**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.

**PCIe packet Reception**

**Packetizer for PCIe Transmission**

**Reading the AP waveforms from output FIFOs**

***Typical Data Acquisition***

***Output Transmission to PC***

**Microelectrode Array Recordings**

**Signal Conditioning**

**Multiple Spike-based Data Reduction units providing parallel processing**

**Temporal and spatial Analysis of the spiking data**

**Time Division Multiplexer A/D converter**

**Control**

**Neural Data**

**Sample-Channel Address Alignment control**

**Deserializer Multi-Gigabit Receiver**

**Addressing and Timing**

**Serializer Multi-Gigabit Transmitter**

***Spiking Analysis in software***

***Data Acquisition high speed serial Interface***

***i***

***Neural Spike Detection Platform***

**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:



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.



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.

**BRAM address generator**

 **ROM\_address generator**

**Input BRAM**

**128 channels x16 samples**

16

BRAM\_data\_in

11

BRAM\_WR\_address

BRAM\_we

 **BRAM Read Control**

BRAM\_RD\_address

 **Channel Status 128x15**

Channel\_status\_out

15

 7

15

Channel\_status\_we

Channel\_status\_address

Channel\_status\_in

 **Spike Detector**

**NEO preprocessor, Threshold selection & Threshold comparator**

3

NEO\_read

NEO\_en

Spike\_detected

11

**FIFO MUX**

2

FIFO\_MUX\_sel

16

BRAM\_data\_out

12

Channel\_ID

16

 Time\_stamp

**FIFO RD\_address generator**

9

FIFO\_RD\_upper\_limit

13

FIFO\_RD\_address

13

FIFO\_WR\_address

FIFO\_we

FIFO\_data\_in

18

**Output FIFO**

**3x36K BRAM**

**128 x 48 words**

**48 words = 2 header + 46 sample AP waveform**

**Multi-Gigabit Receiver**

**Sample Alignment Control**

**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.

FIFO Base address 0

FIFO Base address 1

48 words: 2 word header holding Time Stamp and Channel ID followed by 46 samples AP waveform

FIFO Base address 127

2 Prefix Bits that differentiate between spike waveform, Time Stamp and Channel ID

16 bits holding an AP sample, Time Stamp or Channel ID

Spike Counter

Spike\_detected

Base Address Look-up ROM

128x9

7

9

FIFO Base HI

0

0

0

0

First available FIFO Base address

 (a) (b)

**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

x[n] spike sample

x[n+4] needed for NEO

10 pre-spike samples

x[n] spike sample

x[n+5]

10 pre-spike samples

x[n-11]

First 16 samples in the AP waveform of the detected spike. They are copied in bulk to the 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.

**0**

**BRAM\_we**

**1**

**9**

**Ch\_status\_out = 00**

**Ch\_status\_out= 01**

**Ch\_status\_out = 10**

**Ch\_status\_out = 11**

**11**

**13**

**2**

**Read x[n-4]**

**3**

**Read x[n]**

**4**

**Read x[n=4]**

**5**

**NEO output enable**

**16**

**17**

**6**

**15**

**~Spike\_detected**

**Send Time stamp and channel ID to FIFO and update channel status**

**7**

**Spike\_detected**

**8**

**10**

**Delay states allowing the comparison of NEO output to a threshold**

**Read 16 samples available in the Input BRAM**

**Update Channel status**

**Reading the AP sample by sample as they arrive at the Input BRAM**

**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.

**NEO**

NEO\_RD\_sel

NEO\_we

2

Sample\_in

16

**Absolute**

**value**

**Accumulator& window-size counter**

32

Acc\_ce

Acc\_clear

**Divide by window size N**

**Multiplier**

42

**Comparator**

32

**Threshold RAM 128x32**

Spike\_detected

Threshold\_out

Channel\_order

7

32

Acc\_done

Threshold\_we

Threshold\_in

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

**Input BRAM**

**BRAM**

**Read control**

**BRAM Address Generator**

**Enable\_Disable Queue 256x1**

en\_dis\_status

2

**Reset**

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),



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.

MGT Receiver

Streaming neural signal data

16

BRAM Address

Generator

BRAM

WR\_address

BRAM\_ we

Time\_Stamp

Channel\_ID

Spike Detection and AP extraction

Spike Detection and AP extraction

Spike Detection and AP extraction

Input BRAM

(1)

Input BRAM

(2)

Input BRAM

(20)

Output FIFO

(1)

Output FIFO

(2)

Output FIFO

(20)

11

**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.

BRAM\_data

BRAM\_WR\_address

BRAM\_we\_1

BRAM\_we\_2

BRAM\_we\_20

**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

 Generator.

BRAM data address

Channel ID

Sample number

Channel order on BRAM

Input BRAM ID

 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0

The BRAM address generator operates at a frequency f, where



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.

Data acquisition system model

Stored Test vectors

TDM for 2560 channels

Differential copper wires providing an external link between the MGT

MGT

MGT

Spike Detection unit under test

Output buffer storing channel ID, spike waveform & time stamp

**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% |