural signals containing only one spike were stored on distributed ROMs and read in a cyclic mode. Based on the width of the window, it is possible to control the firing rate of the simulated signal and to determine the exact time stamps. Different Firing Rates (FR) have been tested. For example, for a FR = 125 Hz and sampling frequency of 20 KHz, the total number of samples saved on the ROM was N = 160 sample.

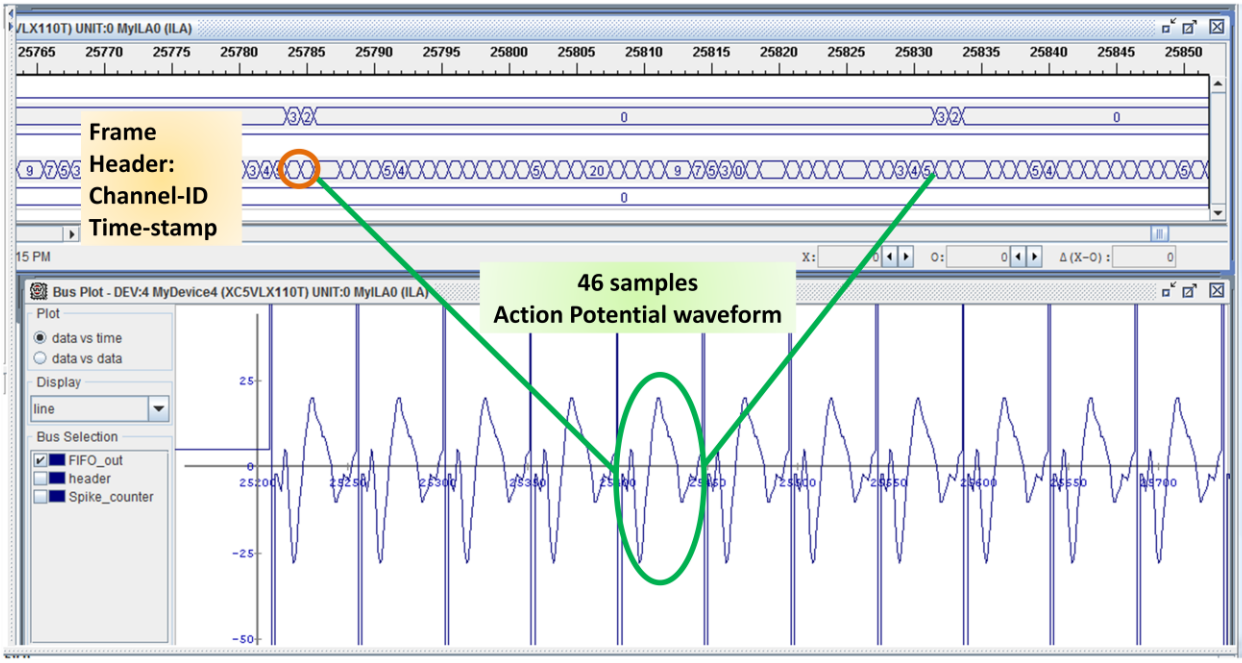


Fig. 3.11: Screenshot from the results obtained on chipscope.

3.7.3 Hardware Usage:

Table I has a design summary describing the hardware usage on the FPGA of the full design integrating 20 spike detection units. The maximum frequency is ~89 MHz. The utilization is based on the Virtex-5 XUP LX110T evaluation board. Virtex-7 FPGAs are expected to have lower utilization percentages, giving more room for design expansion to add spike sorting modules and faster speed.

Slice Logic Utilization	Used	Available	Utilization
Number of Slice Registers	6062	69,120	8%
Number of Slice LUTs	8880	69,120	12%
Number of occupied Slices	2377	12,565	18%
Number of BlockRAM/ FIFO	80	148	54%
Number of BUFG/BUFGCTRLs	1	32	3%
Number of DSP48Es	40	64	62%

References

[1] Deco G, Jirsa VK, Robinson PA, Breakspear M, Friston K. The dynamic brain: From spiking neurons to neural masses and cortical fields. PLoS Comput Biol. 2008 Aug 29;4(8):e1000092.

[2] Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R. The brain activity map project and the challenge of functional connectomics. Neuron. 2012 Jun 21;74(6):970-4.

[3] Lebedev MA, Nicolelis MA. Brain-machine interfaces: Past, present and future. Trends Neurosci. 2006 Sep;29(9):536-46.

[4] Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nat Neurosci. 2011 Feb;14(2):139-42.

[5] Nicolelis MA, Dimitrov D, Carmena JM, Crist R, Lehew G, Kralik JD, et al. Chronic, multisite, multielectrode recordings in macaque monkeys. Proc Natl Acad Sci U S A. 2003 Sep 16;100(19):11041-6

[6] Maccione A, Gandolfo M, Tedesco M, Nieus T, Imfeld K, Martinoia S, et al. Experimental investigation on spontaneously active hippocampal cultures recorded by means of high-density MEAs: Analysis of the spatial resolution effects. Front Neuroeng. 2010 May 10;3:4.

[7] Elaraby, N.; Obeid, I. "A Model Design of a 2560-Channel Spike Detection Platform" IEEE ReConFig 2012

[8] Potter SM. Distributed processing in cultured neuronal networks. Prog Brain Res. 2001;130:49-62.

[9] Buzsaki G. Large-scale recording of neuronal ensembles. Nat Neurosci. 2004 May;7(5):446-51. .

[10] Olsson RH,3rd, Buhl DL, Sirota AM, Buzsaki G, Wise KD. Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays. IEEE Trans Biomed Eng. 2005 Jul;52(7):1303-11.

[11] Gross GW, Rieske E, Kreutzberg GW, Meyer A. A new fixed-array multi-microelectrode system designed for long-term monitoring of extracellular single unit neuronal activity in vitro. Neurosci Lett. 1977 Nov;6(2-3):101-5.

[12] Eversmann B, Jenkner M, Hofmann F, Paulus C, Brederlow R, Holzapfl B, et al. A 128 × 128 CMOS biosensor array for extracellular recording of neural activity. Solid-State Circuits, IEEE Journal of. 2003;38(12):2306-17.

[13] Litke AM, Bezayiff N, Chichilnisky EJ, Cunningham W, Dabrowski W, Grillo AA, et al. What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity. Nuclear Science, IEEE Transactions on. 2004;51(4):1434-40.

[14] Johnson LJ, Cohen E, Ilg D, Klein R, Skeath P, Scribner DA. A novel high electrode count spike recording array using an 81,920 pixel transimpedance amplifier-based imaging chip. J Neurosci Methods. 2012 4/15;205(2):223-32.

[15] Najafi K, Wise KD. An implantable multielectrode array with on-chip signal processing. Solid-State Circuits, IEEE Journal of. 1986;21(6):1035-44.

[16] Frey U, Sedivy J, Heer F, Pedron R, Ballini M, Mueller J, et al. Switch-matrix-based high-density microelectrode array in CMOS technology. Solid-State Circuits, IEEE Journal of. 2010;45(2):467-82.

[17] Imfeld K, Garenne A, Neukom S, Maccione A, Martinoia S, Koudelka-Hep M, et al. High-resolution MEA platform for in-vitro electrogenic cell networks imaging. Engineering in medicine and biology society, 2007. EMBS 2007. 29th annual international conference of the IEEE; ; 2007.

[18] Nicolelis MA, Ghazanfar AA, Faggin BM, Votaw S, Oliveira LM. Reconstructing the engram: Simultaneous, multisite, many single neuron recordings. Neuron. 1997 Apr;18(4):529-37.

[19] Najafi K, Wise KD, Mochizuki T. A high-yield IC-compatible multichannel recording array. Electron Devices, IEEE Transactions on. 1985;32(7):1206-11.

[20] Wise KD, Sodagar AM, Ying Yao, Gulari MN, Perlin GE, Najafi K. Microelectrodes, microelectronics, and implantable neural microsystems. Proceedings of the IEEE. 2008;96(7):1184-202.

[21] Du J, Riedel-Kruse IH, Nawroth JC, Roukes ML, Laurent G, Masmanidis SC. High-resolution three-dimensional extracellular recording of neuronal activity with microfabricated electrode arrays. J Neurophysiol. 2009 Mar;101(3):1671-8.

[22] Perlin GE, Wise KD. The effect of the substrate on the extracellular neural activity recorded micromachined silicon microprobes. Engineering in medicine and biology society, 2004. IEMBS '04. 26th annual international conference of the IEEE; ; 2004.

[23] Obeid I, Wolf PD. Evaluation of spike-detection algorithms fora brain-machine interface application. Biomedical Engineering, IEEE Transactions on. 2004;51(6):905-11.

[24] Bossetti CA, Carmena JM, Nicolelis MAL, Wolf PD. Transmission latencies in a telemetry-linked brain-machine interface. Biomedical Engineering, IEEE Transactions on. 2004;51(6):919-24.

[25] Obeid, I. A wireless multichannel neural recording platform for real time brain machine interface. [Ph.D dissertation]. Durham: Duke Universty; 2004.

[26] Balasubramanian, K. Reconfigurable system-on-chip architecture for neural signal processing [Ph.Ddissertation]. Philadelphia: Temple University; 2011

[27] Zhang F, Aghagolzadeh M, Oweiss K. A fully implantable, programmable and multimodal neuroprocessor for wireless, cortically controlled brain-machine interface applications. J Signal Process Syst. 2012 Dec 1;69(3):351-61.

[28] Jake Wiltgen and John Ayer. Bus Master DMA Performance Demonstration Reference Design for the Xilinx Endpoint PCI Express Solutions.Xilinx, December 2009. XAPP1052 (v2.5).