Digital Signal Processing Techniques and ASIC Development for 3D CdZnTe Gamma-Ray Detectors

by

Damon Anderson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Nuclear Engineering and Radiological Sciences)
in the University of Michigan
2024

Doctoral Committee:

Professor Zhong He, Chair
Professor Michael Flynn
Professor Igor Jovanovic
Research Scientist Yuefeng Zhu

Damon Anderson
damonan@umich.edu
ORCID iD: 0000-0001-5819-8521

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Professor Zhong He, for taking me on as a PhD student and helping me develop certain important attributes of an independent researcher: curiosity, open-mindedness, and efficiency. Opportunities like the ones I've had in the Orion Group are once in a lifetime and only possible as a result of Dr. He's often behind-the-scenes work to continuously bring 3D-CZT to the next level. Notably, the chance to design, fabricate, and test an ASIC with such a unique application was a dream of mine before entering the PhD program, and I feel lucky to have been able to experience it.

This work has benefited greatly from my committee members, Dr. Yuefeng Zhu, Professor Igor Jovanovic, and Professor Michael Flynn. Dr. Yuefeng Zhu sets the standard for technical excellence within the Orion Group and consistently challenges me to contemplate each problem deeper and more meticulously. I am also grateful that Dr. Yuefeng Zhu took the time to provide valuable technical feedback on the majority of this work. Thank you to Professor Jovanovic for providing me with the critical building blocks for radiation detection through NERS 515, and for subsequently taking me on as a GSI. From Professor Flynn, I learned the most important lessons about how to bring an ASIC to life. Practicality, compromise, and relentless attention to details are all fundamental attributes of being an ASIC designer which I can credit to Professor Flynn.

The past and present members of the Orion Group, as well as my collaborator Seungheun Song from Professor Flynn's group have all been essential to helping me grow as a PhD student. Thank you to Dr. Daniel Shy for pushing to include me in fascinating experiments during my first years which helped me begin to understand the 3D-CZT technology. I'm grateful for Dr. Sara Abraham for being willing to provide insights and advice through the various PhD stages. I appreciate the countless hours that Alex Rice endured with me as we developed the PET and SPECT 3D-CZT systems. Thank you to Peter Hotvedt for providing feedback and suggestions on multiple chapters of this work. I would like to credit Seungheun Song for invaluable technical advice on many occasions that molded me as both a PCB and ASIC designer. All of my interactions with Seungheun left me more knowledgeable than before. I would like to extend a special thanks to Dr. Gianluigi de Geronimo for deepening

my knowledge of ASIC design in the field of radiation detection, and for imparting valuable advice on pursuing a career in this path. Finally, I owe a great deal to Jim Berry for his enduring support in so many of the untold aspects of this work. From placing countless purchases, to machining structural components – or amending my errantly designed parts – for PET experiments and for the DAQ-DSP enclosure, Jim is one keystone that keeps the group intact.

I would not have reached this point in my life without the support of my parents. I can only attribute the feeling of confidence and freedom in my life choices to them. I am very privileged to have not been burdened with the tuition costs of my undergraduate education, thanks to my parents. Without such support, it would not have been possible for me to achieve the successes I've had to this point, and I hope to continue to make the most of that privilege.

Finally, I would like to thank Prabs for being an unwavering source of kindness and positivity in my life over the last several years. The best days of my PhD were shared with you, and the most difficult days were brightened because of you. Without you, it would not have been possible. Thank you, I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ADC Analog-to-Digital Converter

APR Automatic Place-and-Route

CPG Coplanar Grid

CZT Cadmium Zinc Telluride

DAQ Data Acquisition

DRC Design Rule Check

DSP Digital Signal Processing

DUT Device-Under-Test

FFT Fast Fourier Transform

FIFO First-In First-Out

FIR Finite Impulse Response

FP Floating Point

FPGA Field Programmable Gate Array

FWHM Full width at half maximum

HPGe High Purity Germanium

IIR Infinite Impulse Response

LDO Linear Drop-Off

LEF Linkable Executable File

LUT Lookup Table

LVS Layout vs. Schematic

MISO Master-In Slave-Out

MLE Maximum Likelihood Estimation

MOSI Master-Out Slave-In

MUX Multiplexer

PCB Printed Circuit Board

PET Positron Emission Tomography

RTL Register Transfer Language

SDF Standard Delay Format

SNR Signal-to-Noise Ratio

SPECT Single Photon Emission Computed Tomography

SPEF Standard Parasitic Exchange Format

SPI Serial Peripheral Interface

SRF System Response Function

UVM Universal Verification Methodology

TOF Time-of-flight

ABSTRACT

Over the last 30 years, CdZnTe (CZT) radiation detectors have been developed into one competitive room-temperature option for gamma-ray spectroscopy and imaging. Electronic readout from triggered anode pixels and a single planar cathode enables the position of each interaction to be located with a resolution of < 500 μm in all three dimensions at 662 keV. Using the 3D position of interaction as a basis for energy calibration, energy resolution of $< 1\%$ for all events, and best energy resolutions of $< 0.35\%$ resolution for single-pixel events at 662 keV have been demonstrated. The unique capabilities of 3D-CZT lend to its use in medical applications such as positron emission tomography (PET). One component of this work is the evaluation of the timing resolution achievable by 3D-CZT, a critical parameter for coincidence systems employed in PET imaging. Prior studies achieved timing resolutions of < 10 ns using 1 cm thick CZT with GHz sampling systems and waveform fitting techniques [18]. In this study, hand-held systems employing 1.5 cm thick CZT, sampling frequencies up to 80 MHz, and linear-fit-based digital signal processing (DSP) techniques yield a best timing resolution of 36.3 ns. Simulations based on the system response waveforms and measured system noise indicate that the measured timing resolution is in agreement with the limits imposed by cathode noise. Towards further optimization of timing resolution in 3D-CZT coincidence systems, investigations on the correlation between cathode noise and detector thickness, area, and applied bias are recommended.

System compactness is at a premium for any fields requiring portability such as military, space, and reactor inspection. For current 3D-CZT electronic readout, a CPU is required to perform complex filtering operations on the anode and cathode waveforms. Discrete analog-to-digital converters (ADCs), and a field programmable gate array (FPGA) are required to read out waveforms from the front-end electronics. To significantly reduce power consumption and area requirements, an application specific integrated circuit (ASIC) is proposed which brings the ADC, data acquisition (DAQ), and DSP all onto one piece of silicon. Four revisions of the DAQ-DSP ASIC have been taped out over the last four years on the TSMC 65nm MS RF GP technology. The 4th iteration of the chip is the first back-end digital ASIC used with 3D-CZT to demonstrate a calibrated energy spectrum, achieving a resolution of

0.6% for single-pixel events at 662 keV. With a die area of 9 mm^2 and a total power consumption of 35 mW, the DAQ-DSPv4 ASIC is a step towards the next generation of low power, hand-held 3D-CZT systems.

CHAPTER 1

Background

1.1 Applications for Gamma Ray Spectroscopy

Gamma rays, much like visible photons, contain valuable information related to their energy – also interpreted as wavelength or frequency. Having energies in the MeV range, though, gamma rays are far more penetrating than visible photons and thus are difficult to detect. Gamma ray sources are ubiquitous, and the application space for gamma ray detection is similarly vast. Naturally, gamma rays are emitted from extreme astrophysical events – gamma-ray bursts, e.g. – as well as from commonly occurring isotopes such as K-40 found in bananas and in humans. With the invention of nuclear technologies, including power reactors and nuclear weapons, the detection of gamma rays has taken on a significant role in safety and preventative measures. In particular, non-proliferation – the prevention of the spread of nuclear materials for ill intent – has become a vital focus in the field of radiation detection. The penetrating nature of gamma rays also provides the basis for useful imaging modalities, such as positron emission tomography (PET) or single photon emission computed tomography (SPECT). In all applications, the spatial and spectroscopic information of the gamma ray source is critical.

Gamma ray spectroscopy is the study of the quantification of energies present in a given gamma ray source. The methods rely on placing high density – or high atomic number – materials in the path of gamma rays such that an interaction between the two is favorable. Once the high energy photon interacts with the material, a certain amount of energy is deposited into the creation of positive and negative charge carriers within the detection volume. By applying an electric field across electrodes placed on the volume, the charge carriers will drift, resulting in a detectable signal. After measuring thousands of interaction signals, the gamma ray signature can be correlated to the known gamma ray emitters to determine both the quantity and identity of the materials present.

1.2 Room Temperature Semiconductors

Characterization of radioactive gamma ray sources begins with the choice of detecting material. It is desirable for the material to be capable of high efficiency and fine energy resolution. Semiconductors comprise a class of materials that achieve these qualities due to their typically high density and atomic number, and low work function. In semiconductor materials, the energy gap between bound (valence) electrons and free (conduction) electrons is on the order of 1 to 3 eV. When an electron is ionized from the valence band, a "hole" is left behind, which serves as the positive charge carrier in the material. The mobility of electrons and holes determines the properties of the semiconductor.

Two of the earliest semiconductors to be used in the field of radiation detection were Silicon and Germanium. While Si provides excellent energy resolution due to its low bandgap of ≈ 1.1 eV [14], it is one semiconductor that has low efficiency due to its low atomic number of 12. Impurities in Si present another challenge since high purity material is required to create a large detecting volume. Ge not only has a higher atomic number (32), but it also can be purified to impurity levels as low as 10^9 $\frac{\text{atoms}}{\text{cm}^3}$ using the traveling molten zone purification technique [14]. As such, High Purity Germanium (HPGe) is considered to be the gold standard for gamma ray spectroscopy. However, the operation of HPGe has additional complications related to leakage current resulting from its low bandgap of ≈ 0.7 eV. To suppress the high leakage, HPGe must be cooled using liquid nitrogen which makes many systems impractical or challenging to operate. As a result, there is clear value in a room temperature semiconductor that yields similar performance to HPGe in both efficiency and resolution.

Cadmium Zinc Telluride (CZT) has proven to be the most competitive room temperature semiconductor option. Unlike HPGe, CZT has extremely limited hole drift, which makes information extraction more challenging [14]. Originally, CZT was considered to be unusable since single-polarity charge sensing methods were not yet invented for semiconductors. In 1994, Luke proposed the idea of using "coplanar grid" (CPG) type electrodes which consists of interdigitated anode electrodes and mimics the Frisch Grid concept used in gas based detectors [17]. Using the CPG configuration, the signal induction by electrons occurs primarily in the vicinity of the anode. As long as the electrons are created outside the anode region, the full electron signal is recovered from each event. This is in contrast to the planar electrode configuration, in which the electron signal is a linear function of the depth of interaction. In 1999, Zhong He introduced the concept of 3D pixelated CZT (3D-CZT) in which a planar

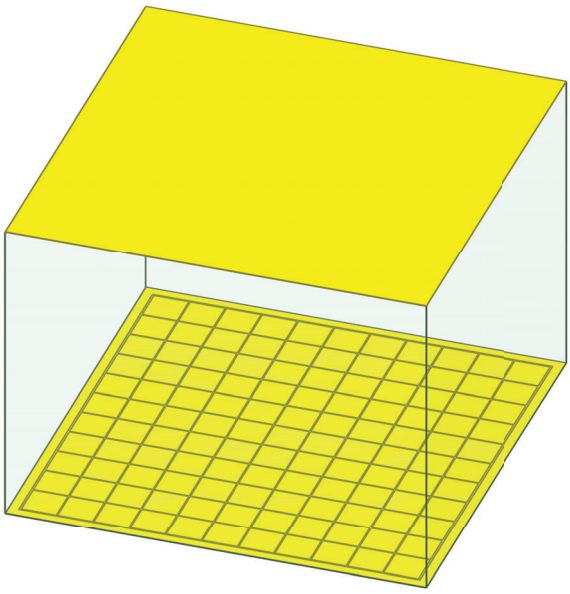


Figure 1.1: 3D model of pixelated CZT generated using Ansys-Maxwell.

cathode and a pixelated anode array are all read out [11] [12]. The electrode configuration, which uses an 11×11 array of pixels, is shown in Figure 1.1. While the electronic complexity significantly increases in this configuration, the high information density allows for the localization of interactions in 3D within the detector. He et. al. showed that by using the ratio of the cathode and anode signals – as is also possible in the CPG configuration – the depth of interaction can be determined. Moreover, the pixelation yields the X-Y position of the interaction. In 2011, Zhu invented the method of subpixel detection in which neighboring anode pixels are used to infer the interaction X-Y location on a scale finer than the physical pixel pitch [34]. By 2019, CZT was shown to achieve, at 662 keV, a best single-pixel resolution of 0.34% and an X-Y subpixel resolution of < 300 μm [32][34]. Using all pixel events, 3D-CZT maintains $< 1\%$ energy resolution and < 500 μm spatial resolution for the 662 keV line.

1.3 The Shockley-Ramo Theorem and Pixelated CZT Signals

The Shockley-Ramo Theorem is a result that simplifies the calculation of induced charge on electrodes in any configuration [26][23]. The theorem states the following:

$$Q = -q\phi_0(x) \tag{1.1}$$

$$i = q\mathbf{v} \cdot \mathbf{E_0}(\mathbf{x}) \tag{1.2}$$

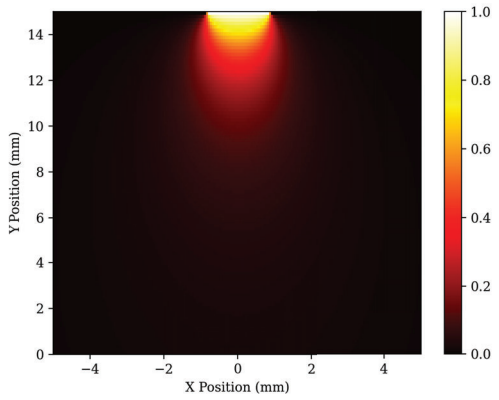
In Equation 1.1, Q is the induced charge on the electrode of interest, and q is the ionized charge in the detecting volume. $\phi_0(x)$ and $\mathbf{E_0}(\mathbf{x})$ are the weighting electric potential and weighting electric field, respectively. To solve for $\phi_0(x)$ and $E_0(x)$, the following routine is used:

1. Set the electrode of interest to 1 V, and set all other electrodes to 0 V.
2. Solve for the resulting electric potential and electric field in the configuration.

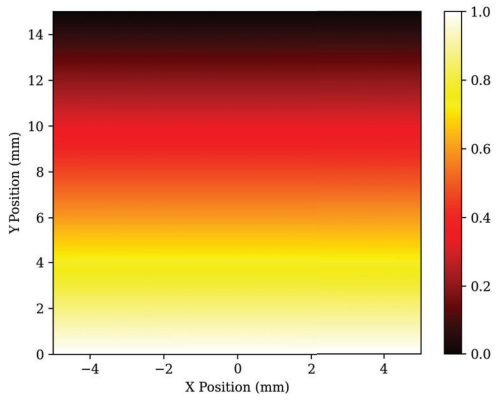
For precise calculations, the weighting potential is determined by a numerical solver. Another value of the Shockley-Ramo theorem is that the resulting potential in a variety of configurations can be understood as a thought experiment using the routine stated above. Moreover, the impact of slight variations – like the addition of conducting metals around or within the detecting volume – can be predicted.

In the pixelated CZT configuration, two weighting potentials are of relevance: the planar potential and the pixel potential. Numerical solutions for both are shown in Figure 1.2. The cathode weighting potential shown in Figures 1.2b and 1.2c is the expected simple result attained by following the Shockley-Ramo Theorem: a linear change in potential such as that in a parallel plate capacitor. The anode weighting potential exhibits the small-pixel effect, as shown in Figure 1.2a. This result can be intuitively understood by considering the potential expected due to a point charge. Of course, the potential will vary as $\frac{1}{r}$, where r is the distance from the charge. The pixelated anode behaves similarly to a point charge relative to the crystal size. In other words, the weighting potential will rise to 1 rapidly near the pixel anode.

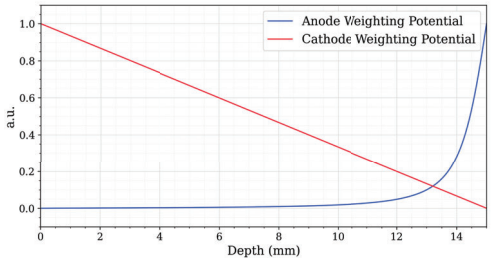
The utility of combining the pixelated anode and the planar cathode is understood by recalling that CZT does not have favorable hole mobility properties. In fact, the holes can



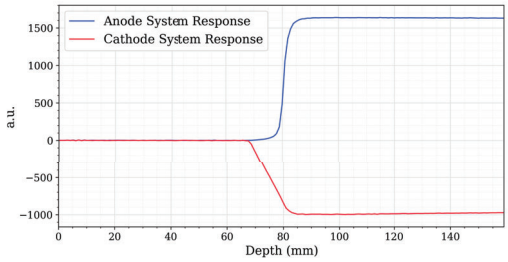
(a) 2D anode weighting potential



(b) 2D cathode weighting potential



(c) Weighting potential depth slice at the crystal center



(d) Anode and cathode system response functions

Figure 1.2: Simulated anode and cathode weighting potentials for pixelated CZT.

approximately be considered to not move at all when electron-hole pairs are created. Then, the signal induced on the cathode by the moving electrons is a linear function of the depth of interaction. On the pixel, however, the signal induction is nearly 0 until the drift charge reaches near to the pixel itself. The induced charge on each electrode, then, is as follows:

$$Q_{CA} \propto q(1-z)$$

$$Q_{AN} \propto q$$

Notably, the ratio of the two charge inductions yields the depth of interaction:

$$\frac{Q_{CA}}{Q_{AN}} = (1-z)$$

Thus far, the cathode and the collecting anode have been considered; however, there are

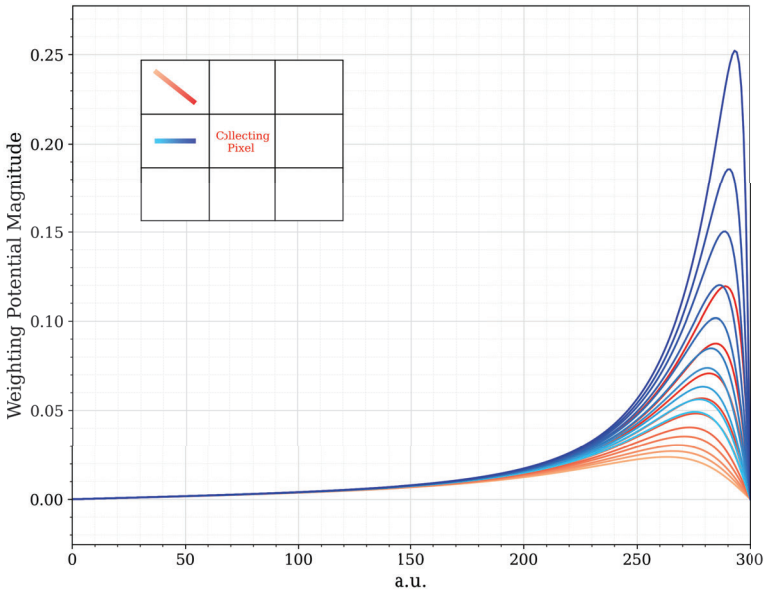


Figure 1.3: Neighboring pixel weighting potential curves for side-neighboring and diagonal-neighboring pixels.

120 other anode pixels yet to be discussed. As shown in Figure 1.2a, the weighting potential extends far into the crystal despite decreasing in magnitude, and the same is true of all other collecting pixels. As the electron charge cloud drifts towards the collecting pixel, some signal must then also be induced on all other non-collecting pixels. Similar to the collecting pixel, the non-collecting pixels will only be influenced when charge is roughly one pixel pitch away. Thus, it is standard to only consider the 8 neighboring electrodes. Figure 1.3 shows an example of the weighting potentials of lateral and diagonal neighboring pixels as a function of the distance from the collecting pixel. The electron charge cloud drifts near to the neighboring pixels at the end of its path causing some signal induction, and then moves out of the neighbor pixel weighting field, causing the induced signal to return to zero. The signal amplitude is a function of the radial distance from the collecting pixel, as expected due to the anode pixel weighting potential characteristic. Although the neighbor signal amplitude is relatively low, the position of the charge cloud within the pixel bounds can be estimated by using the measured amplitude of all 8 of the neighboring signals.

Even the anode pixels with a distance greater than one pixel pitch from the collecting anode will exhibit some response. Figure 1.4 shows the system responses measured for all

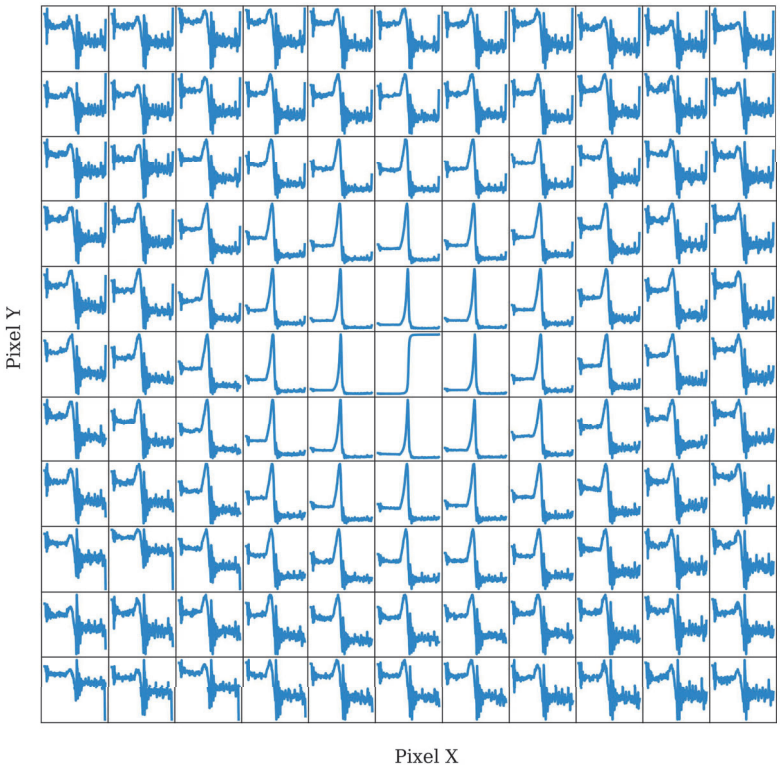


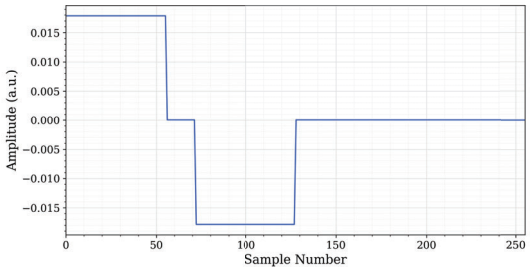
Figure 1.4: Measured neighbor responses for all 121 pixels.

pixels from the Thorium-228 2.6 MeV photopeak at a depth of ≈ 9.4 mm from the anode side of the crystal. The collecting pixel is shown in the center, and the y-scale of all pixel plots is automatically scaled to show that, while minor, all pixels do have some observed signature. Figure 1.4 illustrates that the neighbor signals will have a negative tail, which depends on the depth of interaction. Then, for a perfect energy reconstruction of multi-pixel events, the cross-talk between each unique pair of voxels must be corrected. Methods for calibrating these effects are discussed in [13]. For applications such as high energy gamma-ray spectroscopy, the degradation becomes more significant, and precise methods for removing cross-talk from each pixel pairing have yet to be fully explored.

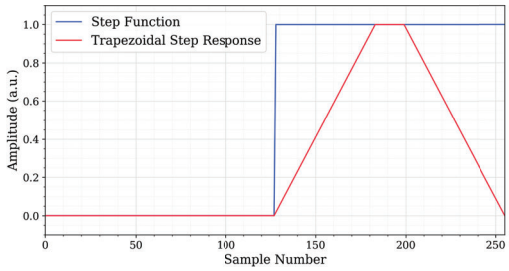
1.4 Digital Filtering Techniques

1.4.1 Amplitude and Timing Filters

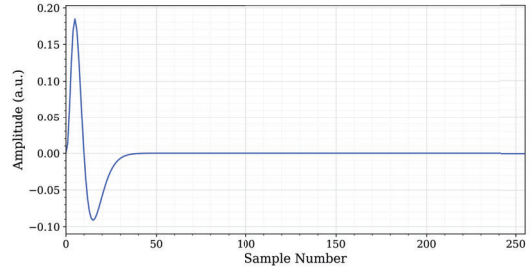
A standard suite of digital filters and parameters has been developed for 3D-CZT, largely motivated by the in-depth study done by Zhu [33]. The goal is to accurately determine the signal parameters of interest that relate to the energy and position of the interaction in the crystal. In particular, the filters are used to extract the timing and amplitude of the cathode and anode, as well as the amplitude of the neighboring signals.



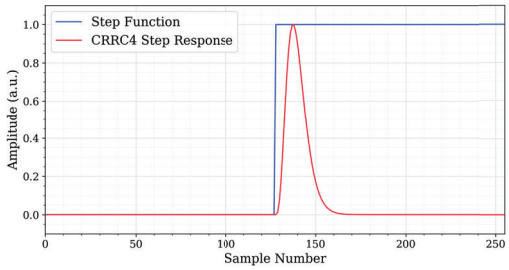
(a) Trapezoidal impulse response



(b) Trapezoidal step response



(c) CRRC^4 impulse response



(d) CRRC^4 step response

Figure 1.5: Overview of standard filters used in 3D-CZT signals processing.

The trapezoidal filter, which emulates a trapezoidal step response as the name suggests, has been reliably used as the amplitude filter of choice for both the anode and cathode waveforms. As shown in Figure 1.5, the trapezoidal filter impulse response consists of two integrating periods separated by a gap of zeros. The two defining parameters are the integrating period lengths, and the gap length. In essence, the trapezoidal filter averages the baseline and waveform tail and subtracts the two in order to determine the amplitude. By applying the standard convolution – or, fast Fourier transform (FFT)-based frequency-

domain filtering method – the average and subtraction is calculated for all points of the waveform. The amplitude is taken to be the maximum value of the filtered waveform. By sweeping both the integration time and the gap time, the optimal filter parameters can be determined based on the system noise.

Simple subtraction is a less rigorous subset of the trapezoidal filter that may be used in situations when signal-to-noise ratio (SNR) is not critical and processing speed is valued, or when transient signals are intended to be ignored. In simple subtraction, the amplitude and baseline are determined by fixed averaging windows, and then subtracted to determine the resulting waveform amplitude. This procedure is nearly identical to trapezoidal filtering, except it is done with fixed baseline and tail windows, rather than convolving the two. While it achieves roughly the same function, the method may be more susceptible to variations in baseline and tail length, and may not always achieve the optimal SNR; however, the process is faster than applying the filtering method.

The CRRC filter – or CRRC^n – is a traditional analog filter used in standard pulse processing chains for a variety of radiation detector readouts. The CRRC^n filter consists of a CR circuit followed by a series of RC circuits. The CR circuit provides a differentiation function, while the RC provides an integration function. When a step function is incident on a CR circuit, the response will be an exponential decay starting at the amplitude of the step function at the moment of the step. Conversely, the RC circuit will respond to a step function by slowly integrating up to the step magnitude over a time period set by the RC time constant. When combined, the CRRC circuit provides a Gaussian-like response, as shown in Figure 1.5. The contrast between the trapezoidal and CRRC filters is clear from Figures 1.5b and 1.5d: the trapezoidal filter is a slow-rising filter with a wide amplitude plateau, whereas the CRRC filter is a fast-rising filter with a limited peaking window. Hence, the CRRC filter is used more optimally for the timing determination.

Neighbor signal processing is fundamentally different than anode and cathode processing because the signal of interest is the transient amplitude as shown in Figure 1.3. In the case of neighboring signals, the timing is not of interest since it is already correlated to the anode timing. The transient amplitude must be calculated by determining the signal maximum as well as the signal minimum. For interactions occurring near the anode side, the signal tail will become negative since the initial weighting potential value is greater than the final value. Thus, two methods are applied to determine the signal amplitude. A first-order CRRC filter is applied, and the maximum value is set as the signal maximum. Then, simple subtraction is used to determine the tail value. The difference between the CRRC maximum and the

tail value is the transient amplitude.

Table 1.1: Standard filter settings

	Amplitude	Rise (ns)	Flat Top (ns)	Timing	Order	Shaping Time (ns)
Anode	Trapezoidal	400	1600	$CRRC^4$	4	100
Cathode	Trapezoidal	800	1400	$CRRC^4$	4	250
Neighbor	$CRRC$	N/A	N/A	N/A	1	100

The standard filtering settings are summarized in Table 1.1. As noted before, the optimal filter settings will depend on the system noise which will be influenced by the readout electronics as well as detector leakage. Table 1.1 is provided as a reference of the typical operation settings.

1.4.2 System Response Functions (SRF)

The standard filtering methods outlined in the previous section are practical and sufficient for generating spectra with excellent resolution; however, in some circumstances, every sample of the waveform must be taken advantage of. The cathode, anode, and neighbor signals all contain information which helps to pinpoint the interaction location and energy. In fact, it is conceivable that every single waveform sample contains some additional bit of information to help improve the position and energy resolution. The SRF is one method for investigating or taking advantage of the minor differences in waveforms across the entire crystal.

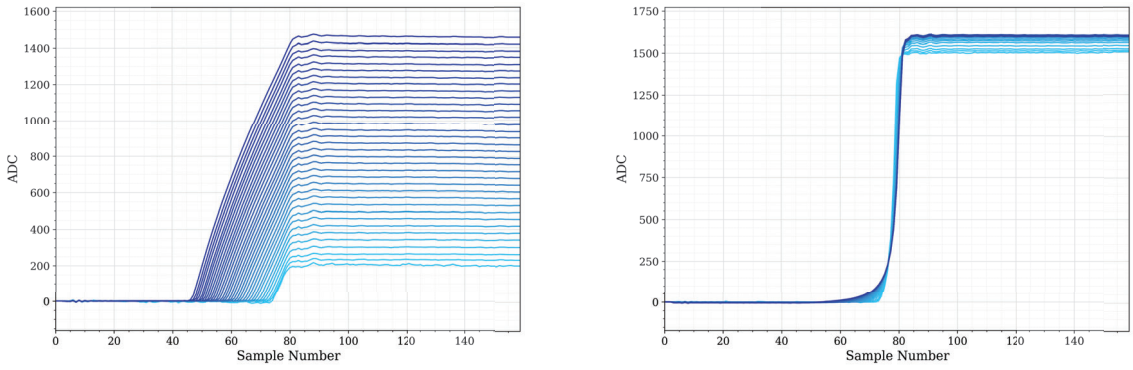


Figure 1.6: Cathode and anode system response functions.

SRFs are generated by categorizing photopeak waveforms into voxelized bins and determining the average waveform in each bin. This yields an estimate of the waveform – or response – expected from each voxel of the crystal. Figure 1.6 shows examples of cathode and anode SRFs, respectively, from varying depths of a single channel. To an extent, SRFs can even be generated on a subpixel basis, implying that the waveform shape can be estimated on a voxelized scale of $\approx (500\ \mu\text{m})^3$. Statistics will limit the viability of this method, though.

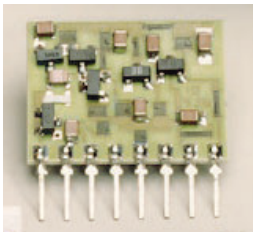
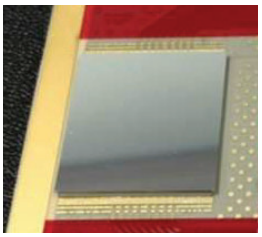
The value of SRFs is two-fold: the results characterize the crystal (uniformity, e.g.), and also provide a method for calibrating subsequent measurements. As an example, the severely warped shape of the cathodes shown in Figure 1.6 indicates that the crystal properties are non-uniform, and the electron cloud drifts slower in the center region of the crystal. Electron cloud distribution estimation is one instance in waveform processing where SRF fitting is useful. In theory, each unique electron shape will induce a slightly different set of waveforms, but the differences are minor and difficult to elucidate using standard filtering techniques. However, using maximum likelihood estimation (MLE) with an SRF-based system model, the electron cloud distribution may be approximately deduced [33][32]. The major drawback associated with SRF-based methods is the added complexity and computational resources required. As a result, SRF fitting is not the standard method for waveform processing.

1.5 Application Specific Integrated Circuits (ASIC) for Detector Readout

Electronics used for detector readout vary considerably with respect to the power consumption, area, and complexity depending on the application. A comparison of pre-amplifiers illustrates the point. The Canberra 2006 pre-amplifier is an example of a rugged and user-friendly module that can be used readily in a variety of NIM-based setups to provide low-noise signal amplification [4]. The form factor of the Canberra 2006 is $7.6 \times 10.2 \times 4.4$ cm^3, and it weighs 0.34 kg. Its power consumption is referenced as 1.296 W. The eV-509x pre-amplifier series is an example of a discrete electronic assembly that offers a lower area and power than full assemblies like the Canberra 2006, but introduces more complexity to the user [3]. The power consumption of the eV-509x pre-amplifier series is reported as 180 mW, but the burden is on the user to design an interface to the system through standard pin headers. The improved compactness may be suitable for systems that contain on the order of 10 channels. For systems requiring on the order of 100 or even several hundred channels, integrated

circuits are required. ASICs make use of MOSFET fabrication technologies which allow trillions of transistors to be placed on a single millimeter squared piece of silicon [19][10]. The Timepix4 ASIC is an example of a chip which implements an array of 512×448 pixels in an area of 24.7×30.0 mm^2 while consuming ≈ 3.5 W [16][2]. This is an equivalent power consumption of ≈ 15.3 $\frac{\mu W}{\text{channel}}$. The inclusion of such an ASIC will increase the design complexity, but is necessary for systems with high channel density. Table 1.2 summarizes the trade-off involving channel complexity, area, and power.

Table 1.2: Power and area trade-off in modern radiation detection electronics [16] [4] [3]

Product	**Canberra 2006**	**eV-509x**	**Timepix4**
			
Power ($\frac{\textbf{mW}}{\textbf{channel}}$)	1296	180	0.153
Form Factor	$7.6 \times 10.2 \times 4.4$ cm^3	2.06×1.52 cm^2	2.47×3.00 cm^2

ASIC technology is a necessity for detectors with high channel density, such as pixelated CZT. Current pixelated CZT systems require at minimum one cathode channel, 121 anode channels, and a guard ring channel, amounting to at least 123 channels. Such a system cannot be implemented using discrete electronic components. The critical readout chain includes the front-end preamplifier, and a means of converting the analog signal to a digital one. This typically consists of either a shaping filter and peak-hold combination, or an analog-to-digital converter (ADC). Along with surrounding control logic and auxiliary capabilities – i.e., test pulse injection and temperature monitoring – this comprises the front-end analog ASIC.

Pixelated CZT occupies a niche between massive-scale projects like ATLAS, and gamma-ray spectrometers with fewer than ten channels. ATLAS, for example, is composed of multiple layers of detectors and several front-end ASICs [25][9][22]. While low-area is important, it is acceptable to occupy large footprints as suggested by the size of the project, and the utmost emphasis is placed on ultra-high data rates [28]. On the other hand, applications involving the CPG CZT configuration, for example, do not require a dense front-end ASIC due to the low channel count, but generally aim for a hand-held form factor. Between the

two, pixelated CZT applications demand both a complex front-end ASIC and ultimately a system that is hand-held. To date, commercial pixelated CZT systems have achieved < 7 W power consumption and a form factor of $10.2 \times 5.7 \times 5.7$ cm^3 [1]. However, to approach the optimization in the most aggressive way, the use of a digital signal processing (DSP) ASIC is proposed in Chapters 4 and 5.

DSP ASICs remain largely untapped in the field of radiation detection, and for clear reasons. In applications that make use of analog peak-hold and time-over-threshold circuitry, there is no need for heavy-duty DSP filtering processes. In experiments like ATLAS, on the other hand, it is desirable and possible to stream out all raw waveform data, even as the scale of throughput reaches tens of Tb/s [28]. With implications for fundamental physics, such data may be reinterpreted and reused in different ways over time, making it most practical to save the raw data. Again, pixelated CZT lies in the middle of the two. In high-rate environments, pixelated CZT systems can stream out gigabytes per minute. Yet, a standardized interpretation of those waveforms is well established. In other words, a processing flow exists that, if successfully implemented on a DSP ASIC, would allow the energy and 3D position of interaction information to be directly streamed out. The power, area, and information density of such a system would be highly valuable in several applications.

1.6 Thesis Overview

The work presented here is divided into two main categories, both related to digital signal processing in 3D-CZT systems: a fundamental study of the limits to timing resolution in CZT, and the design, implementation, and measurement of a digital signal processing ASIC. 3D-CZT is a technology with deep capabilities that continue to be discovered and developed as a result of the information content available in each collected waveform. One less investigated area is the ability of 3D-CZT to operate in coincidence systems, such as those used in PET imaging. The timing resolution, and the factors which limit it, are of interest for such systems. Digital processing methods outside of those detailed in Section 1.4 will be explored in detail in Chapter 2 for this purpose.

Chapters 3 through 6 present the design and test of the DAQ-DSP ASIC, which comprises the bulk of this work. The DAQ-DSP ASIC is a chip proposed to integrate three of the main components of the detector system – the signals processing, data acquisition, and ADC – towards the goals of lower power and area. The implementation of the DAQ-DSP ASIC

spans four revisions over four years, with one final revision still planned. The shortcomings encountered in the first revisions illustrate the challenges associated with ASIC design, and the final successes prove that it is a stepping stone for the next generation of compact 3D-CZT systems.

CHAPTER 2

Timing Resolution Study

2.1 Motivation

Coincidence detection systems are of interest for medical imaging applications, such as PET imaging. In PET measurements, a radioactive positron emitter – or radiotracer – is ingested by a patient. The radiotracer is preferentially taken up by features of interest, such as brain tissue or tumors. The radiotracer emits positrons which then undergo positron annihilation, emitting two co-linear 511 keV photons. By detecting the two photons in coincidence, a line-of-response can be reconstructed. After thousands of interaction lines are reconstructed, an image of the source can be formed.

In coincidence systems, the system timing resolution is a critical performance parameter, as it is one limiting factor for the SNR. In this case, true coincidence events are considered signal, and chance coincidence events are considered noise. For a narrower true coincidence peak – or a finer timing resolution – the integrated chance coincidence is lower, thus improving the system SNR.

System sensitivity, or the achievable count-rate with respect to the source activity, is another crucial parameter of coincidence detection systems. A system that can achieve higher sensitivity will be able to generate a sufficient image in either a shorter time, or with a lesser patient dose, both of which are preferable. The Orion systems used in this study, based on the VAD-UMv2.2 ASIC, suffer from lower maximum count rates compared to analog ASIC based systems since full waveform readouts take significantly longer than just reading out amplitude and timing pick-off information. A trade-off exists here as the analog systems cannot use sophisticated waveform processing techniques to improve timing resolution. In sampling ASIC-based systems, then, it is important to minimize the readout of waveforms which may not be of use. In other words, the system should ideally only read out coincidence events.

State-of-the-art PET systems typically use fast scintillators which are capable of time-of-flight (TOF) methods requiring timing resolution on the order of $\approx$ 500 ps [30]. TOF techniques help to reduce the reconstructed background as the event position is localized to a small portion of the line-of-response depending on the time of arrival of each coincidence photon. The scintillator response time also enables high count-rates. As an example, Siemens Biograph Vision.X PET/CT scanner uses 3.2 mm LSO crystals to achieve 214 ps timing resolution and a sensitivity of 100 cps/kBq [5]. While CZT cannot rival the excellent timing resolution or high count-rates offered by such a system, 3D-CZT has an advantage in terms of energy resolution and position resolution. 3D-CZT achieves $< 1\%$ resolution at 662 keV in contrast to $\approx 10.5\%$ offered by LSO [24]. This provides a unique opportunity for CZT to be used in hybrid SPECT and PET operation when two different tracer energies need to be distinguished. 3D-CZT also has the capability to determine the location of interaction to < 500 μm in all three dimensions. This capability allows single large crystals to be used without any septa while still theoretically enabling mm-scale position resolution. These characteristics outline a niche for 3D-CZT to occupy in pre-clinical, small animal PET where TOF capabilities are not useful. Thus far, 3D-CZT has demonstrated timing resolution of < 10 ns using GHZ sampling and selective waveform fitting techniques [18]. The objective of this study is to investigate the coincidence timing resolution limits of a using a practical research system as one part of evaluating the merits of using CZT in PET systems.

2.2 Coincidence Detection System

2.2.1 UM-VAD Orion systems

The overall detector system consists of 18 pixelated CZT modules split between two detector systems operated in coincidence. Each CZT module is the combination of one $2 \times 2 \times 1.5$ cm^3 CZT crystal and a direct-attached VAD-UMv2.2 ASIC, jointly developed by the University of Michigan and Integrated Detector Electronics AS (IDEAS). The CZT crystal has a planar cathode on one face, and a 121 anode pixels on the opposing face. The pixel pitch is 1.72 mm, with a pixel dimension of 1.66×1.66 mm^2. A guard ring of width 0.54 mm surrounds the anode pixels. The direct-attached front-end ASIC handles the readout of the individual pixels and single cathode channel.

The two detection systems – named Orion-α and Orion-β – are twin systems, each containing a 3 x 3 array of CZT modules. One field programmable gate array (FPGA) and 3

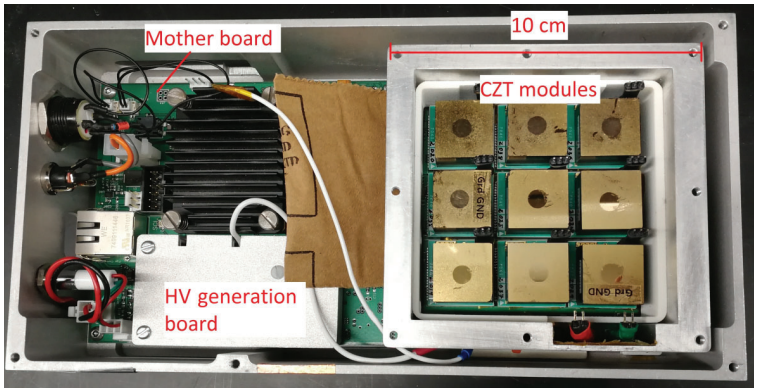


Figure 2.1: Orion-β system interior with 3x3 array of CZT modules [32].

ADCs are used to receive data from each of the 9 modules per system. A 100 MHz system clock is generated on Orion-α and sent to Orion-β via inter-system cabling to provide synchronization between the two systems.

The IDEAS front-end ASIC is capable of sampling at multiple frequencies. Ideally, the best timing resolution will be achieved with the highest sampling frequency, as the signal start can be scrutinized more finely. However, a fixed number of samples are always acquired, meaning that an increased sampling frequency will reduce the number of samples in the signal baseline and tail. The baseline and tail amplitudes are critical parameters for determining the interaction location and energy, which means the increase in sampling frequency comes with a trade-off.

2.2.2 Inter-system Synchronization

The Orion systems are synchronized using connecting cables which carry the clock signal. However, the synchronization method is of particular interest in this study as it is a limiting factor in the timing resolution. The original system used a basic synchronization design, shown in Figure 2.2. Here, each system FPGA generates an 80 MHz clock, which is subsequently used to drive the mixed-mode clock manager (MMCM). The MMCM generates several system clocks which are used for all functions, including waveform sampling and inter-system clock generation. The inter-system clock is shared through an I/O buffer (IOB), and whether the system drives or receives the clock is decided by the system IP address, which may be modified through a physical switch.

This synchronization scheme is sufficient, but suffers from timing jitter between the two master clocks. Any clock-domain crossing from the inter-system synchronized clock to the

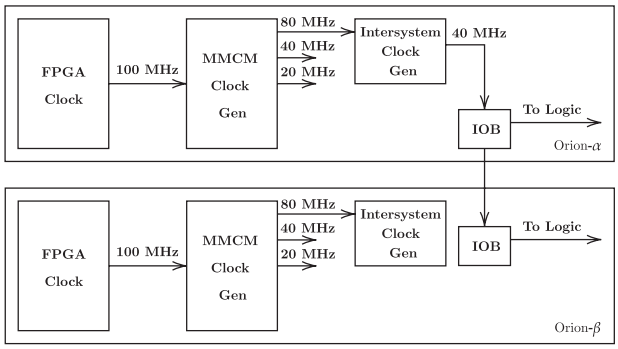


Figure 2.2: Original Orion-α to Orion-β synchronization scheme.

system master clock would introduce significant timing errors. As such, an improved scheme was designed in which the same IOB sharing is used, but prior to the primary clock generation. Ultimately, both systems are driven from the same clock source, eliminating any clock jitter artifacts.

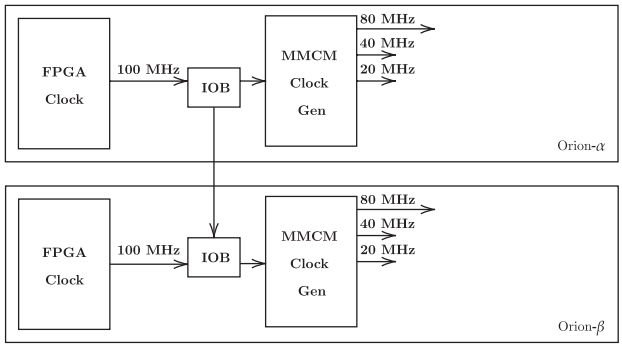


Figure 2.3: Revised Orion-α to Orion-β synchronization scheme.

2.2.3 Inter-system Triggering

The triggering scheme employed for this study is designed such that each system will only read out when both systems trigger. This implementation requires two additional inter-system cables that convey the trigger status of each system. When a trigger arrives for either Orion-α or Orion-β, the respective system will wait for a programmable time frame to see if the other system triggered. If no coincidence trigger is detected, the event is discarded and the system is rearmed for the following trigger. System sensitivity is not quantitatively measured in this study, but this is one measure used to improve the efficiency

of the experimental setup.

2.3 Coincidence Timing Pick-off Methodology

For standard coincidence imaging systems employing fast scintillators, an accurate timing pick-off may be attained using methods like leading-edge discrimination [30]. To achieve the best timing resolution possible, time-amplitude walk should be corrected using constant-fraction pick-off methods. In CZT based systems, the electron drift time may be up to ≈ 1 μs, and the anode fast rising edge is delayed from the time of interaction. As shown in Section 1.3, the cathode signal will start to rise at the time of interaction, while the anode signal rapidly increases when the electrons are less than one pixel-pitch from the collecting anode. Thus, the cathode must be used to determine the actual time of interaction. The cathode signal is linear, though, and the slope significantly depends on the charge deposition, which means traditional methods like constant-fraction pick-off can only yield a timing resolution on the scale of $\approx$ 100 ns. For example, Figure 2.4 shows a coincidence timing spectrum generated by using the standard CRRC^4 and constant fraction pick-off technique. The full width at half maximum (FWHM), determined using a Gaussian fit, is only 81.9 ns. A method, first proposed by Zhu, involving a linear fit of the cathode signal is elaborated on here.

The concept is illustrated in Figure 2.5: two linear segments are fit to the cathode, and the intersection of those line segments is the drift time start. The fitting algorithm must use carefully calculated bounds to achieve the best possible timing resolution. As shown in Figure 2.9b, the cathode signals often exhibit non-linearity related to material non-uniformity. If all cathode rising samples are included in the linear fit, the result will become inaccurate. Another important question is where to start the linear fit from. Ideally, the fit range will start from the drift-time start and range to a fixed number of samples above that. Of course, the drift-time start is not known *a priori*, so it must be estimated by the algorithm. With these considerations in mind, the following algorithm is proposed.

First, the drift-start time is estimated using a fractional pick-off. The cathode amplitude is calculated using simple subtraction as explained in Section 1.4, and the waveform is searched in reverse for the 10% fractional threshold crossing. The waveform must be searched in reverse for this crossing to minimize false pick-offs on noise. In theory, the threshold may be at a fixed amplitude that is safely above the noise; however, a fractional threshold of 10% was shown to sufficiently reduce false threshold crossings. Furthermore, to improve the

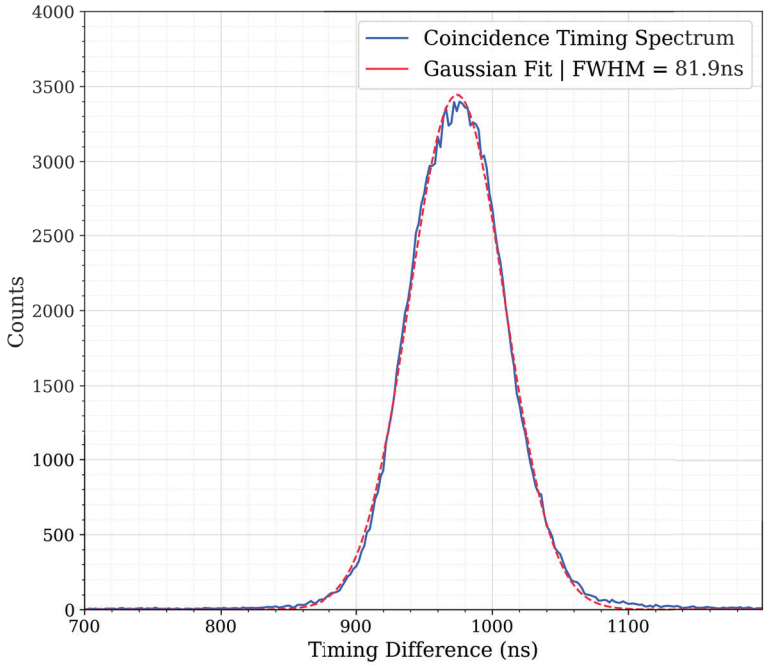
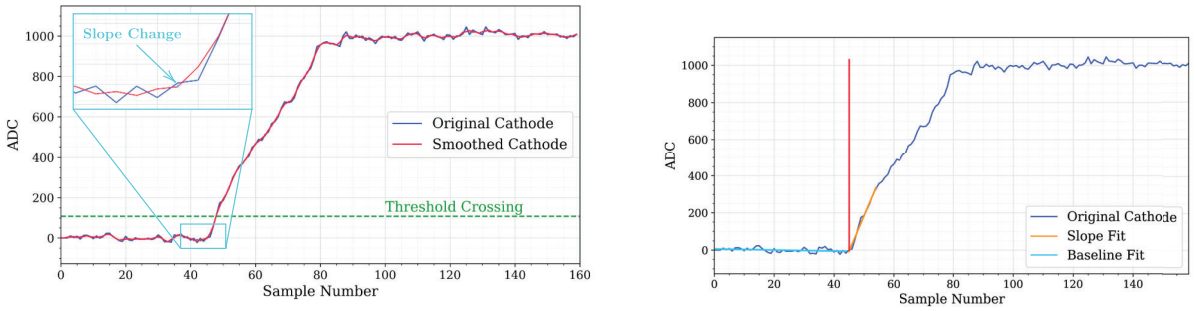


Figure 2.4: Best timing resolution achieved using constant fraction timing pick-off.

accuracy of this search process, a smoothing average filter can be applied to the waveform to reduce the noise. The optimal number of averaging points needs to be considered, as a high number of smoothing average points will start to warp the linearity of the signal around the turning point of interest. After determining the threshold crossing, the estimation of the drift start must be improved by searching sample-to-sample. At this point, the cathode slope is estimated using several points from the threshold crossing onward. Now, still using the average-smoothed waveform, the slope at each sample is checked moving towards the drift-start. Once the slope has deviated significantly from the estimated slope, the search is stopped and the resulting sample is used as the lower bound for the slope fit. A fixed number of samples are included above that starting point. Finally, the baseline, which may include a slight slope itself, is also estimated using a linear fitting procedure. The intersection of these two linear fits is used as the cathode drift start.



(a) Smoothing and slope search procedure

(b) Linear fitting and intercept solving

Figure 2.5: Example of linear fitting procedure used for cathode start pick-off.

2.4 Noise Measurement and Timing Resolution Limits

In order to determine the expected timing resolution achievable by the system, noise measurements are combined with SRFs. Two typical methods can be used to estimate the noise contribution to a given filtering method. First, the method can be applied to baseline noise, and the result can be evaluated at a fixed sample. For the linear fit method, this clearly cannot work because baseline noise contains no linear signal. The second method is to combine a fixed SRF, which represents the expected signal from a given voxel of the crystal, with the baseline noise dataset and apply the pick-off method. This method, illustrated in Figure 2.6 is used here to evaluate the noise contribution to the linear fitting procedure.

Two types of base signals – SRFs and ideal linear waveforms – are combined with noise to understand the expected noise performance. The SRF method accurately captures the expected cathode signals and provides a simple way to probe the effects of non-linearity. However, the only variable parameters using this method are depth of interaction and channel. Varying the depth modifies the signal start sample, as shown in Figure 2.7a, but it also simultaneously varies the slope in the vicinity of the drift start due to crystal non-uniformity. Although the signals shown in Figure 2.7a are quite linear, the slope still slightly varies over different depths. As a result, the effects of the drift start sample and non-uniformity cannot be probed independently using this method. This motivates the ideal signal method, which allows the independent evaluation of the effect of the start sample and signal slope on the timing resolution. An ideal signal, in this case, is an exactly linear signal with a controllable slope and starting sample. Figures 2.7b and 2.7c show examples of the start sample variation and slope variation, respectively. One important note is that the ideal signals are passed

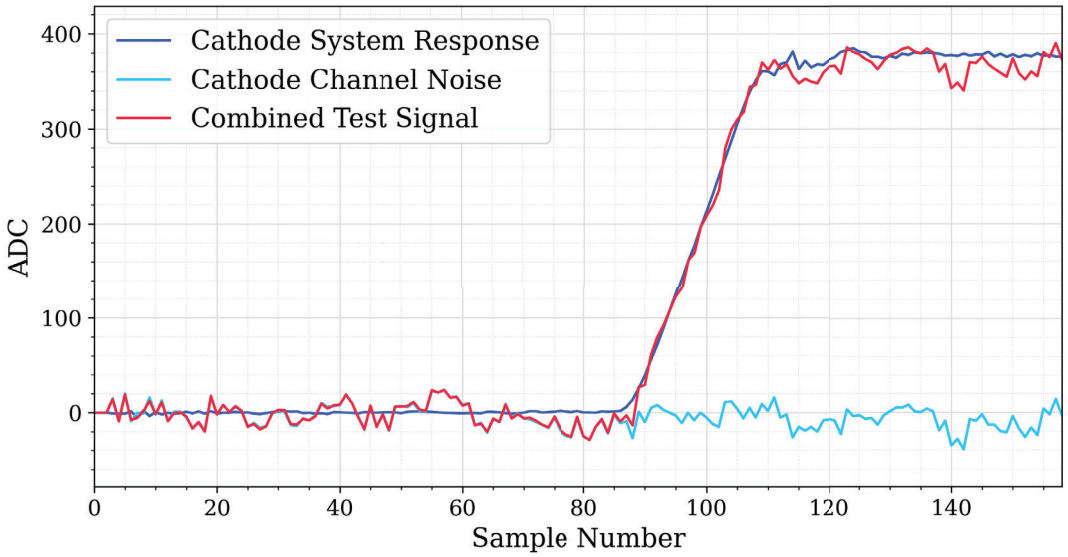
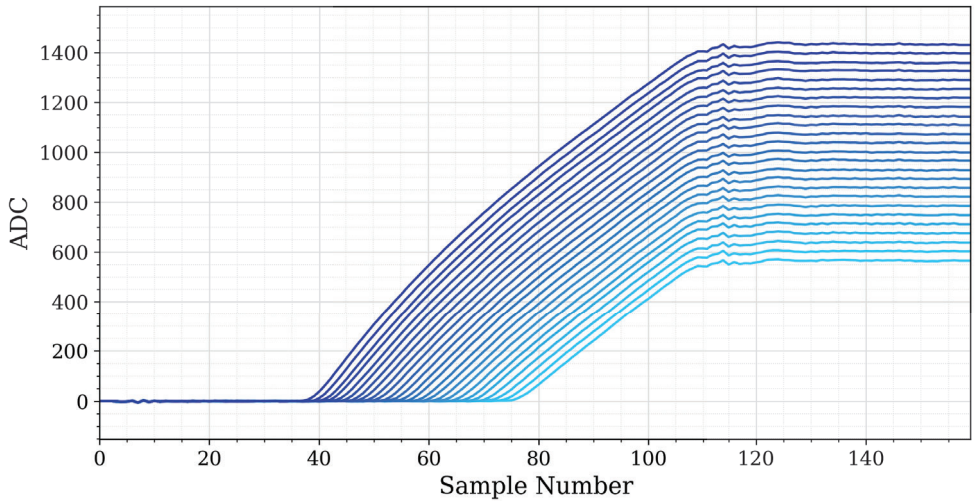


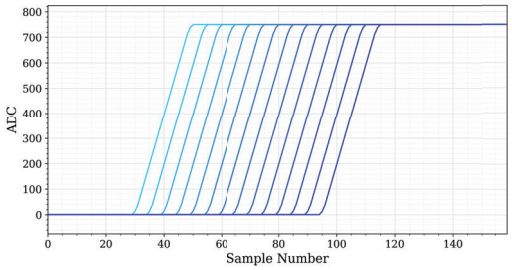
Figure 2.6: Combined SRF and system noise used to evaluate linear fit performance.

through an averaging filter to replicate the smooth turning at the start of the cathode signal, which is typical.

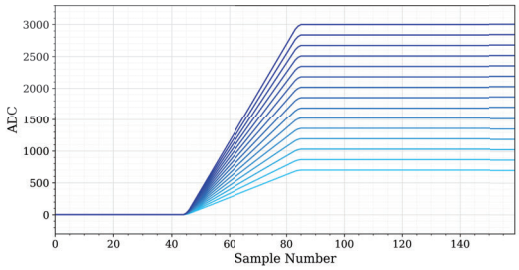
Using the SRF method, several different system settings were evaluated to determine the best timing resolution. The VAD-UM ASIC allows four dynamic ranges: 700-keV, 3-MeV, 7-MeV, and 9-MeV. Since the coincidence experiments involve 511 keV annihilation photons, 700-keV dynamic range may be the best way to maximize the SNR. However, it was found experimentally that the 700-keV dynamic range offered no significant advantage in terms of SNR, and signal saturation at this setting caused further reconstruction issues. The source used experimentally was Na-22 which has a 511 keV annihilation line, and a 1274 keV gamma-ray line. In general, events with energy greater than the dynamic range result in cathode saturation, which may sometimes result in incorrect reconstruction. Although processing techniques can be used to remove errant events, it was deemed not worth the effort since the 700-keV dynamic range appeared to offer no significant SNR advantages. Therefore, the 3-MeV dynamic range was used for all experiments. The more interesting system parameter is the sampling frequency, which includes usable options of 20 MHz, 40 MHz, and 80 MHz. The typical setting is 40 MHz, but it was hypothesized that the 80 MHz option may offer a timing resolution advantage due to the finer time steps. The 20 MHz setting was tested to help verify the performance, although it was expected to perform



(a) Cathode SRFs from the cathode electrode (dark blue) to the anode electrode (light blue)



(b) Ideal signal: start sample variation



(c) Ideal signal: slope variation

Figure 2.7: Basis signals used for estimation of noise contribution to the cathode timing pick-off methodology.

worse.

Before evaluating the combined noise and SRF performance, the baseline noise was measured as a function of the sample index, as shown in Figure 2.8. The reduction in FWHM between samples 10 to 20 has not been explored because those samples will generally be inconsequential to the recorded waveforms. The general trend indicates that the storage time until a sample is read out increases the noise. As such, the 80 MHz sampling frequency is again expected to slightly outperform 40 MHz, and the 20 MHz setting is expected to do the worst. Figure 2.8 also introduces the hypothesis that the timing resolution may degrade for waveforms with later drift start samples. However, it will be shown in Figure 2.11b that the start sample does not actually affect the timing resolution. Aside from drift start sample,

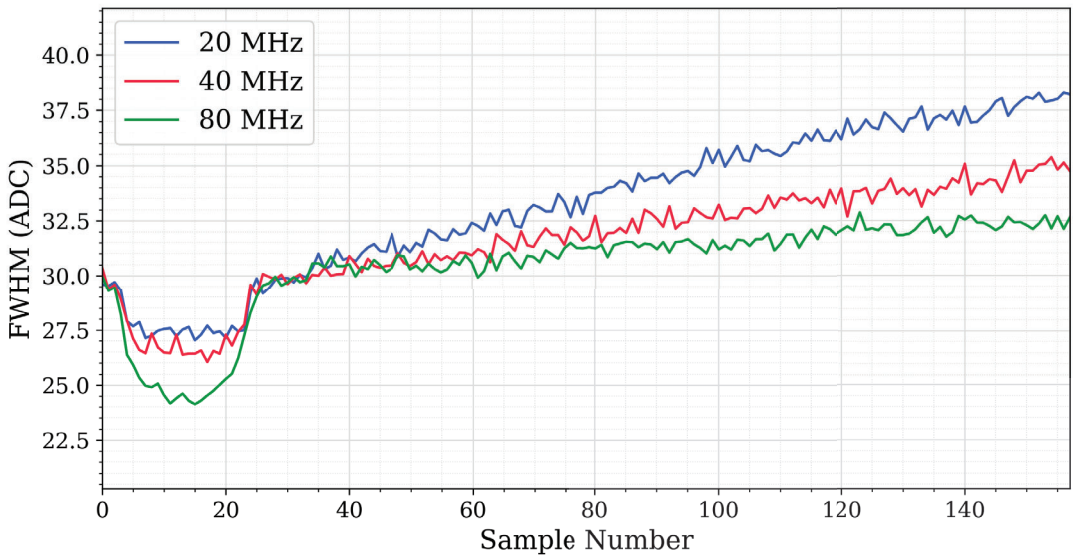
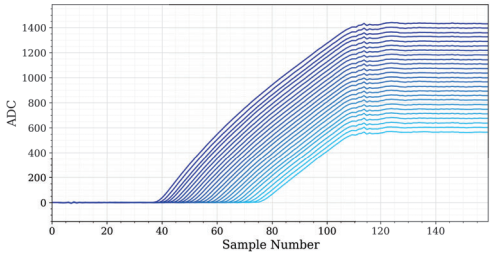
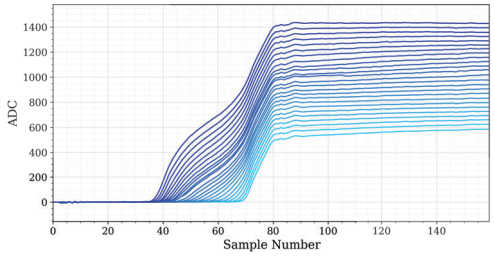


Figure 2.8: Unfiltered sample FWHM measurement using forced readout.

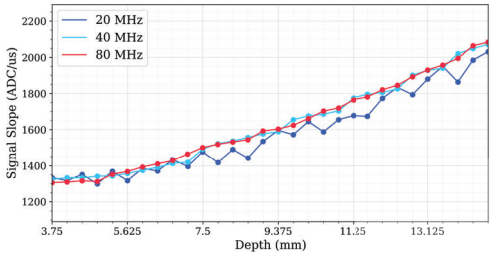
the drift slope is naturally expected to impact the timing resolution as a slow rising signal will be more obscured by noise.



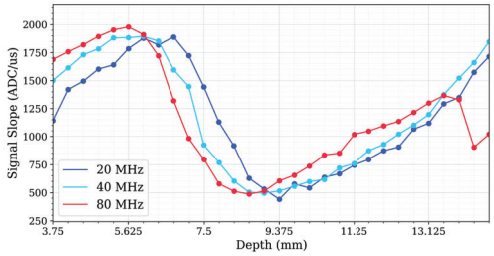
(a) Linear cathode SRFs (40 MHz)



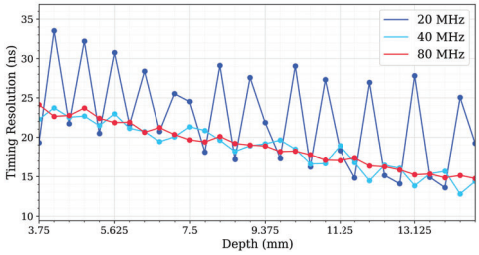
(b) Non-linear cathode SRFs (40 MHz)



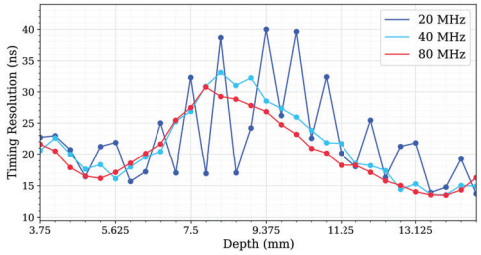
(c) Linear cathode slope variation



(d) Non-linear cathode slope variation



(e) Timing resolution depth sweep for (a)



(f) Timing resolution depth sweep for (b)

Figure 2.9: SRFs, slope variation, and timing resolution variation for two different cathode SRFs.

Figure 2.9 shows the timing resolution as a function of depth for two different cathodes and each sampling frequency using the method outlined. Figures 2.9a and 2.9b show examples of linear and non-linear cathode responses, respectively, as a function of depth. Figures 2.9c and 2.9d show the corresponding slope measurements as a function of depth. Despite being considered relatively uniform, the cathode response shown in Figure 2.9a still appears to have non-negligible slope variation. Figures 2.9f and 2.9e show the timing resolution as a function of depth. For the non-linear cathode response shown in Figure 2.9b, the timing resolution degrades for depths corresponding to the middle of the crystal due to the severe non-uniformity.

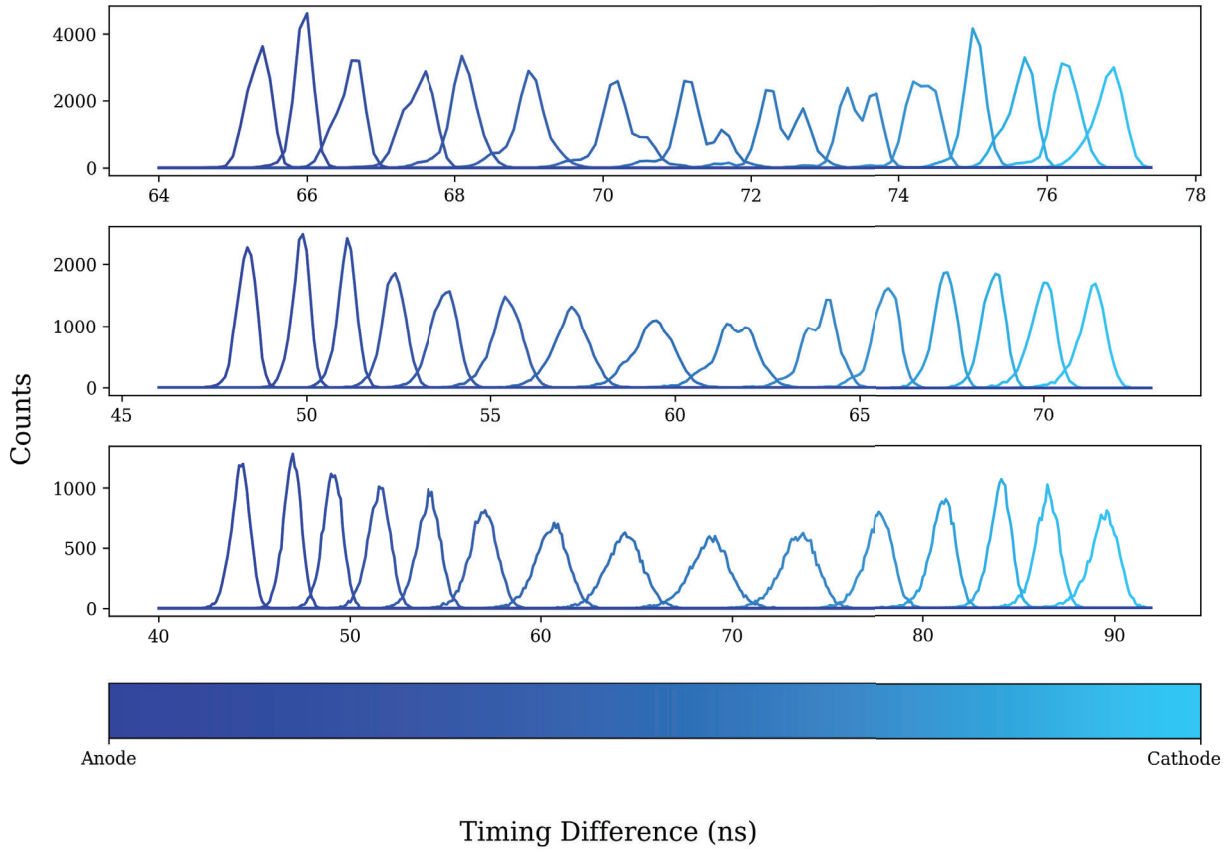
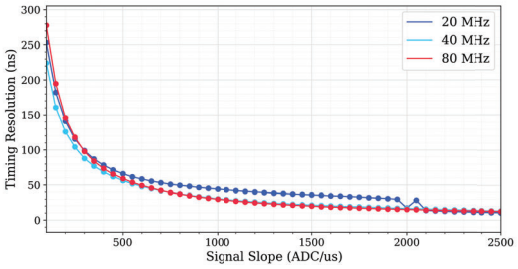


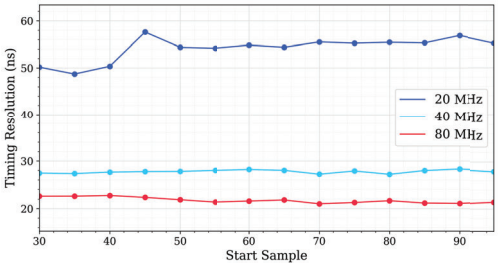
Figure 2.10: Timing spectra from anode to cathode corresponding to each point in Figure 2.9f. Top: 20 MHz, Middle: 40 MHz, Bottom: 80 MHz.

The erratic performance of the 20 MHz sampling frequency is characteristic to the linear fitting technique. The slopes of the 20 MHz cathode signals are relatively sharp, and thus could achieve fine timing resolution. However, the effect of noise is more dramatic on the 20

MHz signals, since each sample is a significant (50 ns) change in time. Moreover, as shown in Figure 2.8, the 20 MHz sampling frequency should suffer from higher noise. The result is that the drift start sample distributions from the 20 MHz setting become bi-modal when using this noise evaluation technique, as displayed by Figure 2.10. The bi-modality is due to the fact that the slope fit must use an integer number of samples, and the algorithm for determining the lower bound of the fit will then be altered by integer samples due to noise effects. Then, two peaks form corresponding to the different slope-fit bounds found by the linear fitting algorithm. The bi-modal effect can be observed only slightly in two of the 40 MHz spectra, while it appears to completely disappear in the 80 MHz spectra. This is explained by the magnitude of the noise relative to the time per sample. For the 20 MHz setting, each sample is 50 ns while the noise may go as low as 30 ns – indicated in Figure 2.11a – which yields a clear distinction. For the 80 MHz setting, each sample is 12.5 ns, but the noise only approaches $\approx 20\ ns$, so the effects described will blend together. Ultimately, it is the accuracy of the method that matters, and the artificially low FWHM manifested by the bi-modal effect does not capture this.



(a) Slope Variation:
Start sample = 45



(b) Start sample variation:
Slope = 1400 $\frac{\text{ADC}}{\mu\text{s}}$

Figure 2.11: Timing resolution measured using the ideal signal slope and sample start variations methods.

Finally, the ideal signal method is used, as shown in Figure 2.11, to verify the degradation contributions from the slope and start sample variation. Figure 2.11a indicates the expected strong correlation between slope and timing resolution, whereas Figure 2.11b shows that, despite the noise variation by sample shown in Figure 2.8, the start sample actually appears to have no considerable effect on the timing resolution. Note that the drop in FWHM for the 20 MHz setting shown towards the end of Figure 2.11a is again due to the bi-modality issue. This result also helps to clarify that 40 MHz and 80 MHz settings should achieve

similar timing resolution.

2.5 Na-22 Coincidence Measurements

Measurements at 20 MHz, 40 MHz, and 80 MHz sampling frequencies were conducted using a $\approx 1\ \mu Ci$ Na-22 source in order to evaluate the timing resolution performance in the Orion-α and Orion-β coincidence system. The reason for using a relatively weak Na-22 source was to prioritize true coincidence events, and to reduce any degradation effects due to chance coincidence or pile-up. An example experimental setup is shown in Figure 2.12, although in practice, the detectors were moved as close together as possible to increase the count rate. The two CZT planes were positioned antiparallel, and the Na-22 source was centered between the planes. Each experiment was run for 18 hours to collect sufficient statistics.

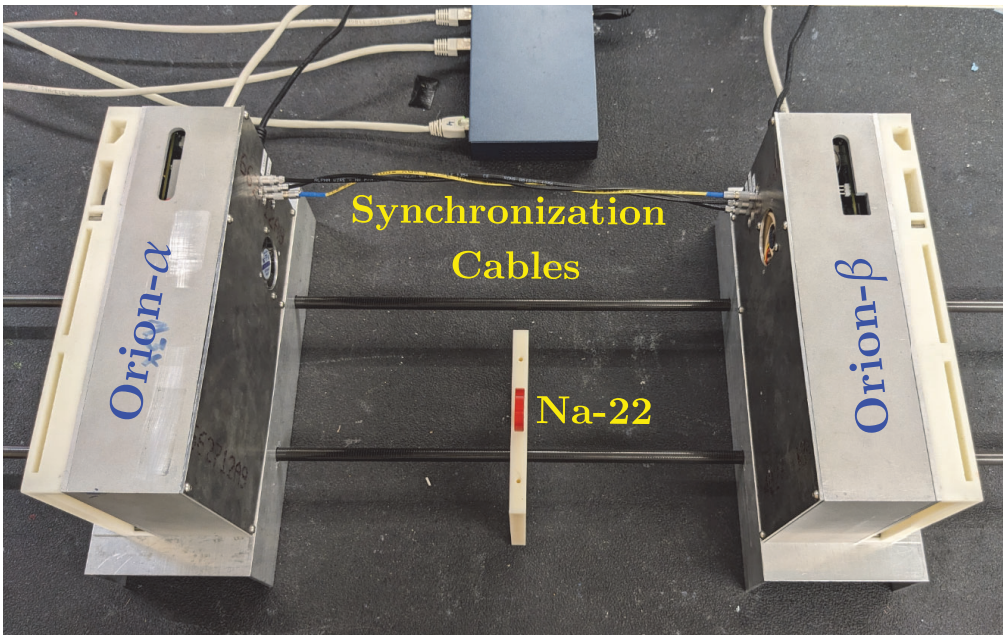


Figure 2.12: Orion-α and Orion-β coincidence Setup for Na-22 Measurements.

Using the same pick-off methods outlined in Section 2.3, the experimental data was processed and the corresponding version of Figure 2.11a was recreated as shown in Figure 2.13a. In experiment, four timing values are measured as opposed to the one cathode pick-off used in simulation. The timing values consist of two cathode pick-offs, and two timestamps for each respective detector system. The timestamps are clocked on the master clock, and correspond to the time at which each system's anode triggered. To determine the exact time

difference between the two interactions, Equation 2.1 is used.

$$\Delta T = T_{mclk} \times (t_{s_0} - t_{s_1}) + T_{samp} \times (t_{c_0} - t_{c_1}) \tag{2.1}$$

$$\sigma_{\Delta T} = \sqrt{T_{mclk}^2 \times (\sigma_{t_{s_0}}^2 + \sigma_{t_{s_1}}^2) + T_{samp}^2 \times (\sigma_{t_{c_0}}^2 + \sigma_{t_{c_1}}^2)} \tag{2.2}$$

Here, T_{mclk} and T_{samp} are the periods of the master clock and the sampling clock, respectively; t_{s_0} and t_{s_1} are the timestamps for each system; and, t_{c_0} and t_{c_1} are the cathode pick-off samples for each system. The expected uncertainty, $\sigma_{\Delta T}$ is a function of the uncertainty associated with the timestamps and the cathode pick-offs as shown in Equation 2.2. For this experiment, the timestamps are taken to exactly represent the time of each anode trigger. However, the anode threshold crossing can occur at any time within the clock period preceding the reported timestamp, which means there is some error associated with this assumption. To estimate the incurred error, it is assumed that the anode trigger occurs uniformly randomly within the clock period. Then, the uncertainty for each timestamp is just the uncertainty of a uniform random variable, or $\frac{1}{\sqrt{12}}$. The master clock was run at 80 MHz for this experiment, which means the uncertainty contribution from the timestamps is given by Equation 2.3.

$$\sqrt{T_{mclk}^2 \times (\sigma_{t_{s_0}}^2 + \sigma_{t_{s_1}}^2)} = \sqrt{(12.5 \text{ ns})^2 \times 2 \times \frac{1}{12}} = 5.1 \; ns \tag{2.3}$$

Since 5.1 ns is well below the expected timing resolution for this experiment, as shown in the previous section, this contribution can be considered negligible. Then, the simplified expected timing resolution is given as follows:

$$\sigma_{\Delta T} \approx T_{samp} \sqrt{\sigma_{t_{c_0}}^2 + \sigma_{t_{c_1}}^2} \tag{2.4}$$

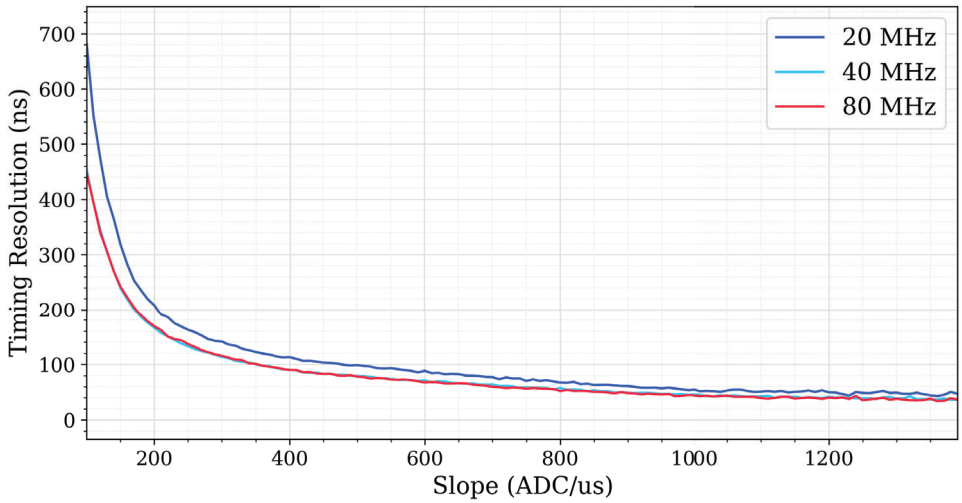
The result shown in Equation 2.4 highlights the fact that the expected timing resolution is a function of *both* cathode slopes, and therefore a different timing resolution is expected for all unique pairings of cathode slopes. To use the simulation data to accurately verify the experimental results, two methods may then be used. First, the experimental data may be binned into a 2D structure based on both measured cathode pick-offs. Then, from the simulated data shown in Figure 2.11a the expected timing resolution for each pairing can be calculated and compared to the measured resolution. The issue with this method is that the statistical requirement to generate a full 2D binned structure is too high for the current

measurement data. Since the source strength for this experiment was only ≈ 1 μCi, it would take a significantly longer measurement period to generate enough to data. In the future, the results may be verified more accurately by taking a measurement with a stronger source activity. Another method that may be investigated is to bin the experimental results based on the expected timing resolution determined by the simulated data on an event-by-event basis. However, it is preferred to keep the simulation and experimental results independent.

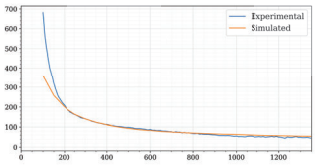
In experiment, it was decided to use an approximation in which the minimum of the two cathode slopes is used as the binning variable to determine if the results roughly match the simulated values. For coincidence events where the minimum slope is small, the greater of the two slopes can range anywhere from being similar to significantly higher. Then, those low-slope bins may end up being skewed to lower average timing resolutions. For events with a high minimum slope, the corresponding greater slope will tend to be similar, so the high-slope bins should approximately reflect the expected results. Since it is the timing resolution limit – or high slope domain – which is of interest, this approximation is considered appropriate for the current experimental dataset. To compare the simulated and experimental results, the simulated values are multiplied by $\sqrt{2}$, again by approximation. Note that the experimental definition of cathode slope is the slope of the fitted region.

Direct comparisons between the experimental observations and predicted performance are shown in Figures 2.13b through 2.13d. Notably, each of the sampling frequencies show relative agreement with the expected resolution in the high-slope limit. Here, error bars are not included due to the inherent uncertainty in the simulation methods, and difficulty with estimating error in the experimental method. As discussed in Section 2.4, the SRF method is a useful estimate of the expected performance, but does not perfectly capture all effects that may occur on an event-to-event basis. The ideal signal method is a rough estimate, since the real signals will contain non-linearity. In experiment, Figure 2.13a is generated by binning the events based on the lower of the two cathode slopes involved in a coincidence event, and then fitting a Gaussian to the timing peak in each bin. To generate error bars for each timing bin would require several repetitions of the experiment, which is not practical since each iteration requires 18 hours. The result shown in Figure 2.13 is intended to prove that the experimental results are approximately limited by the system noise.

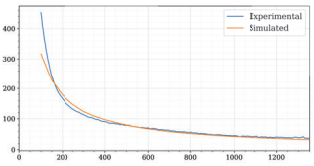
The achievable timing resolution in CZT is measured for all events, photopeak – or, "high-slope" – events, and for the best modules. Each data selection is relevant for different experimental situations. In typical PET systems, full energy depositions are primarily used to generate the image to reduce background artifacts. This constraint also helps to achieve



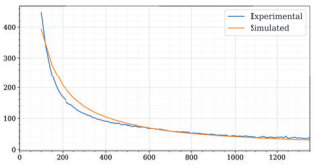
(a) Experimental slope variation



(b) 20 MHz comparison



(c) 40 MHz comparison



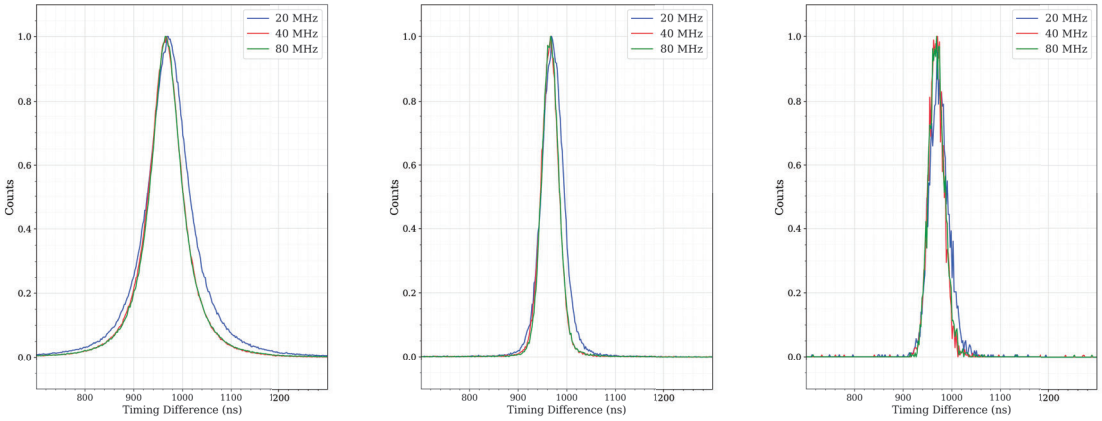
(d) 80 MHz comparison

Figure 2.13: Comparison of experimentally measured slope variation to simulated slope variation.

the best timing resolution, since the signal magnitude is the greatest. However, one benefit of using CZT in PET systems is the ability to – in many events – not only distinguish the position and energy of independent interactions in Compton scatter full-energy depositions, but to also determine the initial interaction in those events. The advantage of using this capability is to improve the system sensitivity, but at the same time the timing resolution will degrade due to the lower signal amplitude of partial depositions. This offers a trade-off in which a greater efficiency may cause the timing resolution to trend toward the all events category presented in Table 2.1. The best module performance captures the timing resolution between the two modules with presumably the lowest cathode noise and most linear cathodes. Note that events with slopes greater than 1000 $\frac{\text{ADC}}{\mu\text{s}}$ roughly correspond to events with energy > 500 keV, although there is not necessarily a one-to-one relationship.

Table 2.1: Summary of Timing Performance for Various Sampling Frequencies

Event Category	Timing Resolution Performance (ns)		
	20 MHz	**40 MHz**	**80 MHz**
All Events	105.7	86.8	84.6
High Slope Events (>1000 ADC/μs)	51.2	41.7	40.5
Best Module and High Slope	45.4	36.3	37.7



(a) All events spectra (b) High slope spectra (c) Best module spectra

Figure 2.14: Summary of timing resolution performance for 20, 40, and 80 MHz sampling frequencies.

Figure 2.14 and Table 2.1 summarize the measured timing spectra and FWHMs achieved in each category. As expected, 20 MHz performs the worst, while 40 MHz and 80 MHz are similar, with 80 MHz achieving the best resolution in most cases. Although the best module performance at 40 MHz appears to be slightly better than at 80 MHz, the spectrum shows that the two are almost identical. Using the best module pairing significantly reduces the usable statistics, and this is likely responsible for the discrepancy between the two sampling frequency results.

2.6 Cathode Noise Limiting Factors

Thus far, the cathode noise is the main limit to the timing resolution, so it is necessary to outline the factors relating to the optimization of both the cathode noise and signal. In this case, the signal is the cathode slope, which is defined as the induced charge over the drift time. As discussed in Section 1.3, the cathode signal is a linear function depending on the depth of interaction. The drift time is defined by charge velocity, which is a function of the mobility and electric field. Thus, the slope can be defined by Equation 2.5 where q is the deposited event charge, D is the detector thickness, V is the applied bias, and μ is the electron mobility. Note that the units for Equation 2.5 are $\frac{C}{s}$. The result highlights two parameters of interest: the applied bias, and the detector thickness. From a signal perspective, the thickness should be minimized, and the applied bias should be maximized.

$$\text{slope} = \frac{Q}{t_{drift}} = \frac{q(\frac{d}{D})}{\frac{d}{\mu \mathbf{E}}} = \frac{q\mu V}{D^2} \tag{2.5}$$

Next, the various sources of noise are considered. The noise sources are broken into parallel, series, and $1/f$ noise. The parallel noise contributors consist of the detector leakage, the feedback capacitance, and the charge amplifier input node. The series and $1/f$ noise sources are only from the charge amplifier input. Reverse biased CZT is considered to behave like a resistor, which means that the leakage current is defined by the applied bias and the detector dimensions. The formula for the leakage noise spectral density in units of $\frac{A^2}{Hz}$ is shown in Equation 2.6, where e is the electron charge, D is detector thickness, and A is the detector area. Here, to reduce leakage related noise, the detector thickness should be increased, and the applied bias should be decreased, contrary to the result shown in Equation 2.5 prior. The leakage formula also indicates that the detector area should be minimized to reduce leakage noise. Leakage has also been shown to be a function of the CZT temperature, which is not modeled in Equation 2.6 [7]. To reduce the leakage further, the CZT can be operated in a controlled temperature environment.

$$S_L = 2eI_L = 2e\frac{V}{\rho\frac{D}{A}} \tag{2.6}$$

Apart from leakage, the other noise factors – front-end amplifier and feedback capacitance – can be considered constants and neglected for the purpose of this discussion. However, the series noise from the front-end amplifier will be contributed across the total capacitance at the input node of the amplifier. Again, the front-end amplifier and feedback capacitance

can be considered constant, but the capacitance contributed by the CZT itself will vary as shown in Equation 2.7. This result also suggests that detector thickness should be increased, and area should be decreased.

$$C_{CZT} = \frac{k\epsilon_0 A}{D} \tag{2.7}$$

The relationship between the slope magnitude, noise factors, and the resulting timing resolution is challenging to establish analytically. No transfer function can be associated with the linear pick-off method, which makes it difficult to assess the theoretical impact of noise. However, the relations shown in Equations 2.5 through 2.7 reveal that the detector thickness, area, and applied bias are factors that need to be experimentally explored to understand the optimal settings for cathode noise. Reducing the detector thickness and area also comes at a cost to efficiency. The same efficiency may be maintained by including more CZT crystals, but this then comes at the cost of electronic complexity and power. Experiments with varied applied bias and detector thickness are recommended for future work to provide insight into the quantitative trade-offs.

In addition to the theoretical contributions, the cathode noise can be attributed to several practical system features, including:

- The filtering scheme used for the high voltage trace,
- System grounding configuration,
- Cathode to high voltage connection quality,
- High voltage trace wear, degradation, and cleanliness,
- High voltage source noise,
- Interference from digital signals.

From a design perspective, the most important aspect for cathode noise is the filtering scheme. The filtering scheme used to control the high voltage trace must remove any potential sources of pick-up. Usually, a strong RC filter with $\tau \approx 0.1$ s is placed physically at the location where the high voltage trace enters the high voltage enclosure to remove any external pick-up. At least two more RC filters are included between each cathode connection made. The connection itself must be made cleanly, with no added resistance to the cathode surface. The system grounding configuration will always play a key role in the electronic noise, and

must also be carefully considered. Usually, this should be enough to keep the cathode noise under control, but for the Orion-α and Orion-β systems used in this study, many of the other considerations noted above will play a role. Since Orion-α and Orion-β are research systems, they have been deconstructed, modified, and used in various circumstances over the years. When humidity levels are high, sparking has been observed, requiring high voltage components to be replaced. Over time, the high voltage traces acquire significant wear and gather debris which degrades the high voltage signal quality. Additional sources of degradation include noise from the high voltage source itself or interference from digital signals in the system, although in Orion-α and Orion-β, these are not considered to be present. For commercial systems, issues like trace degradation and interference should be eliminated.

2.7 Conclusions and Future Investigation

The limiting timing resolution of CZT in the current Orion-α and Orion-β research system was evaluated and compared to the estimated limit as a function of system noise. This investigation serves as one component of a feasibility study of CZT for the use in PET systems. CZT offers advantages over state-of-the-art systems due to its excellent energy and spatial resolution capabilities. Although it's known that the timing resolution of CZT will not be capable of TOF PET, the limits of the timing resolution were not yet well understood, motivating this study.

The primary experimental variable for the VAD-UMv2.2 ASIC-based systems is the sampling frequency, which may be set to 20, 40, or 80 MHz. Using the linear pick-off method to determine the cathode drift start, the best timing resolutions were found to be 45.4, 36.3, and 37.7 ns, respectively. Using experimentally measured noise in combination with SRFs and ideal linear signals, the measured timing resolutions appeared to show relative agreement to the limit set by system noise. Thus, it is critical to reduce the system noise if the best timing resolution is to be reached. To determine the optimal detector configuration with respect to system noise, the impact of detector thickness, detector area, and applied bias should be empirically modeled by a future experiment.

A remaining, important investigation is to determine the limiting sensitivity of CZT, which is another fundamental parameter that determines the system viability. While the intrinsic efficiency of CZT is not necessarily in question, the readout electronics and maximum count-rate are the challenging optimizations that remain. The maximum drift time

in 15 mm-thick CZT biased at -3000 V is currently on the order of ≈ 1 μs which suggests that event rates greater than 10^6 cps are likely impossible barring sophisticated pile-up reconstruction. In experiment, the effective noise-equivalent count-rate – or rate of useful true coincidence events – will be far below this limit due to readout time, and losses relating to chance coincidence and scatter background. The VAD-UMv2.2 used for this study, for example, is limited to the order of 10^4 cps when using triggered and neighboring anode waveform readout. Then, after removing chance coincidence and scatter background, the best noise-equivalent coincidence rate measured from the Orion-α and Orion-β system is only ≈ 400 cps so far. State-of-the-art fast scintillators are capable of noise-equivalent count-rates as high as 306×10^3 cps [29], a rate then equivalent to an unreasonably high 13770 CZT crystals. However, certain compromises must be investigated to understand what the optimal CZT-based system can achieve. The H3DD-UM ASIC, for example, includes the capability to modify the number of samples read out from each waveform. This may allow the user to strike a balance between the digital waveform information offered by sampling ASICs, and the maximum count-rate expected from the analog ASICs. Another angle for optimization is the detector thickness. By reducing the maximum drift time, the acquisition time of each waveform can be decreased, the slope will increase if the bias is kept the same, and the detector saturation rate will increase. On the other hand, decreasing the detector thickness will start to reduce the interaction rate. A search for the optimal thickness may yield interesting results. Further investigation on this optimization is necessary to fully evaluate the merit of pixelated CZT for PET systems.

CHAPTER 3

65nm Design Methodology

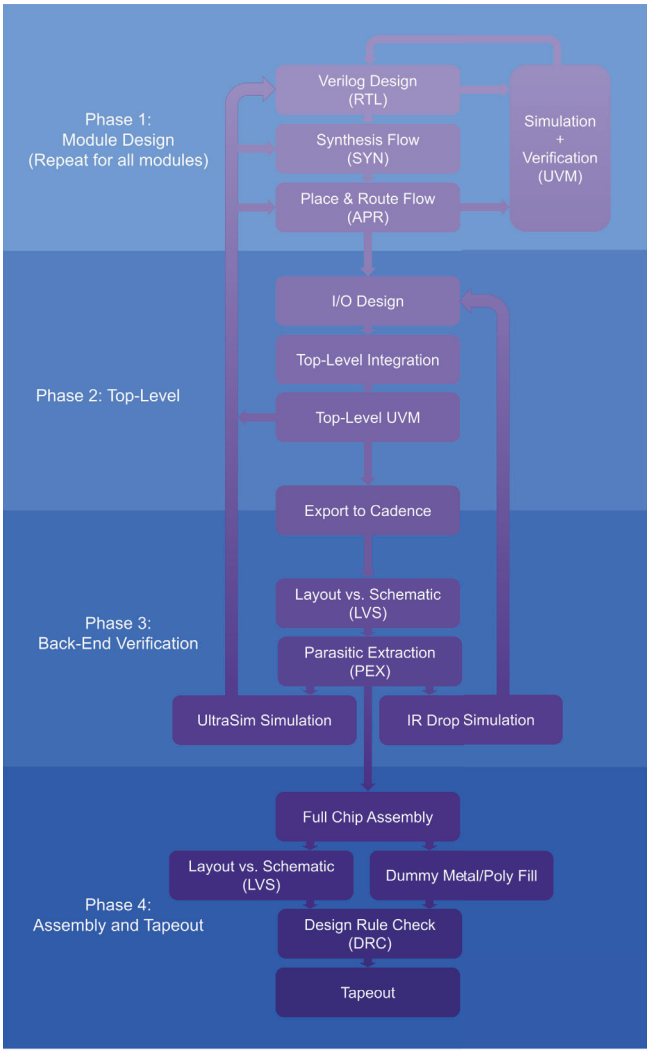


Figure 3.1: 65nm design flow overview.

In the following chapters, the design of a DAQ-DSP ASIC is outlined in detail as one of the major components of this work. The DAQ-DSP ASIC was designed using the 65nm MS RF GP TSMC process, and follows a standard design flow of synthesis, automatic place-and-route (APR), and back-end assembly. A nine-metal stack-up including metal one, six standard layers, one thick layer, and one ultra-thick layer was used. The different metal layers are referred to as M1 through M9, with M9 being the ultra-thick layer. The design flow is illustrated in Figure 3.1 and will be described in this chapter.

3.1 Phase 1: Module Design

The initial module design loop begins when the chip functions have been well-defined. The top-level chip is divided into several abstraction layers in which the base consists of simple, coherent register transfer language (RTL) modules, and, for the most part, intermediate layers only interface those blocks together in various ways.

All individual RTL blocks are cycled through the module design loop shown in Figure 3.1. First, the module functionality is written into Verilog. At the same time, the testbench and test cases are elaborated to help constrain the design. The industry standard Universal Verification Methodology (UVM) was used to execute the testbench. UVM is a SystemVerilog based flow which outlines standard logical blocks to be contained in each testbench. The principles underlying UVM are re-usability and reliability. Standardized modules should be able to be used in other designs as applicable, and the UVM testbench should rigorously confirm the intended module behavior. The standard components that make up the UVM testbench are detailed as follows, and the relations are shown in Figure 3.2.

- **Interface:** The interface contains the instantiations of all input and output signals of the device-under-test (DUT).
- **Sequencer:** The sequencer receives sequences of items – known as sequence items – and passes those items to the driver.
- **Driver:** The driver receives sequence items and drives them through the interface using a user-defined RTL routine.
- **Monitor:** The monitor continuously watches the interface to record signals of interest and send signal-packets to the scoreboard.

- **Scoreboard:** The scoreboard receives signal packets from the monitor and contains user-designed functions that check whether the output is correct with respect to the inputs.

In addition to the components listed, there are also two additional abstraction boundaries: the agent and the environment. The agent encompasses the sequencer, driver, and monitor. The purpose of abstracting the agent is that, if the same DUT is used anywhere else in the design, the agent can be inserted directly into other parts of the design, enabling re-usability. The environment encompasses the agent and the scoreboard. For certain designs, the communications between the monitor and the scoreboard, or the scoreboard design itself, may change. For example, in a more complicated system, the scoreboard may link together monitor outputs from multiple devices to verify the overall functionality. However, if the DUT remains independent when placed into another system, then the entire environment can be directly reused. Each base-level module must have an associated UVM testbench with rigorous test cases. The UVM testbench and the RTL are developed in tandem until the module achieves the expected functionality and passes all test cases.

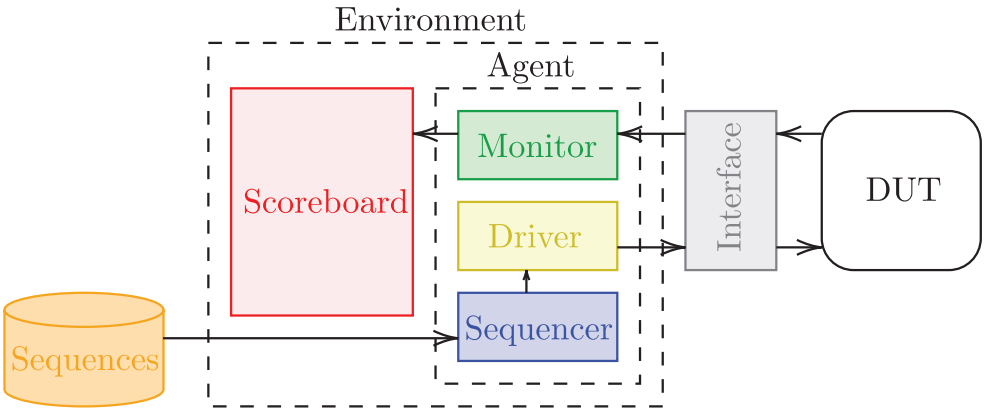


Figure 3.2: Standard UVM component block diagram.

Once the module meets the specifications, the Verilog is synthesized, translating from behavioral RTL to structural gate instantiations. In the synthesis flow, the critical step is to outline the design constraints, such as the desired clock frequency. Depending on the design, more specific constraints may be required, and the synthesis may be optimized against area or power. Although the testbench can be run at the synthesis step, it is most common to only run the testbench again after the final APR step.

APR converts the list of gate instantiations and connections into a layout on silicon. Several critical steps are executed:

1. **Power Ring and Power Mesh**: The power routing layers are chosen, and VDD and GND are routed in a ring and stripes over the design area.
2. **Placement**: All gates are placed in silicon.
3. **Clock-tree Synthesis**: The clock-tree is synthesized based on the defined frequencies. Buffer insertion is used to balance drive capabilities, and skew is optimized.
4. **Filler Insertion**: Standard filler cells are added to meet density constraints.
5. **Routing**: The critical routing step is executed. Several iterations are run to route connections between modules based on timing, signal integrity, and design rules.
6. **RC Extraction and Output**: The design is finalized, and the RC delay parameters are fully extracted to a standard delay format (SDF) for use in simulation. The netlist and linkable executable file (LEF) are exported for simulation and back-end flow.

When all steps of APR are satisfactorily complete, the module design is ready for a final pass of the testbench. The exported SDF is linked to the UVM testbench to simulate the expected delays, and the same test cases are run. It is also critical at this stage to check the reported timing slack and clock tree parameters to ensure there are no violations. If any issues persist, the RTL and UVM is revised, and the cycle continues until all test-benches pass.

Figure 3.3 shows an example output of Innovus – the APR tool – for one of the filters used for the DAQ-DSP ASIC. The power ring and power mesh are routed on the top thicker layers, M8 and M9. Depending on the density of the design, even the power routing layers may be used for signal routing, as seen in the center of Figure 3.3. During the design flow, it is essential to push each module through the APR flow to determine the required area and to estimate how difficult the timing constraints of the design are to meet.

3.2 Phase 2: Top-Level Design

When phase 1 is complete for all modules, the final top level assembly flow is executed. For this design, a top-down flow is used, meaning that synthesis and APR are run at the

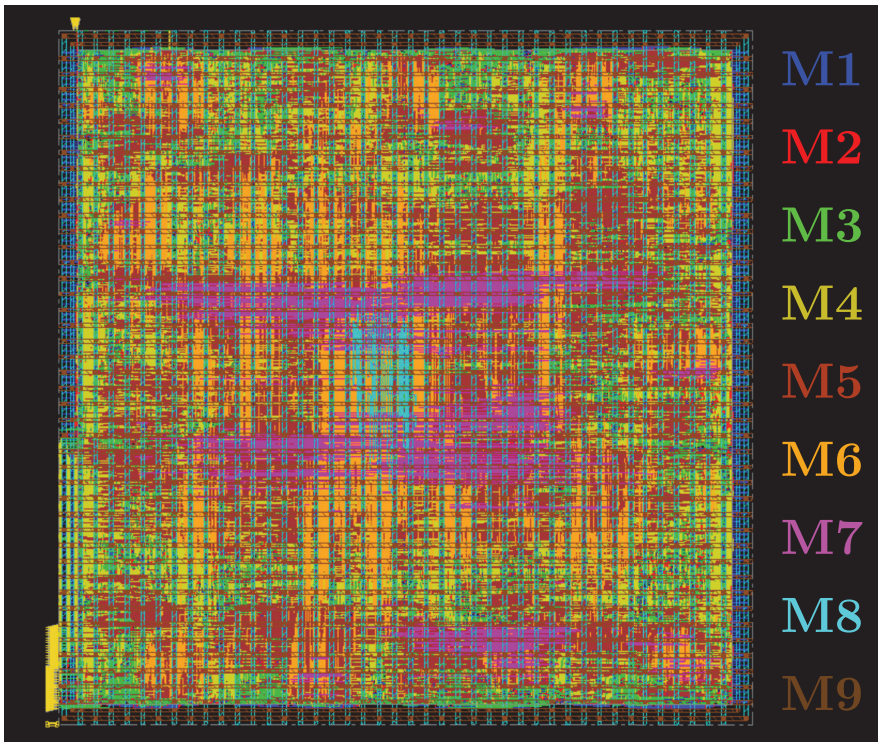


Figure 3.3: Sample of a completed APR result for a CRRC filter.

top-level without inclusion of any other synthesized or placed design files. In other words, the synthesis and APR tools are aware and have full control over all gates in the design. As long as sufficient computing resources are available, this is the recommended method.

At the top-level, the I/O cell layout is a new inclusion. Several design choices guide the I/O placement, including:

- Power distribution
- System-level PCB designs and ease of signal routing
- Internal constraints such as module placement and proximity to the chip boundary

A first pass of the I/O cells is done at this stage, and all modules are integrated in the top-level RTL file. An equivalent phase 1 flow is executed for the top-level: the UVM testbench is passed at the behavioral level, synthesis and APR are executed, and UVM is passed at the APR stage. The critical difference at the top-level is that the APR flow now reads in the core and pad cells as LEFs and places them in the floorplan. The routing phase of APR

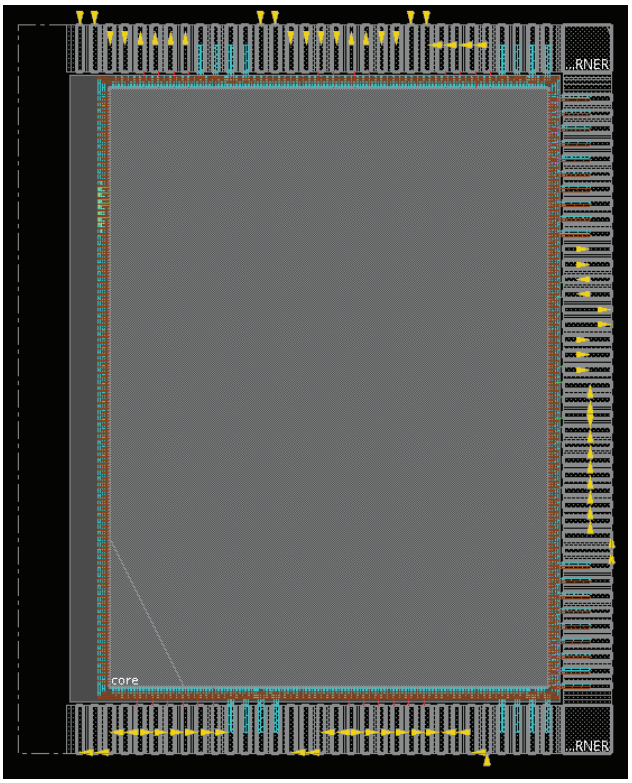


Figure 3.4: Sample of the completed top-level design for the DAQ-DSP ASIC.

simply links the I/Os to the core pins. The final design is exported to a GDS file which is imported by Cadence for the final back-end flow.

Figure 3.4 shows an example of a completed top-level design. The core design is shown as the central block, and then various I/O pad cells are instantiated around the core. The only shown connections are the signals being brought from the I/Os to the core pins – or power – and vice versa.

3.3 Phase 3: Back-End Verification

The back-end process is done in Cadence using Calibre to ensure that the design meets manufacturing constraints – such as metal density – and is connected as expected. Moreover, Calibre tools – parasitic extraction (PEX), e.g. – provide more accurate estimates of the performance than the strictly digital tools. An initial layout vs. schematic (LVS) is run to ensure that the netlist connectivity matches the Cadence extracted netlist. The LVS flow does a top-down, one to one comparison of the two following netlists to ensure equality:

1. The schematic-view file (`.cdl`) generated by running `v2lvs` on the post-APR netlist

2. The `.sp` netlist extracted from the layout view in Cadence

At this stage, several options are available for additional verification. There is technically no issue with skipping the remaining steps shown in phase 3 and moving to final back-end flow in phase 4 directly, but to be fully confident in the design, 3 additional simulations are recommended. First, PEX should be run in Cadence to provide an accurate SDF (or SPEF) representation of the RC delays associated with the digital circuit. The most accurate PEX process models parasitic resistances, parasitic capacitance and coupling capacitance (R+C+CC). This level of accuracy is usually not possible due to the magnitude of the digital circuit; however, it is typically possible to extract parasitic capacitance and coupling capacitance (C+CC) values, which are the most important sources of degradation. The extracted C+CC results are more accurate than those reported from Innovus. Using these extracted values, two simulation options are available: UltraSim – or any fast Spice simulator – and IR Drop simulations. UltraSim performs the same function as the post-APR simulation described in phase 1, but at a more accurate level. UltraSim can treat the signals as purely analog – which most likely requires far too significant computational resources – or as varying approximation levels to digital signals. Using UltraSim, a transient simulation on the scale of milliseconds may be completed to confirm the expected behavior.

The IR drop simulation – done using the Voltus IC Power Integrity Solution – is a valuable method for verifying the power distribution network. In particular, this simulation may inform the placement of the power pins in the I/O ring. For the DAQ-DSP, the IR drop simulation was skipped since an abundance of power pins were available for the design. However, in pin-limited designs, it may be necessary to run an IR drop simulation to ensure that no power hot-spots exist on chip. Figure 3.5 shows the IR drop (left) and layout (right) for a small section of one of the filters used in the DAQ-DSP ASIC. The left panel shows the overlaid IR drop, where brighter colors indicate that VDD is below the nominal level. In this example, the brightest spot – or the largest drop – is approximately 0.001 V, indicating that there should be no IR drop issues for this design which uses a 1.0 V rail. As discussed, the IR drop simulation becomes more valuable when done in the top-level design with I/Os to understand how to better place the I/O cells.

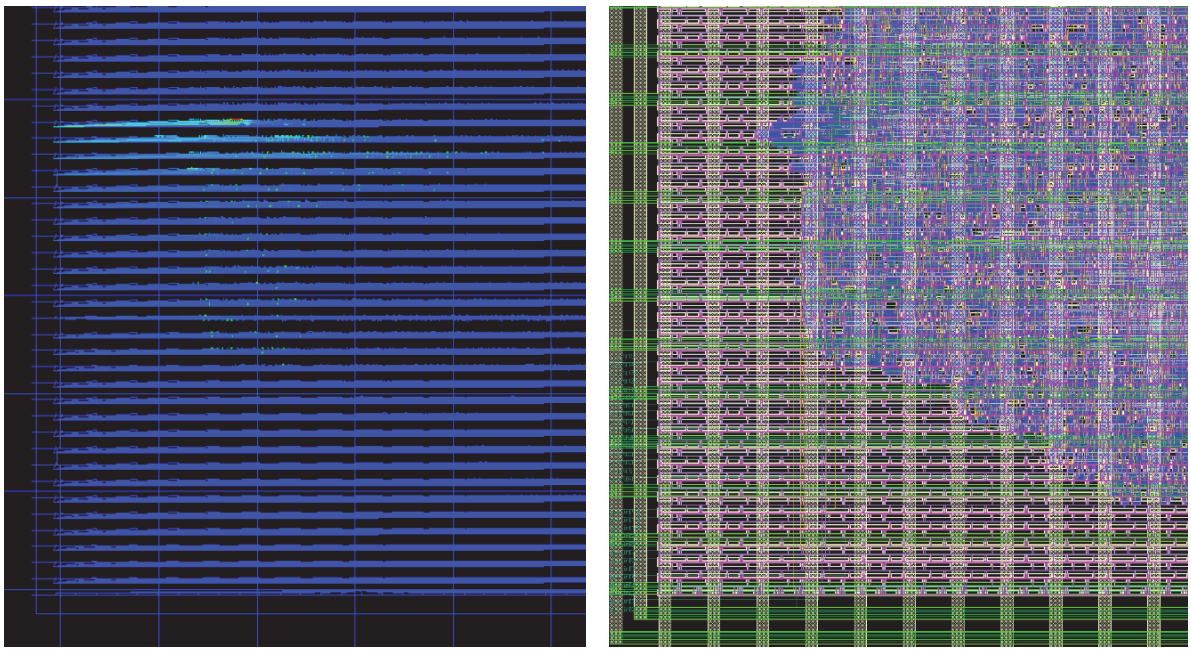


Figure 3.5: Left: Magnitude of IR drop over the layout. Right: Metal layers of the same sample layout.

3.4 Phase 4: Assembly and Tapeout

The final steps of the design flow involve the assembly of any separately tested and synthesized components. At this stage, the seal-ring is also added around the chip to set the final area.

Once the full-chip is assembled and verified, the remaining steps include a final run of LVS, dummy fill scripts, and design rule checks (DRC). LVS is run again using the final top-level schematic. Dummy metal and poly filler scripts are executed to ensure the density of all layers meets foundry requirements. Figure 3.6 shows an example of an ASIC ready for tapeout with the seal-ring and filler cells inserted. The zoomed in portion highlights that the empty spaces are filled with dummy metal and poly cells to meet foundry constraints. Finally, DRC is run with the foundry rulesets to ensure that the densities are satisfactory, and that there are no other fabrication issues such as antenna violations. When no DRC issues remain, the final design is exported to a GDS file and sent for tapeout.

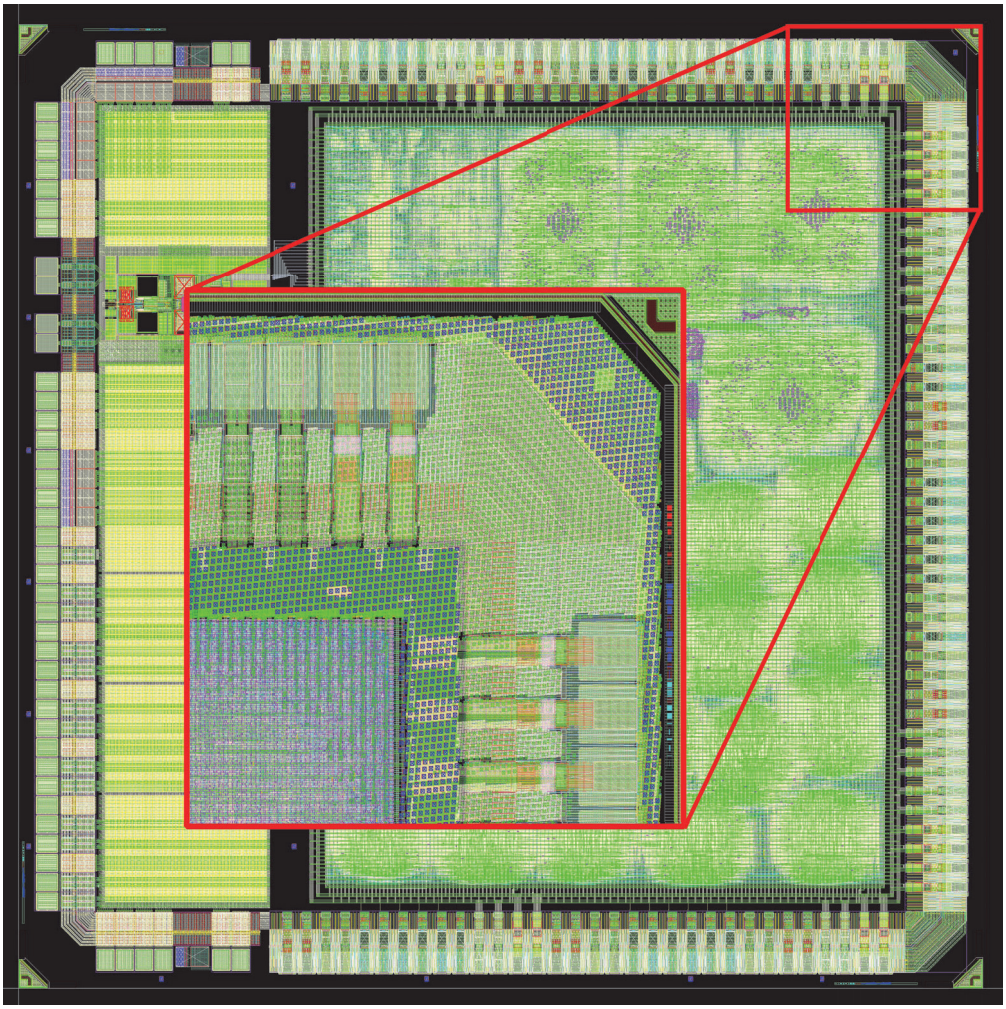


Figure 3.6: Tapeout ready ASIC with dummy fill metal and poly.

CHAPTER 4

DAQ-DSP ASIC Revisions 1 through 3

4.1 Motivation

3D-CZT readout systems traditionally consist of several electronic components including the front-end pre-amplifier ASIC, an ADC, an FPGA, and a CPU – usually in the form of a personal computer – to completely process the event stream into a clean and calibrated energy spectrum. The Orion-α and Orion-β systems described in Section 2.2.1 are examples of the standard processing chain. However, to best take advantage of the 3D-CZT capabilities, it is desirable to implement a compact and low-power processing system. The ADC, FPGA, and CPU are all components which have high-quality, commercially available options, but the resulting system size is large and power consumption is significant. It is possible to reduce the size and power consumption by taking advantage of a custom ASIC design. Moreover, the digital signal processing flow used for 3D-CZT has become relatively standardized, as noted in Section 1.4. With only a handful of parameters, it is conceivable to mix together fixed filters with a degree of flexibility to create an ASIC which offers all the desirable functionality. Figure 4.1 conveys the reduction which motivates the DAQ-DSP ASIC project. The DSP replaces the CPU-based waveform processing, the DAQ replaces the FPGA controller, and the ADC is combined on the same chip.

In total, four revisions of the DAQ-DSP ASIC have been completed during the initial project contract. While the bulk of the design and results presented in this work will relate strictly to the 4th revision, it is important to preface those results with the initial three revisions. The motivation for design choices made regarding the first three revisions of the DAQ-DSP ASIC are outlined in [8] and [21]. Although several modifications were made to the design following the 2nd revision and enabling the 3rd revision to successfully provide interesting results, the fundamental blueprint of the chip did not change until the 4th revision. Thus, only a brief, high-level description of the DAQ-DSP is provided here, and the remaining

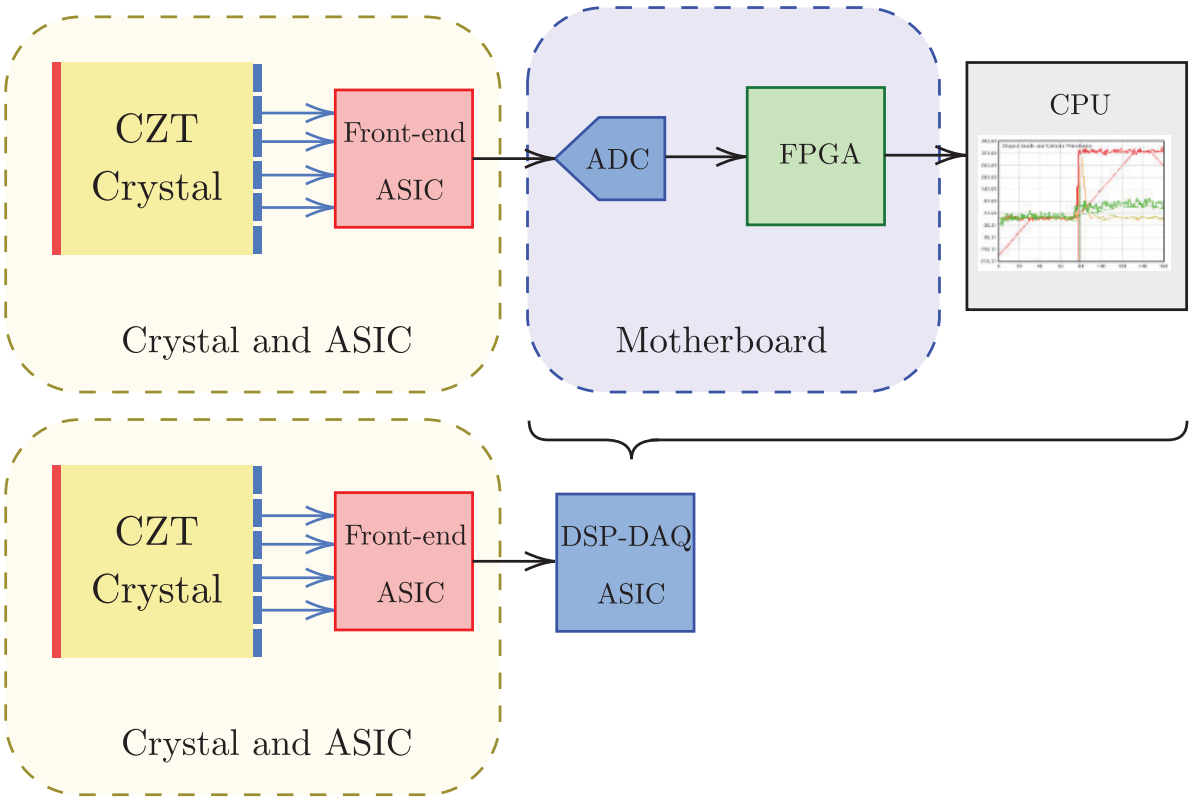


Figure 4.1: The goal of the DAQ-DSP ASIC is to bring the ADC, controls, and waveform processing onto one chip.

details can be found in [8] and [21]. Then, a description of the revisions made for the DAQ-DSPv3, and the results attained will follow. The 4th revision DAQ-DSP functionality and design will be discussed in granular detail in the following Chapter 5.

4.2 General Description

The DAQ-DSP ASIC consists of the two main sections: the DSP Core, and the DAQ, which includes an on-chip ADC. The ASIC interfaces with the H3DD-UM ASIC, which is a front-end chip responsible for acquiring analog waveform samples from a CZT crystal. The DAQ consists of a component, called the H3DD Core, which controls the H3DD-UM front-end ASIC, and a 13-bit ADC. The 13-bit ADC is responsible for converting the analog samples into digital samples. The DSP Core receives digital samples from the ADC and performs a variety of filtering operations to determine the amplitude, trigger timing, and type of the

incoming waveform. The ASIC described in this document can be used in place of FPGA controllers and CPU-based waveform processing in order to significantly improve the power cost, area cost, and processing speed.

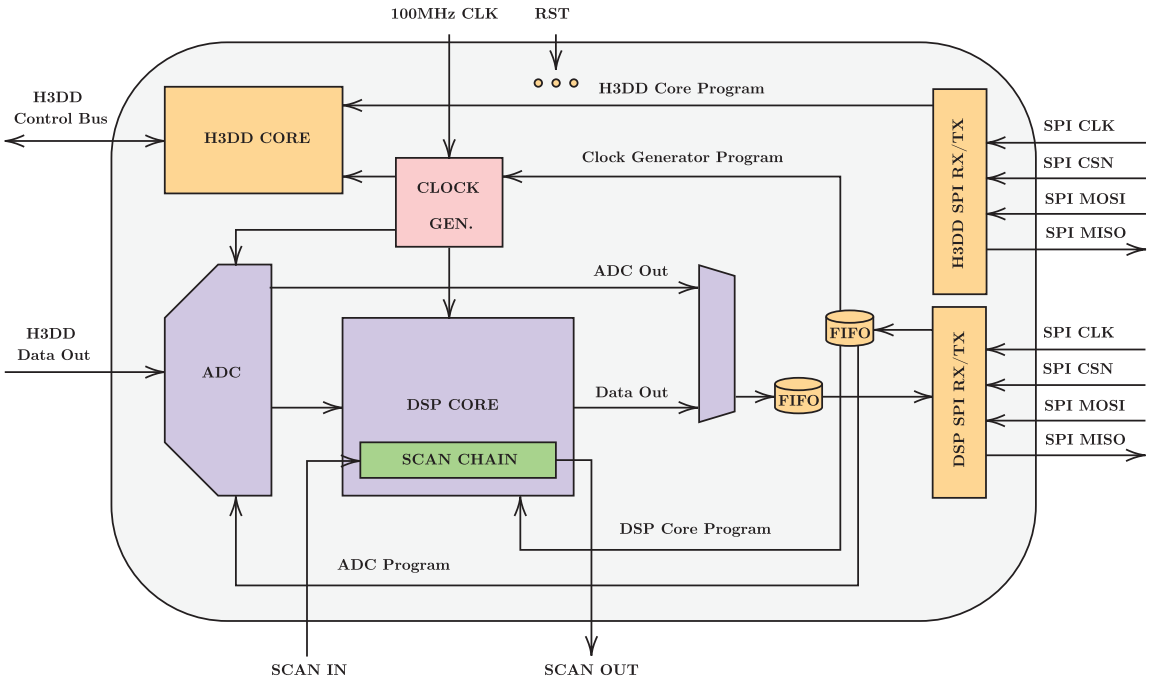


Figure 4.2: DAQ-DSPv3 ASIC top-level functional diagram.

A top-level diagram is shown in Figure 4.2. The flow of data follows the path indicated by the dark blue components. During operation, the H3DD-UMv4 ASIC feeds analog samples to the on-chip ADC that then converts those samples into 13-bit digital samples. The digital samples are directly fed into the DSP Core which is responsible for all signals processing functions. At the output of the DSP Core, several data are released for each waveform. The waveform amplitude and timing – the primary processed results – are passed to the output as 16-bit integer values. Meta-data including the channel number, trigger information, and timestamp are formatted together with the amplitude and timing to make a 6×32-bit word output packet. An alternate option is available to directly send the ADC output to the data first-in first-out (FIFO) buffer. The output packet is transferred from the FIFO to the outside world via the serial peripheral interface (SPI) bus.

All control related components are shown in orange. Two SPI transceiver blocks serve as the communication interface between the DAQ-DSP ASIC and the user. The use of two

SPI channels is motivated by the different SPI conventions used by the H3DD Core and the DSP related components. The H3DD Core is an inherited component from the FPGA based control system RTL, designed by Zhu. To reduce the complexity of the project, it was deemed more effective to leave the H3DD Core exactly as it was in the FPGA system. Within the H3DD Core, multiple SPI endpoints exist including direct throughput to the H3DD-UM ASIC, configuration registers, and control bit registers. Among those endpoints, the sending convention even varies slightly. As such, a completely different method was used on the DSP side. The DSP SPI channel responds to standard 32-bit words which are divided between the address and data package as elaborated in [8]. The endpoints associated with the DSP SPI are shown in Figure 4.2: the clock generator, the DSP Core, and the ADC. An RX FIFO and a TX FIFO are used to moderate the data flows between the SPI bus and the DAQ-DSP. Since the reset signal drives all internal reset patterns, the connections are not shown in detail.

The clock related blocks are shown in red. The chip is run from a single 100 MHz clock, and internal clocks are generated using the master clock. The main internal clocks are the ADC clock, the DSP clock, and H3DD-UM related clocks. Further details on the clocking scheme are included in [8].

Finally, the primary debugging capability of the 3rd revision is shown in green. The DSP scan chain allows the user to ensure that all registers of the DSP Core are working as expected. However, the scan chain function also flushes all the programmable parameters out of the DSP which means that the scan chain can only be used to verify connectivity of internal registers, and it cannot be used to verify that the DSP Core is programmed as expected during operation. This drawback motivates revisions that are discussed in Chapter 5.

4.3 Revisions 1-2

The 1st and 2nd revisions of the DAQ-DSP ASIC were taped out on Mar. 17th 2021, and Dec. 7th 2022, respectively. The initial revision first suffered from an error related to bringing the core voltage onto the chip. It was found, using fine-tip probes, that the resistance between different core VDD pins was as high as 5.5 $M\Omega$. This showed that the VDD pins were not connected on the chip. Focused ion beam (FIB) modifications were made after fabrication to connect metal layers 8 and 9 in the VDD pads, enabling the chip to power on. It was subsequently found that the critical control sequences for the H3DD-UM ASIC, debugged by

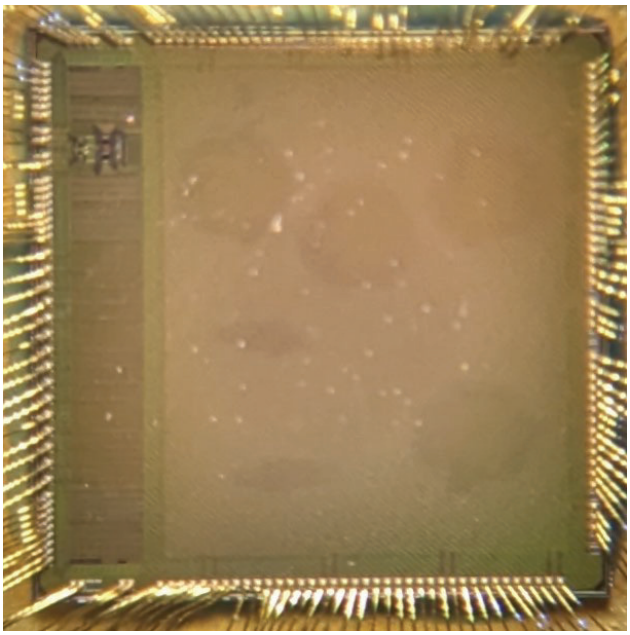


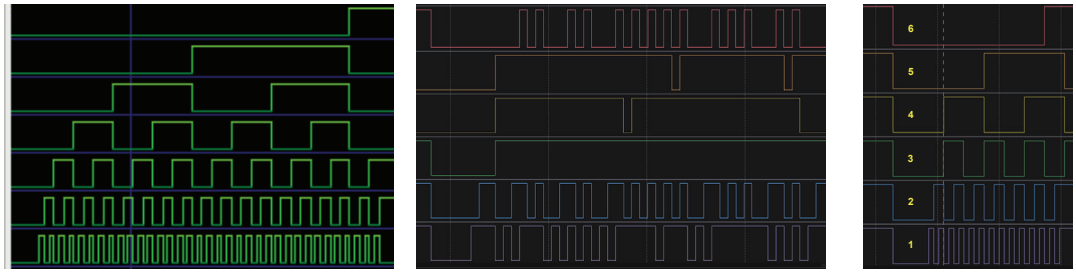
Figure 4.3: DAQ-DSPv2 sample die microscopic image.

observing the control I/Os, did not follow the expected patterns. Although the ADC signals could be read out, there were issues with observing spurious zero codes which were attributed to the FIFO design. Interestingly, the ASIC core operation only seemed to work reasonably well when the core voltage was raised to 1.3 V as opposed to the nominal 1.0 V. Later, it was found that the chip would heat up during testing; however, during the design of the 2nd revision test system, an error in the PCB landing pattern for the chip was discovered which may have resulted in shorting between certain pads. The ADC functionality was roughly confirmed using the ADC readout, and no other major conclusions were drawn from this revision.

In the second revision, the number of monitor ports was expanded from four outputs and three selection bits ($4 \times 2^3 = 32$) to sixteen outputs and four selection bits ($16 \times 2^4 = 256$), and marginal improvements were made to the chip's critical path timing under the hypothesis that narrow timing margins may have caused the control sequence errors detailed above. The monitor signals were selected to cover the majority of critical signals relating to the clocks, H3DD Core, and DSP Core.

The initial testing protocol was to send a data byte using one SPI channel, and to observe that the SPI word came out on the monitor port. While the SPI words could be successfully observed on the monitor ports, the first issue appeared when the registered SPI address did

not exactly reflect the incoming SPI word. This behavior appeared to be highly dependent on the clock frequency. For example, only frequencies as low as 12.5 MHz – not the nominal 100 MHz – would show reasonable functionality. Interestingly, it was found coincidentally that *lowering* the core voltage to 0.6 V apparently restored much of the errant behavior. In other words, at 0.6 V, the SPI address registering process appeared to work exactly as expected.



(a) Simulated signal (b) VDD = 1.0 V (c) VDD = 0.6 V

Figure 4.4: Simulated and experimental comparison of the `asic_gReset_cnt` signal.

Following the low bias anomaly, the testing was continued by attempting to verify the H3DD Core behavior using the respective monitor ports. It was observed on the monitor ports that one internal counter responsible for the H3DD-UM ASIC global reset, `asic_gReset_cnt`, behaved irregularly. Figure 4.4 shows the simulated counter behavior, the measured result when applying 1.0 V, and the result when applying 0.6 V. As shown in Figure 4.4b, the signal was incomprehensible at 1.0 V core voltage. Again, at 0.6 V, the signal appeared to work properly. However, upon closer investigation, even Figure 4.4c is not correct. Bits 2 and 3 both go high at the same time, indicating a counting sequence of $0, 1, 2, 3, 12$. The clock signal is not shown in Figure 4.4, but eventually it was found that `asic_gReset_cnt` toggled on the rising and falling edges of the internal clock. This was verified by running the chip with clocks of varying duty cycles and then observing how the incrementing of `asic_gReset_cnt` relatively matched the duty cycle. After this finding, it was concluded that the internal clock must have an issue. By further scrutinizing the synthesis and APR commands, it was found that one line of code errantly caused the clock tree synthesis step to be skipped. The violating line was written as follows: `ccopt_design -check_prerequisites`. Apparently, by including the `-check_prerequisites` flag, it was possible to entirely skip the clock tree synthesis phase, which had happened in the first two revisions. To resolve the issue, it was only required to run `ccopt_design -cts` instead.

Importantly, it was noted that the synthesized clock tree should always be visually observed using the Innovus clock tree tool as one check in the final tapeout checks. The UltraSim simulations detailed in Section 3.3 are another sign-off method for verifying that the clock works as expected on chip. Once the clock tree synthesis issue was discovered, no further testing was done as it was clear that all internal clocks would be too capacitively loaded to perform any interesting tests.

4.4 Revision 3

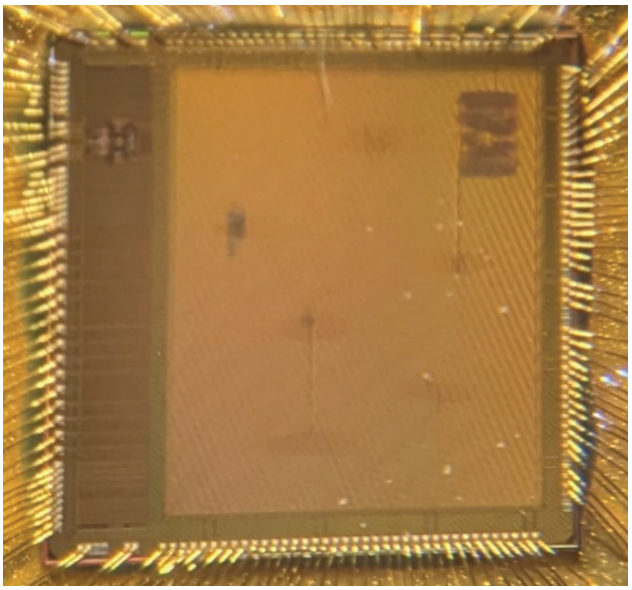


Figure 4.5: DAQ-DSPv3 sample die microscopic image.

The goal for the DAQ-DSPv3 was to achieve any functional readout to prove the merits of the project. Although it was known at the time that certain fundamental design issues existed, the bare minimum was done to ensure that a result could be delivered. First, the clock tree synthesis issue was resolved, requiring the change of only one line of code. Moreover, multiple design errors were revised for the DAQ-DSPv3 to reach the minimum deliverable goal. Even if the clock tree synthesis had been executed correctly, the existing design flaws would have prevented any reasonable functionality.

4.4.1 RTL Revision

The first of these issues related to the reset programming in the DSP Core. Two possible methods are used for applying resets to the chip: synchronous and asynchronous. The synchronous reset, as shown in Figure 4.6 makes use of registered logic which resets the register through the data stream, whereas asynchronous resets directly access a reset port of the register. The asynchronous reset is typically a long pulse relative to the clock frequency, which means that the reset port has timing requirements that prevent the reset pulse from being too short. Conversely, the synchronous reset may change with the clock cycle, which means that it can potentially be a fast changing signal. In the revision 1 and 2 designs, the DSP Core used a problematic mix of the two techniques. Synchronous-type resets – or fast switching signals – were sent directly to the asynchronous reset port of all registers due to the RTL coding style. Simulations revealed that it was not possible to determine the behavior of the chip due to this issue, and all resets were rewritten in the synchronous style to match the intended signal convention.

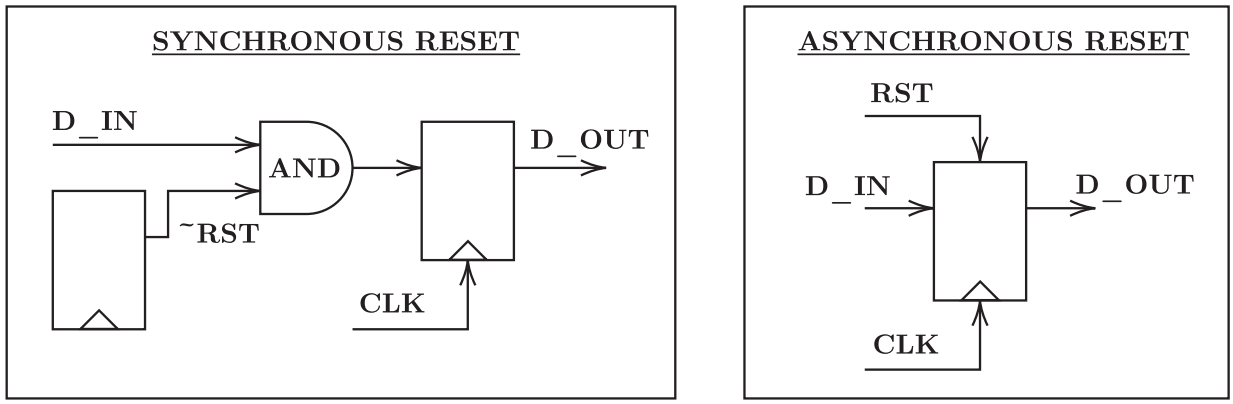


Figure 4.6: Synchronous and asynchronous reset techniques.

Two more categories of errors were associated with the DSP Core: baseline programming errors and sampling alignment errors. The intended baseline functionality is to subtract a baseline value which is averaged over many cycles. The initial RTL, though, had a timing error such that the average baseline value would never be updated, making the true baseline never subtracted. Sample alignment errors were also found in the RTL which caused 3 samples from the waveform to be lost. The baseline error was easily resolved by modifying the clock edge associated with the baseline update logic. The sample alignment error, though, was deemed too difficult and not significant enough to resolve without rewriting a

considerable portion of the DSP Core, and thus was left in place for the 3rd revision.

Three errors were corrected relating to the SPI communications and output interface. In revisions 1 through 3, each output packet consisted of 6 32-bit words. When writing the output packet into memory, the write address was only incremented by 2 each time, resulting in all but the first two words of the output packet being overwritten each time. The output packet was also incorrectly gated depending on whether the ASIC was currently reading out data from H3DD-UM front-end chip. If output data was ready while a waveform was being read out, the output packet would be completely lost. Since the waveform readout and data output phases are independent, there was no reason to include this gating process. Finally, the output packet included a timestamp to help identify the time of interaction. The timestamp was not locked at the correct time, though, and so it only corresponded to the time that the output packet was released, which was misleading. All three described errors were resolved for the 3rd revision.

Although major revisions were made to the chip, two prominent issues remained on the DAQ-DSPv3 ASIC as they were considered difficult or risky fixes, but would not prevent a proof of concept. Those two problems were the clock-domain-crossing at the FIFO, and the systematic rounding and cycle-timing issues of the DSP Core. The FIFO serves as the boundary between on-chip data and off-chip data. In order to accommodate differences in data production and data consumption speeds – which can in part be attributed to clock speed differences – the FIFO allows for temporary storage of data. Since the two sides of the FIFO use different clocks which have no relationship (i.e. asynchronous clocks), the timing must be handled carefully. In revision 3, the asynchronous clock-domain crossing was not handled, and therefore the chip would suffer from severe data loss and paralyzability. In addition, the DSP Core accuracy was degraded due to several instances of incorrect rounding, and off-by-one sample losses. The result is that the DSP incorporated "digital-noise" and sample losses into the resulting output. These issues were systematic and widespread, necessitating a complete overhaul of the DSP Core in the 4th revision.

4.4.2 Design Flow Upgrades

The design flow discussed in Chapter 3 is the most up-to-date version, which was used for the 3rd and 4th DAQ-DSP ASIC revisions. Prior to the third revision the design flow contained critical differences. One error, as already discussed, was that the clock-tree synthesis was not run during the APR flow. Another drawback was that the top, thicker metal layers – M8 and M9 – were not used on the digital side of the design in the first two revisions. This

is an important section of real estate which offers more optimal power routing, and also frees up additional layer space for signal routing. This difference is visually clear in Figures 4.3 and 4.5 as the top metal layers are used to create a power mesh across the digital block.

One important flag related to the filler cell instantiation, `-ecoMode true`, missed in the previous revisions, was now enabled. The flag allows filler cells to be removed and re-added during timing optimization. As explained in Section 3.1, filler cell insertion typically happens after clock tree synthesis, and before routing is done. This may be sufficient in some cases, but in a large scale design like the DAQ-DSP ASIC, the routing phase may encounter DRC errors or timing constraints which can only be resolved by relocating filler cells. If the filler cells are completely fixed in place, the optimal routing and placement will not be achieved. The inclusion of proper flags for the filler cells in the APR phase is a significant improvement for the 3rd and 4th revision.

The final top-level synthesis and APR flow was also revised to use a complete top-down approach for the 3rd and 4th revisions. In the first revision, the major blocks of the ASIC were routed and placed independently [8][21], and then placed into the top-level design by including the LEF of each block. Starting with the second revision, a hybrid approach was used in which the major blocks were synthesized independently, and then the synthesized netlists were included as non-modifiable blocks during the top-level synthesis run. While this saves computational effort, it prevents the synthesis tool from optimizing those included blocks with respect to the surrounding top-level logic. If possible, the best strategy is to run synthesis and APR completely using the top-level netlist without including any fixed blocks. The only limit is whether the computing resources available can manage the entire netlist in a reasonable time frame and without crashing. The machine used for revision 3 and 4 was indeed capable of this method. The major advantage of using a complete top-down approach is that the design flow tools always have full control over the netlist, which allows any signal path to be modified as appropriate for the optimal timing and area results.

4.4.3 Layout and Placement Results

The bare and annotated tapeout layouts for the DAQ-DSPv3 are shown in Figure 4.7. The total chip area is 4.2×4.2 mm^2 $= 17.64$ mm^2. The critical path slack reported was 1.066 ns on a 100 MHz clocked path for the 3rd revision, which is considered to a healthy margin to ensure correct chip functionality.

As shown in Figure 4.7, a significant portion of the layout on the digital side was left empty in order to ensure that sufficient area was available during the first three revisions.

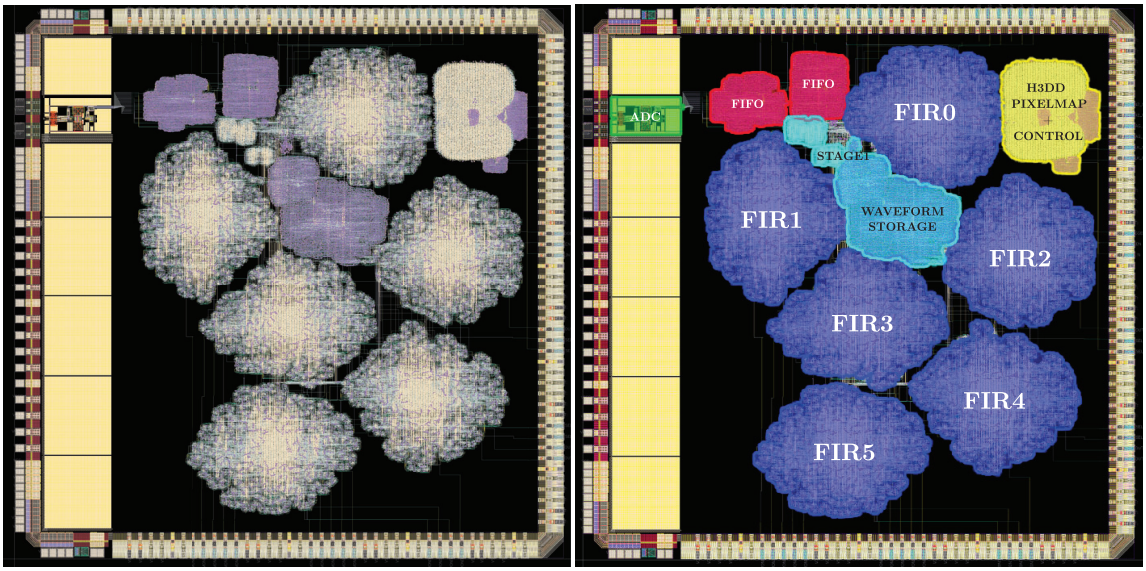


Figure 4.7: The tapeout ready DAQ-DSPv3 design. Left: Final DAQ-DSPv3 ASIC design. Right: Annotated layout.

The annotated layout also provides insight into which blocks consume the most area and power. The six finite impulse response (FIR) filters, shown in blue, occupy a majority of the space on chip and are expected to be the most power consuming components as well. In addition to the FIR filters, the three other main components are the FIFOs, the H3DD-UMv4 pixel mapping, and the DSP Core waveform storage. Each of the design blocks, and its trade-offs, are discussed in contrast to the Revision 4 design in Chapter 5. The rectangular layout on the left-hand side of the chip contains the analog section, with the ADC being highlighted in green.

4.4.4 Test System

The DAQ-DSPv3 test system consists of a single motherboard printed circuit board (PCB), and a commercial Zynq 7000 SoC based FPGA board. The motherboard PCB was designed to have debugging capabilities while still demonstrating that the overall system could be relatively compact. Figure 4.8 shows the block diagram of the overall test system, as well as the top and bottom layers of the PCB designed to test the revision 3 ASIC. The block diagram shown in Figure 4.8 only indicates the functional blocks of the test system, and not supporting circuitry like power regulation or debug pins. The power distribution scheme is carefully designed to optimize the performance. As shown in the PCB layout, three SMA

connectors bring in power for the digital and analog sections of the board. The top-left SMA connector brings in 3.3 V which is regulated down to two separate 1.2 V used for the ADC and the H3DD-UM front-end ASIC. Another regulator is used to generate the 0.6 V common mode reference for the ADC. On the digital side, 3.3 V is supplied for the I/O cells, and 1.2 V is supplied for the H3DD-UM digital circuitry. The 1.2 V supply is also regulated down to 1.0 V for the DAQ-DSP ASIC core operation.

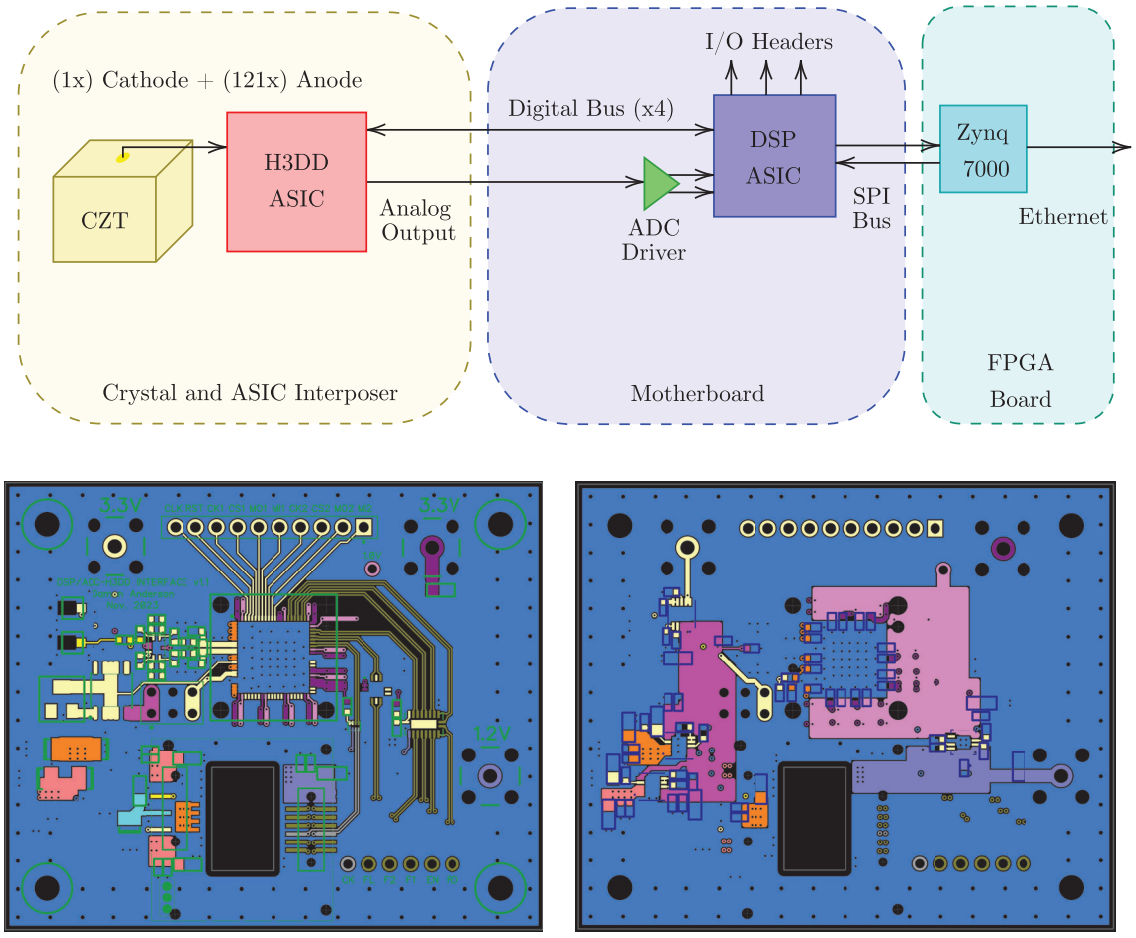


Figure 4.8: Top: DAQ-DSPv3 Test system block diagram. Left: Motherboard PCB top layer. Right: Motherboard PCB bottom layer.

The test system was used in two configurations within a test box as shown in Figure 4.9. The motherboard and FPGA boards were mounted within a test box, and connections were made to allow the DAQ-DSPv3 ASIC to be controlled by the FPGA SPI channel. Power was brought into the enclosure through three SMA cables. The DAQ-DSP ASIC basic functionality was confirmed by inserting a bare H3DD-UMv4 ASIC. Subsequently, a

CZT crystal was attached to the H3DD-UM front-end ASIC, and the module was brought to -3000 V bias using the high voltage board shown on the right-hand side of Figure 4.9.

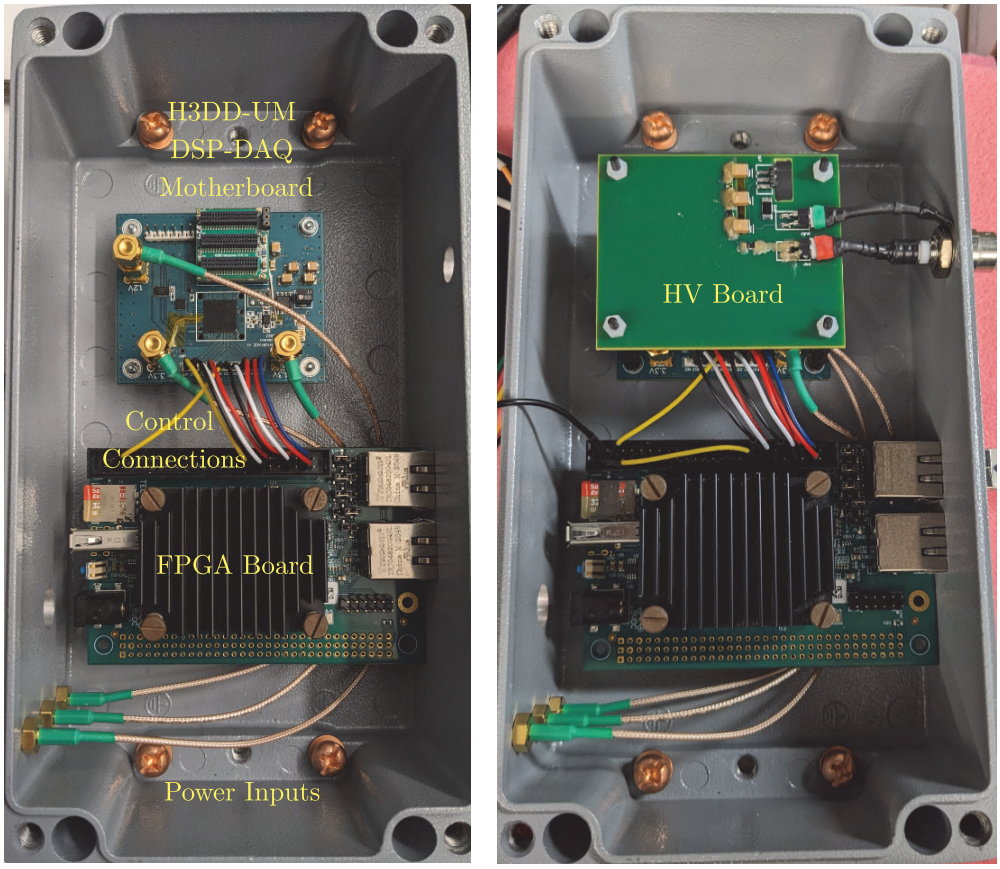


Figure 4.9: Revision 3 test box. Left: Motherboard with mounted H3DD-UM ASIC. Right: High voltage board assembled.

4.4.5 Measurement Results

The 3rd revision of the DAQ-DSP ASIC taped out on August 3rd, 2023, and testing began on October 30th. Using the fabricated test-system, three major results are reported from this iteration of the DAQ-DSP ASIC. First, a test-pulse scan was collected without any crystal attached, proving that the H3DD Core control and basic DSP Core filtering functions work as expected. The result of the test pulse scan is shown in Figure 4.10. The scan result shows that the H3DD-UM ASIC can be automatically controlled as expected, and the DSP Core is properly filtering the test pulse and posting correct results.

Next, the crystal was biased to -3000 V and a rough spectrum was collected to further

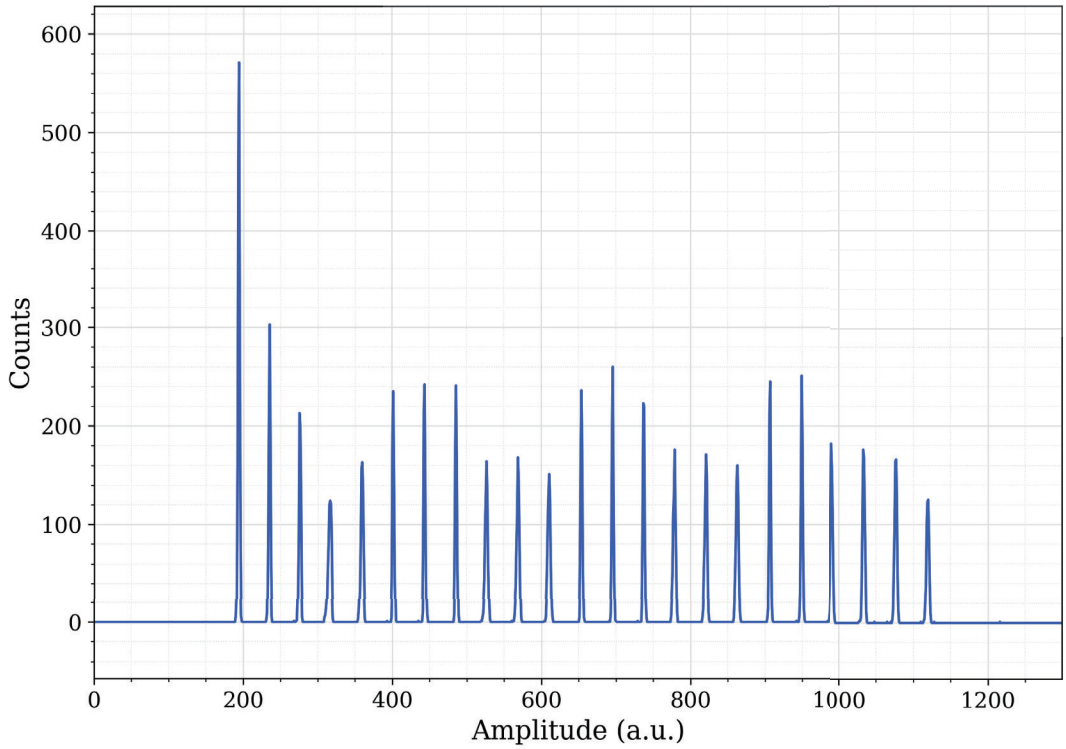


Figure 4.10: Test pulse spectrum measured from the DAQ-DSPv3.

illustrate that the output signals were of the expected format. While Figure 4.11 bares resemblance to a raw Cs-137 spectrum, it is not high quality. Low steady-state data throughput was one major issue which plagued the revision 3 ASIC. During measurement, it was found that as the source was brought closer to the test system, the readout would become completely paralyzed. This was partially expected as noted in Section 4.4.1 due to the FIFO design. At high rates, the FIFO would be flooded with data, causing severe timing issues as the FIFO is constantly being filled and read out from both sides. In a typical FIFO design, this is not an issue, but because the asynchronous clock-domain crossing was not handled for the third revision, this type of behavior is hard to avoid. The resulting maximum count rate was only around ≈ 100 cps, and it was deemed too impractical to try to collect a full spectrum for calibration. Even the limited spectrum shown in Figure 4.11, which is only from single-pixel interactions, is broad and uncharacteristic. The relatively high threshold and the DSP Core issues noted in Section 4.4.1 are expected to contribute to the spectral

degradation. It was not deemed necessary to fully debug and understand the spectrum shape observed in Figure 4.11 as the DSP Core was already intended to be revised for the 4th revision.

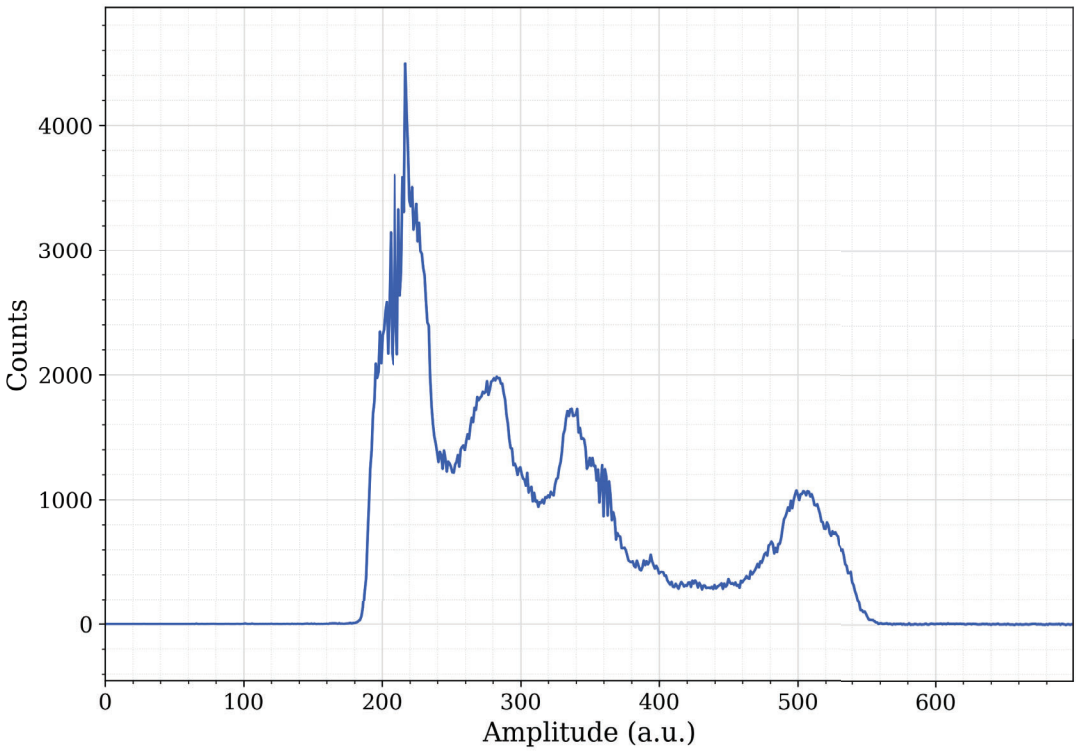


Figure 4.11: Cs-137 spectrum measured from the DAQ-DSPv3.

Finally, the operational power of the system could be measured now with a working version of the chip. The power consumption was measured using a standard power supply (RIGOL DP832A). The contributions from the I/O supplies and core voltage supplies are summarized in Table 4.1.

4.5 Conclusions

Three iterations of the DAQ-DSP ASIC were fabricated with the intent of replacing the bulkier ADC-FPGA-desktop style system with a single chip. The DAQ-DSPv3 can ultimately be considered a successful step in the direction of a functional processing ASIC,

Table 4.1: DAQ-DSPv3 Power Consumption Summary

	Power Consumption (mW)
Core (1.0V)	58
I/O (3.3V)	30
Total	**88**

while still having shortcomings. After failing to record any meaningful results from the 1st and 2nd revisions, alterations to the ASIC design flow and modifications to the RTL were included in the 3rd revision. The fabricated ASIC was placed in an enclosed test system consisting of a motherboard for the processing and front-end ASICs, and an FPGA board used to control the system. The intended capabilities of reading out events and controlling the H3DD-UM ASIC were confirmed. Following the successes of the 3rd revision, the 4th DAQ-DSP ASIC was planned to bring the project close to completing the goal of a fully functional ASIC.

CHAPTER 5

DAQ-DSP ASICv4–Design

5.1 General Description

The design of the DAQ-DSPv4 ASIC was motivated by the remaining errors in the 3rd revision, as well as the desire for lower power and area. The updated high-level diagram is shown in Figure 5.1. The notable differences are summarized here, and the design details and further motivation are discussed in the following sections.

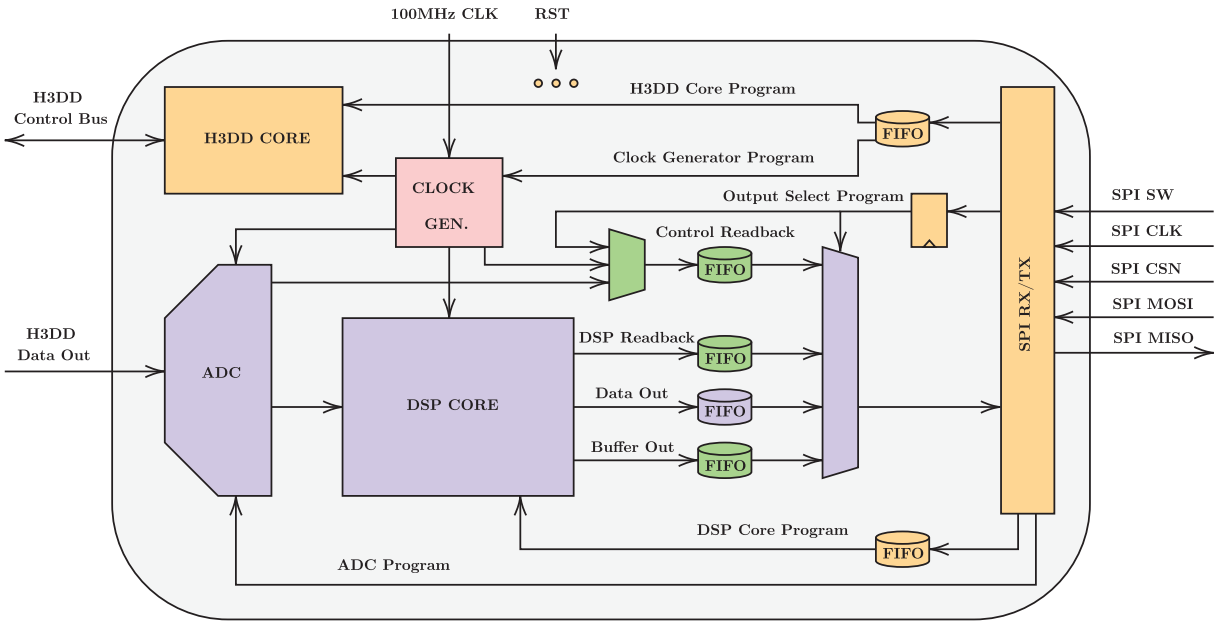


Figure 5.1: DAQ-DSPv4 ASIC top-level functional diagram.

First, the debug capabilities are shown in green. Any block with programmable registers was modified to include a read-back path so that any register can be independently verified

after programming. The read-back function is described in more detail in Section 5.3.3. In addition, a buffer path is included as one output of the DSP Core. The buffer is a method for enabling the user to see full waveforms from all outputs of the DSP Core. This was deemed a critical method for ensuring that the filters are functioning as expected. The buffer is elaborated on in Section 5.2.5.

In the 4th revision, more FIFOs – with an updated design – are used throughout. Since one of the critical FIFO functions is to provide a reliable transition between clock domains, one FIFO is included for each clock-domain crossing. On the receiver side, two separate FIFOs are used: one for the DSP Core which relies on the specific DSP clock, and one for all other control registers which operate on the master clock. On the transmit side, four separate FIFOs are used for varying purposes. Two readback FIFOs serve the same purpose, but again for separate clock domains. All master clock related read-back sequences go through the control readback FIFO, while all DSP related read-back sequences go through the DSP readback FIFO. Then, the data-out FIFO is used for the standard output path from the DSP Core. Finally, the buffer has its own high-depth FIFO since it requires enough depth for an entire waveform. The output path is configured using an output selection multiplexer (MUX) between the four options.

The data flow, shown in dark blue, is updated with respect to the third revision. A 13-bit SAR-assisted pipeline ADC was designed by Seungheun Song, and the architectural details are reflected in [15] and [27]. The ADC no longer has a direct readout path to the output FIFO; rather, the ADC readout is handled using the debug readout buffer in this case. Again, one important update to the data path is the inclusion of an independent FIFO.

At the SPI communication interface, the two different SPI channels of the 3rd revision were merged into a single one to reduce the SPI pin count. While the internal conventions still differ, the two channels are now managed by a digital switch – `SPI_SW` – to determine which destination the SPI information goes to internally.

5.2 DSP Core

The DSP Core receives 13-bit waveform samples from the ADC and processes the result into an output packet containing signal amplitude, timing, and trigger information. In addition, meta-data including channel number, module number, and timestamps are passed along at each stage of the DSP Core to be incorporated in the output packet.

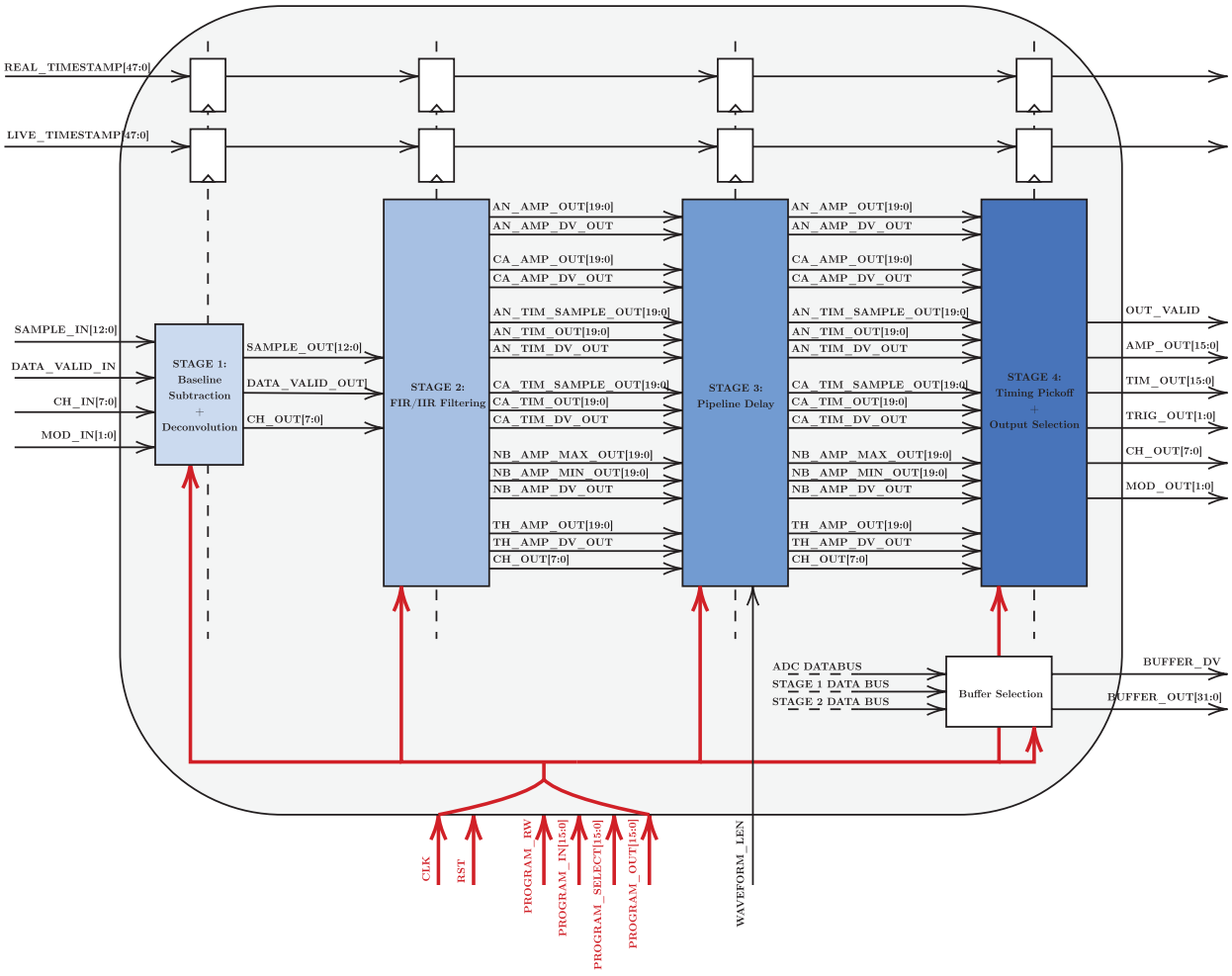


Figure 5.2: DSP Core functional diagram.

The DSP Core consists of four primary stages as shown in Figure 5.2. First, the waveform baseline is removed and the exponential decay from the H3DD-UM ASIC preamplifier is removed by deconvolution. Then, the primary filters – trapezoidal and CRRC^4 – are applied to the incoming waveform samples to prepare for amplitude and timing pick-off. Amplitude pick-off is performed during the filtering operation by tracking the maxima and minima of the filter outputs, where applicable. After the filtering stage, the output samples of timing related filters are delayed by a variable amount determined by the waveform length. The relevant filter maxima and minima are locked and passed along during the third stage. In the final stage, the constant fraction timing pick-off is performed by searching through

the waveform for a programmable percentage threshold crossing. The final stage is also responsible for providing the data in a usable packet to the respective FIFO. In addition, output samples from eight possible sources may be sent to an output buffer, for debugging purposes. The eight sources are: ADC output, stage 1 output, and the six filter outputs from stage 2. The accompanying data valid and channel signals are passed to the buffer selector. When the corresponding data is valid, and the corresponding channel matches a programmed option, then the sample is written to the output buffer FIFO. The entirety of the DSP Core relies on one clock which is nominally operated at 25 MHz.

5.2.1 Stage 1: Baseline Subtraction and Decay Deconvolution

Stage 1 of the DSP Core is responsible for calculating and removing the waveform baseline, and deconvolving the exponential preamplifier decay.

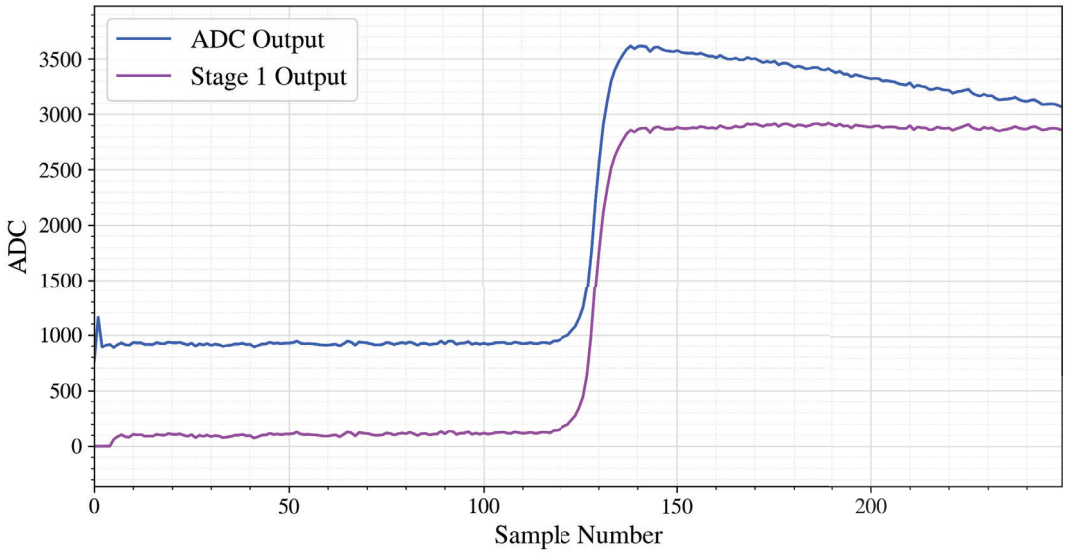


Figure 5.3: Example of baseline subtraction and decay deconvolution.

The baseline subtraction scheme relies on a rolling average for each channel. The current 13-bit baseline for every channel of each module is stored in memory, and initialized to 0. During the signal valid phase, the current stored baseline is subtracted from each incoming 13-bit sample. In parallel, a new baseline is calculated from the incoming samples based on the programmed baseline start and length. The baseline start is an 8-bit field which specifies the sample count after which the baseline calculation should begin. The baseline length

is a 4-bit, one-hot encoding in which the value is multiplied by 8 to determine the actual length. As such, baseline lengths of 8, 16, 32, and 64 are possible. If the baseline length field is set to 0, then a default length of 8 is used. Finally, the updated baseline is calculated using the following rolling average formula where B_{new}, B_{curr}, and B_{calc} are the new, current, and calculated baselines, respectively. After the baseline calculation is complete, the new baseline is stored in memory and used for subsequent waveforms.

$$B_{new} = \frac{255 \times B_{curr} + B_{calc}}{256}$$

The exponential decay deconvolution is implemented using an infinite impulse response (IIR) block which will be described in Section 5.2.2. In order to account for different input polarities, a programmable polarity switch is included, denoted S in Equation 5.1. In addition, the exponential decay constant, denoted d in Equation 5.1, can be programmed to match each channel, as well as different sampling frequencies. The resulting output of stage 1, including baseline subtraction, is given by Equation 5.1. The scaling factor of $\frac{1}{1024}$ is chosen as a power of 2 so the division can be implemented as a simple shift. Then, d is in the range $[0, 1024]$.

$$\begin{aligned} x_B[i] &\triangleq x[i] - B_{curr} \\ y[i] &= (-1)^S \times (y[i-1] - x_B[i] + \frac{d}{1024} \times x_B[i-1]) \end{aligned} \tag{5.1}$$

The IIR block operates in the floating point (FP) domain. However, the FIR filter and IIR trapezoidal filter used in stage 2, also elaborated on in Section 5.2.2, both operate on signed 13-bit integers. As a result, two outputs are provided by stage 1. The 32-bit FP results are maintained and used for the IIR CRRC filters, and a rounded 13-bit result is passed to the FIR and IIR trapezoidal filters. Operating on the FP domain in stage 1 is necessary to preserve resolution while doing the exponential decay deconvolution.

5.2.2 Stage 2: Filtering and Amplitude Pick-off

The second stage of the DSP Core implements the 8 filters, summarized in Table 5.1. The FIR filters are fully programmable and can be used for any purpose. The IIR trapezoidal filters are used primarily for preparing the waveform for amplitude pick-off. The IIR CRRC

Table 5.1: Stage 2 Filter Summary

Filter Type	Quantity	Purpose
FIR	2	Any
IIR-Trapezoidal	3	Anode/cathode amplitude
IIR-CRRC^n	3	Anode/cathode timing; Neighbor amplitude

filters are used primarily for preparing the waveform for timing pick-off. In addition, one of the CRRC filters is used for the amplitude determination of neighboring signals. Six of the eight available filters are programmed to drive the following required outputs, as shown in Figure 5.4.

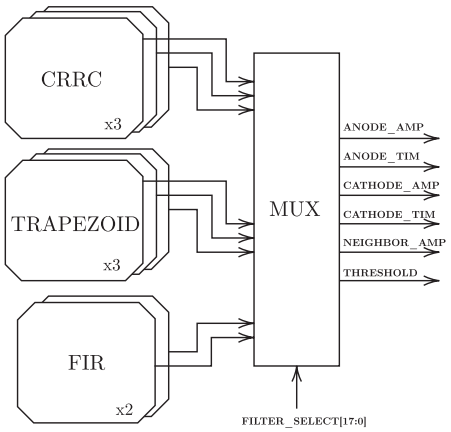


Figure 5.4: DSP Core stage 2 functional block diagram.

- Anode Amplitude
- Cathode Amplitude
- Neighbor Amplitude
- Anode Timing
- Cathode Timing

- Threshold Crossing

To calculate the anode and cathode amplitudes, the maximum is tracked during the filtering process. Similarly, the neighbor amplitude is calculated by tracking the maximum *and* the minimum. The difference between the two yields the amplitude. The output of the anode and cathode timing filters are passed to stages 3 and 4 where constant fraction pick-off is used to determine the signal timing. The threshold filter is used to determine if the signal crosses a programmable threshold. The threshold determination is unique in that the *final* sample, rather than the maximum sample, is used to determine if the threshold filter output is greater than the programmed threshold. In other words, the threshold crossing is determined by the multiplication and sum-reduction of the filter impulse response and the input signal, rather than the convolution and maximum.

5.2.2.1 FIR Filter

The FIR filter is implemented as two shift registers of length 256. One shift register is used to contain 16-bit programmable parameters, and the other shift register receives 13-bit inputs from the DSP Core stage 1. During each clock cycle, one input is shifted in, and the multiplication and sum-reduction of all elements is calculated.

5.2.2.2 IIR Trapezoidal Filter

The IIR trapezoidal filter is implemented using the recursive formula in Equation 5.2.

$$y[i] = M \times (y[i-1] + (x[i] - x[i-k_0]) - (x[i-k_1] - x[i-(k_0+k_1)])) \tag{5.2}$$

Here, M is a programmable scaling factor to make best use of the dynamic range, and $x[i]$ and $y[i]$ are the inputs and outputs, respectively. The constants k_0 and k_1 set the integrating and gap periods of the trapezoidal filter, as discussed in Section 1.4. The required elements include a 256 length delay register containing 13-bit unsigned integers, four adders, and one 20-bit register for the accumulated output as shown in Figure 5.5.

The internal filter output width is arbitrarily maintained at 20 bits. At the final output of the DSP Core, the amplitude and timing outputs are rounded to keep the 16 most significant bits in order to form more convenient integer words. Between filtering and amplitude determination, no additional operations are done. For timing determination, the fractional pick-off procedure may benefit from the additional four internal bits. As a result, it may be beneficial in future revisions to reduce the internal storage length to 16-bit for amplitude

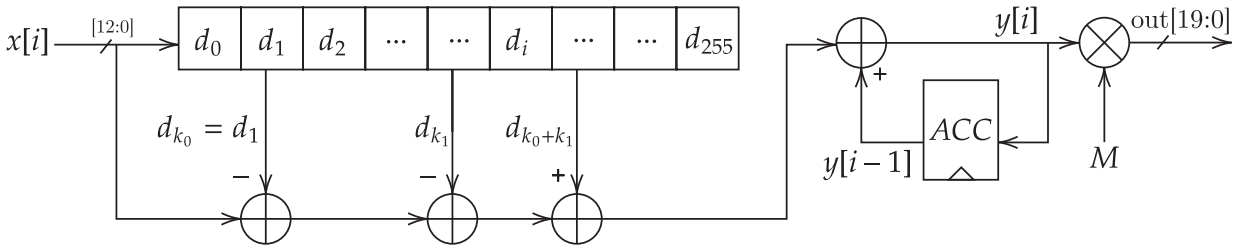


Figure 5.5: IIR trapezoidal filter block diagram.

related filters, and maintain a bit-length of 20 for timing related filters. Figure 5.6 shows an example of the trapezoidal filter step response.

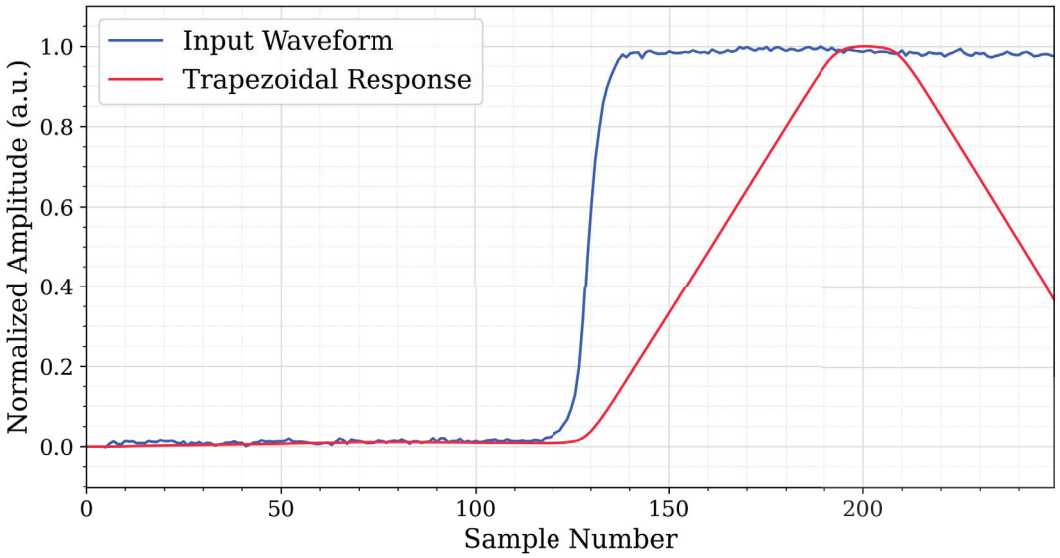


Figure 5.6: Example of IIR trapezoidal filter waveform.

5.2.2.3 IIR CRRC Filter

The IIR CRRC filter is composed of several IIR blocks that implement the general recursive formula shown in Equation 5.3:

$$y[i] = p_0 \times x[i] + p_1 \times x[i-1] + p_2 \times y[i-1] \tag{5.3}$$

Here, p_0, p_1 and p_2 are 32-bit programmable floating point (FP) parameters. Since the

IIR CRRC filter relies on specific fractional parameters for the recursive formula, FP must be used to maintain precision. The FP convention does not match the standard FP-32 IEEE 754 protocol [6]. Options like NaN, infinity, and denormalization are not included, for simplicity. Since the maximum output value is only allowed to be $2^{19} - 1$ – the maximum value for a 20-bit signed integer – only 6 exponent bits are required. This allows a range of exponents from 2^5 to 2^{-5}. Then, the remaining 25 bits are used for the mantissa. The maximum relative rounding error associated with floating point arithmetic is the well-known result defined in Equation 5.4, where $fl(x)$ is the floating point representation of the real number x, β is the radix, and p is the precision. With base-2, and a precision of 26 – one greater than the number of mantissa bits – the maximum incurred relative error is 1.49×10^{-8}. If the maximum value allowed at the filter output is $2^{19} - 1$, then the maximum absolute error is $(2^{19} - 1) \times 1.49 \times 10^{-8} \approx 0.008$. Since this is well below the closest-integer rounding of 0.5 which will occur at the final output, the error margin is considered acceptable.

$$\epsilon_{max} \triangleq \max_{x} \frac{|x - fl(x)|}{|x|}$$
$$\epsilon_{max} = \frac{1}{2}\beta^{1-p} \tag{5.4}$$

The IIR formula shown in Equation 5.3 can be shown to exactly represent one CR or RC block using the Bilinear Transform [20]. The analog system response of an RC circuit, for example, is the well known formula 5.5 where $\tau = RC$ and s is the complex variable jw with j as the imaginary number.

$$H_a(s) = \frac{1}{1 + s\tau} \tag{5.5}$$

The Bilinear Transform states that a continuous response on the s-plane can be transformed to the discrete z-plane using the relation stated in equations 5.6 and 5.7, where T is the sampling period. Equation 5.6 shows the transform solved for z, and Equation 5.7 shows the transform solved for s which requires an approximation [20].

$$z = e^{sT} \tag{5.6}$$

$$s = \frac{2}{T}\frac{z - 1}{z + 1} \tag{5.7}$$

In order to avoid a phenomenon known as frequency warping, a pre-scaling (pre-warping) factor can be included which modifies Equation 5.7 to Equation 5.8, where ω_0 is a constant frequency of interest, such as the cut-off frequency of the RC filter.

$$s = \frac{\omega_0}{tan(\frac{\omega_0 T}{2})}\frac{z-1}{z+1} \tag{5.8}$$

Plugging in Equation 5.8 to 5.5 yields the result shown in Equation 5.9.

$$\begin{gathered} A = \frac{1}{tan(\frac{\omega_0 T}{2})} \\ H_d(z) = \frac{(\frac{1}{1+A})(1+z^{-1})}{1+\frac{1-A}{1+A}} \end{gathered} \tag{5.9}$$

An equivalent form of the transfer function is used to understand how Equation 5.9 is transformed into the difference equation. Equations 5.10 and 5.11 are the canonic representations of the discrete transfer function and corresponding difference equation, and the transformation between the two is the discrete Fourier transform. The transfer function is composed of the sum of any power of z in the numerator and denominator weighted by respective b_i and a_j coefficients. In the time domain, those coefficients reflect weights to the past inputs and outputs, respectively.

$$H_d(z) = \frac{\Sigma_{i=0}^{P} b_i z^{-i}}{\Sigma_{j=0}^{Q} a_j z^{-j}} \tag{5.10}$$

$$y[n] = \frac{1}{a_0}(\Sigma_{i=0}^{P} b_i x[n-i] - \Sigma_{j=1}^{Q} a_j y[n-j]) \tag{5.11}$$

Using this transformation, Equation 5.9 is rewritten as the final result in Equation 5.12. The general recursive relation shown in 5.3 is sufficient to implement that shown in Equation 5.12. The CR transfer function, similarly, can be shown to yield Equation 5.13.

$$y[i] = (\frac{1}{1+A}) * x[i] + (\frac{1}{1+A}) * x[i-1] + (\frac{1-A}{1+A}) * y[i-1] \tag{5.12}$$

$$y[i] = (\frac{A}{1+A}) * x[i] + (\frac{-A}{1+A}) * x[i-1] + (\frac{1-A}{1+A}) * y[i-1] \tag{5.13}$$

The top-level IIR CRRC filter contains five such IIR blocks: one CR, and four RC. Each block contains the 32-bit FP parameters p_0, p_1, and p_2. Since each parameter is considered to be 16-bit, the CR and RC blocks are considered to have 30 total parameters, where each 32-bit FP parameter is composed of two 16-bit parameters. Two additional programmable parameters are included in the IIR CRRC top-level. First, a parameter corresponding to the filter order selects which RC output is actually used. For example, if the order is set to 4, then the filter is programmed to be a CRRC^4. The second parameter is a gain knob, much like the trapezoidal filter parameter detailed in Section 5.2.2.2. This allows the CRRC output to be scaled to use the full dynamic range. The comprehensive block diagram is shown in Figure 5.7.

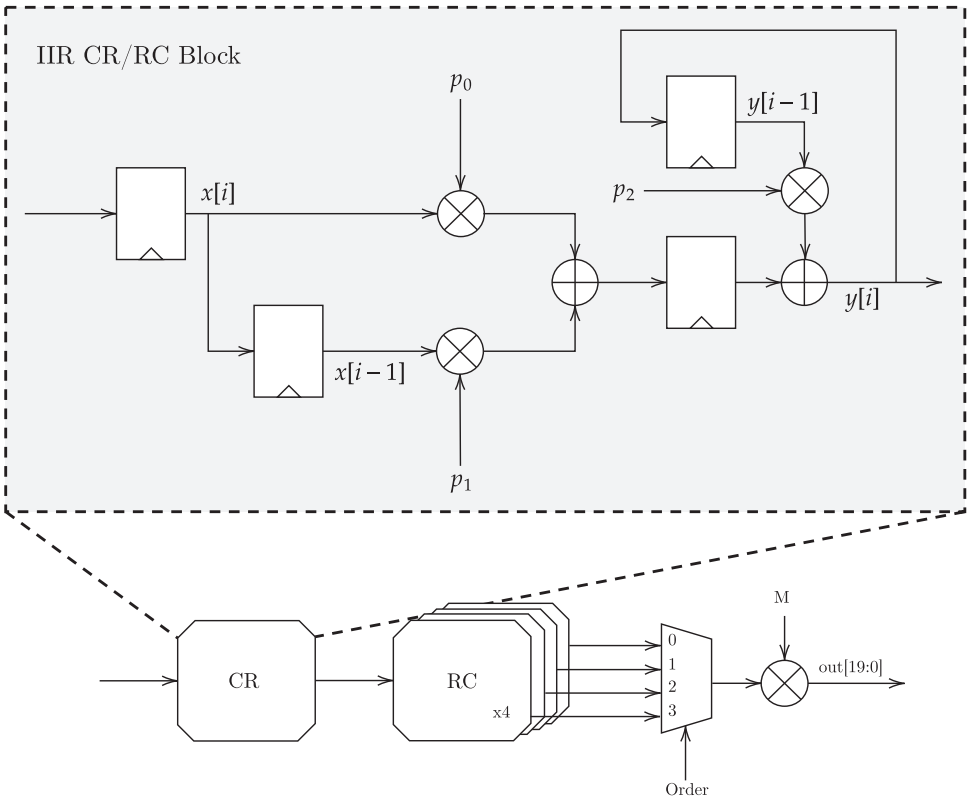


Figure 5.7: IIR CRRC filter functional block diagram.

Figure 5.8 shows the response of a standard anode waveform to several different CRRC shaping orders. It is important to note that the CRRC shaping time is adjusted for each

order such that the peaking time of each CRRC filter is held constant.

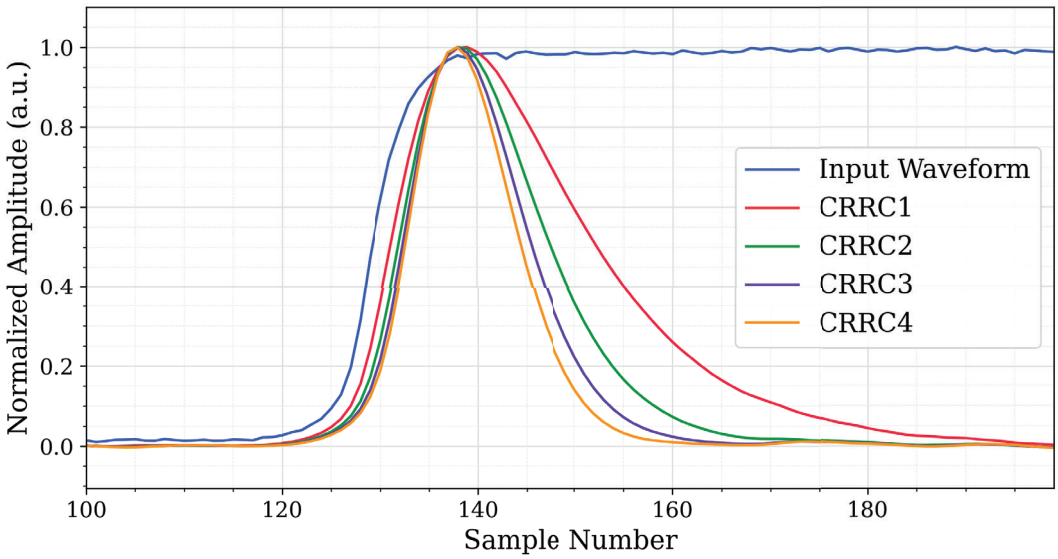


Figure 5.8: Example of IIR CRRC filter responses.

5.2.3 Stage 3: Pipeline Delay

Stage 3 of the DSP Core is a simple pipeline delay, which is included to allow the constant fraction pick-off stage to receive the signal amplitude before starting the threshold search. Two shift registers are included: one for the cathode timing samples, and one for the anode timing samples. The shift registers are a maximum of 256 length, but the actual delay length corresponds to the waveform readout length, which is a variable parameter passed from the H3DD Core. Single integer parameters – including maxima, minima, and meta-data – are locked and held until the pipeline delay is complete.

5.2.4 Stage 4: Constant Fraction Timing Pick-off

The final stage of the DSP Core performs two main functions:

1. Perform timing pick-off on the timing filtered waveforms
2. Package the received amplitudes, timing results, and meta-data into an output packet

To perform the fractional timing pick-off, anode and cathode filtered waveforms are received sample-by-sample. The key difference between stage 4 and stage 2 is that the maximum amplitude information is now available so that the fractional threshold is known as samples arrive. The fractional threshold, f, is a programmable value from 1 to 256, and the resulting fractional threshold is set to be $\frac{f}{256}$. If f is set to 256, the timing pick-off will correspond to the sample at which the waveform maximum is reached. An additional parameter is used to set the sample at which the threshold crossing starts. In order to improve the precision, a linear interpolation is performed between the two samples nearest to the threshold crossing. The linear interpolation is performed to the nearest $\frac{1}{16}$ between the two samples. Figure 5.9 shows the fractional pick-off method, where the sample number associated with the 50% crossing is returned as the waveform timing.

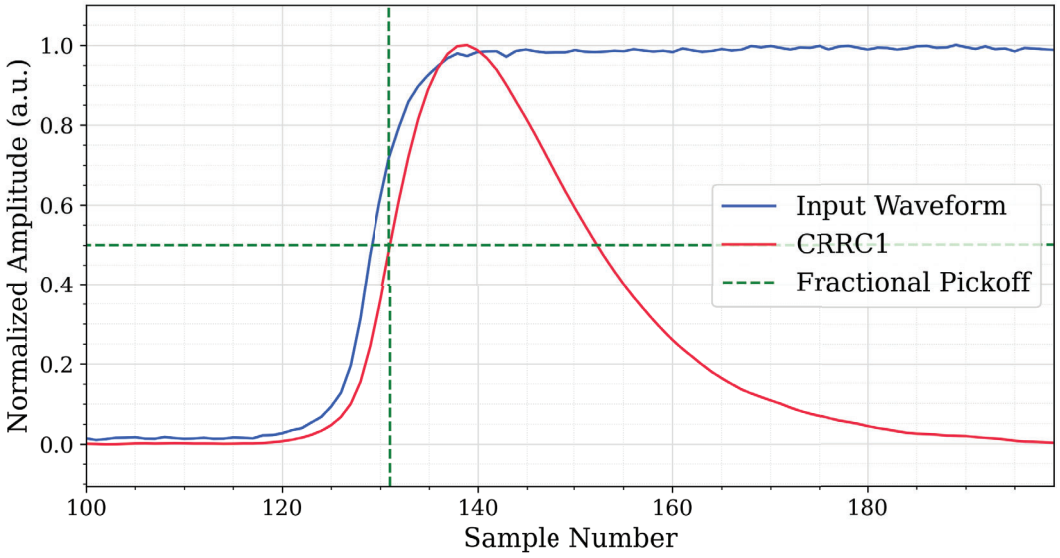


Figure 5.9: Example of the fractional pick-off technique used in stage 4.

Once the timing results are available from the anode and cathode pick-offs, the data is packaged into an output packet. The output packet consists of four 32-bit words, as shown in Figure 5.10:

The upper two bits of each output word are used as an index to ensure that the contents can always be interpreted when received. Each field shown in Figure 5.10 is defined below:

- **Module:** The index of the crystal from which data is being read out.

0	2'b00	TIMING[3:0]	AMPLITUDE[15:0]	PIXEL[7:0]	MODULE[1:0]
1	2'b01	REAL_TIMESTAMP[15:0]		TRIG[1:0]	TIMING[15:4]
2	2'b10	LIVE_TIMESTAMP[9:0]		REAL_TIMESTAMP[35:16]	
3	2'b11	LOSS_COUNT[3:0]	LIVE_TIMESTAMP[35:10]		

Figure 5.10: DAQ-DSPv4 output packet format.

- **Pixel:** The channel within the crystal that is being read out.
- **Amplitude:** The amplitude of the waveform.
- **Timing:** The timing pick-off of the waveform.
- **Trig:** The type of trigger being read out. The options are cathode, triggered anode, or neighbor waveform.
- **Real Timestamp:** The real operation time tracked by increments of the master clock.
- **Live Timestamp:** The live operation time tracked by increments of the master clock will readout is **not** occurring.
- **Loss Count:** A dynamic counter for the number of events lost will waiting to send the output packet. This field is included to help debug the system operation.

5.2.5 Buffer Output

One buffer output is included as a means of debugging the DSP Core and setting the proper filter coefficients. When the buffer input data stream is valid, each sample that arrives will be sent to a dedicated FIFO that can then be read out. Samples will arrive at the readout frequency which may be as high as 25 MHz, whereas readout from the FIFO requires one SPI transaction. As a result, the samples will arrive much faster than they can be read out. The buffer has a depth of 256 samples so that one entire waveform can be stored and then readout subsequently. This is the intended use for the buffer: receive one event, and then fully readout the samples before another event arrives. This is best controlled using a pulser for debug.

Eight different internal data streams may be sent to the buffer input:

- ADC Output (13-bit)

- Stage 1 Output (13-bit)
- 6x Stage 2 Outputs (20-bit)

The expected buffer readout sequence is outlined as follows:

1. Send an instruction to set the output stream switch to the DSP buffer FIFO.
2. Send an instruction to configure the buffer selection to the desired output and desired channel.
3. Fully halt the readout of the H3DD-UM ASICs by disabling the trigger flag for all included front-end ASICs.
4. Send 256 SPI instructions to flush the current contents of the buffer FIFO.
5. Enable the readout of the desired H3DD-UM ASIC by enabling the trigger flag.
6. Send 256 SPI instructions to read out buffer FIFO which will be filled by the first waveform encountered after the prior step.

The resulting collected data will represent the waveform at the given stage selected. The protocol may be repeated rapidly with a test-pulse for each buffer output option to verify that the filters are operating as expected. The buffer also requires that the input channel matches a programmable channel setting. Otherwise, the first received channel, rather than the channel of interest, would always be sent to the buffer.

5.3 Control Blocks

The supporting control blocks perform 3 primary functions: control the H3DD-UM front-end ASIC, generate all internal clocks, and communicate between internal blocks and off-chip devices.

5.3.1 H3DD Core

The H3DD Core is a pre-existing design done by Zhu which was imported, for simplicity. The primary function is to program and control the H3DD-UM front-end ASIC for standard readout. Since the H3DD-UM functionality will not be described in this work, it is not

necessary to describe the H3DD Core in detail here. However, one important upgrade made for the 4th revision is the migration from the operation of a single front-end ASIC to the operation of four front-end ASICs. In order to interface with four H3DD-UM ASICs while only making use of a single ADC, the implemented readout scheme is such that all four ASICs are readout any time a trigger is detected. While this does not achieve the maximum throughput of four H3DD-UM ASICs, it is a step towards the goal of increased throughput. The main change required for this functionality was to increase the number of I/O pins to accommodate 4 ASICs.

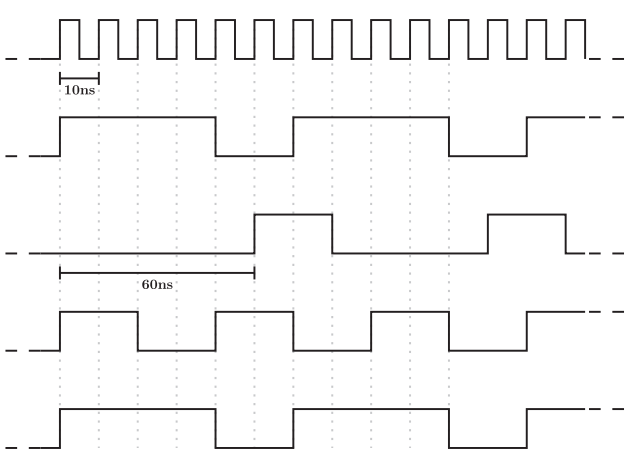
5.3.2 Clock Generation

One external clock, at a nominal operating frequency of 100 MHz, is required to drive the DAQ-DSP ASIC. Internally, multiple derived clocks are used to drive separate components, as shown in Table 5.2. The DSP Core and H3DD readout must run on exactly the same clock for this revision since there is no FIFO between the H3DD-UM sample read-in and the DSP Core pipeline. In other words, it is necessary for one sample to enter the DSP Core for every sample exiting the H3DD readout. The H3DD sampling clock controls the frequency at which waveform sampling occurs, and this frequency may be varied depending on the experimental setting. The ADC clock, on the other hand, typically includes a delay and a reduced duty cycle. The duty cycle is set by the designer, Seungheun Song, and the delay is controlled such that the ADC samples at the correct time. During readout, the H3DD-UM ASIC provides the analog samples after the falling edge of the readout clock. Since the ADC samples on the rising edge, the clock is delayed to match the sample readout, as shown in Table 5.2.

Each clock is generated from the master clock using a programmable lookup table (LUT) containing the waveform edge sequence. During operation, a counter is used to access the LUT to determine whether the clock output is high or low in the next cycle. Therefore, the output clock is actually a division of the input master clock. For example, if the LUT is programmed to contain `4'b1100`, then the output clock will go high for two master clock cycles, and then low for the following two clock cycles. If the master clock has a period of $T_M = 10$ ns, then this programming implements a clock with period of 20 ns and 50% duty cycle. Using this LUT programming implementation, the required computation during each clock cycle is minimal and a wide variety of clocks can be programmed.

The length of the LUT is set by the maximum possible waveform period – or minimum frequency – allowed by design. In this revision, a LUT length of 64 is used which allows a

Table 5.2: DAQ-DSP Clock Specification

Clock	Typ. Frequency (MHz)	Typical Timing Diagram
Master (f_M)	100	
DSP Core (f_D)	25	
ADC	f_D	
H3DD Sampling	50	
H3DD Readout	f_D	

maximum period of $64 \times T_M$, or 640 ns with the typical driving clock at 100 MHz.

5.3.3 SPI Communication

Data is transferred off the DAQ-DSP ASIC using a standard SPI interface comprised of MISO, MOSI, CLK, and CSN signals. Within the chip, the H3DD Core uses a different SPI receiving convention than the DSP Core and other control blocks, such as the clock generation circuit. This discrepancy exists due to the pre-designed nature of the H3DD Core, as noted in Section 5.3.1. To switch between the two SPI endpoints, one additional SPI control bit called `SPI_SW` is included. In future revisions, this convention may be resolved so that the SPI communication requires only the 4 standard bits. In prior revisions, two entire SPI channels (8 bits total) were used. For the following explanation, the two SPI channels will be called the "H3DD SPI" for all H3DD Core related endpoints, and the "DSP SPI" for the DSP Core and all other control blocks besides the H3DD Core.

All DSP SPI transactions use 32-bit words. When the 32-bit word arrives, the top 16-bits are interpreted as the `program_select` field, and the bottom 16-bits are interpreted as the `program_in` word.

The `program_select` field is broken into several different fields, as shown in Figure 5.11. Bits 23 to 16 are used for the DSP addressing, and each stage uses a subset of those bits, as

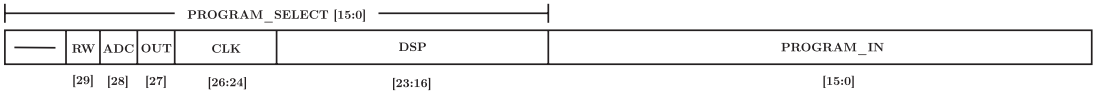


Figure 5.11: DSP SPI 32-Bit word decoding format.

detailed in Table 5.3. The clock generator requires 3 bits, 26 to 24, to address each of the generated clocks. All other sub-fields are one bit enables to various programmable blocks. Bit 27 is used to program which FIFO is connected to the output. Bit 28 is used to enable or disable the ADC, and bit 29 is used to signal that, instead of writing to memory, the selected address should be read out to the respective FIFO.

Table 5.3: DSP Core programmable parameters and addresses

	Address Bits	Register Data	Sub-address	Number of Registers	Total Bits
Stage 1	[1:0]	Baseline Start [6:0] Baseline Length [10:7] Polarity Switch [11]	1	1	12
		Decay Coefficient [10:0]	2	130	1300
Stage 2	[5:2]	FIR Parameter [15:0]	1, 2	256 (x2 Filters)	8192
		IIR Trapezoidal Parameter [15:0]	3, 4, 5	3 (x3 Filters)	144
		IIR CRRC4 Parameter [15:0]	6, 7, 8	32 (x3 Filters)	1536
		Output Assignment [2:0]	9, 10	3 (x6 Outputs)	288
Stage 4	[6]	Trigger Threshold [15:0] Anode Search Start [23:16] Anode Timing Threshold [31:24] Cathode Search Start [39:32] Cathode Timing Threshold [47:40]	1	1	48
Buffer	[7]	Buffer Option [2:0] Channel Selection [10:3]	1	1	11

For debugging purposes, a memory readback option is included. When the incoming instruction has bit 29 (RW) set to 1, the contents of the given address will be sent to the respective readback FIFO. For several addresses, data is shifted in rather than existing in a simple 32-bit register. For example, the FIR filter parameters for a single filter in the DSP Core have only one address for 256 parameters. To select which parameter is read out, the `program_in` field is used to select the address. In this case, if `program_in` is set to 32, then the 33rd FIR parameter will be read out. A sample protocol using the memory readback to verify the FIR programming is outlined here:

1. Send 256 sequential instructions with the FIR address and the desired FIR parameters to program the filter.

2. Send 256 sequential instructions with the FIR address, the "read-back" flag enabled, and sub-addresses (0 through 255) of each FIR parameter.

3. Send an instruction to set the output stream switch to the DSP read-back FIFO.

4. Send 256 empty instructions and record the output via the SPI MISO signal.

The 32-bit words recorded on the SPI MISO should exactly match the programmed parameters, otherwise an error occurred during programming or readback.

5.3.4 FIFO Interfaces

The FIFO interface used for the 4th revision follows the established design outlined in [31]. The functional diagram is shown in Figure 5.12. The notable characteristics of the design are the two register pairs – `sync_r2w` and `sync_w2r` – between the write side to the read side, as well as the gray code counter that is not shown at the top-level. The register pair is a standard method for guaranteeing that one bit cannot take on a meta-stable state, causing unknown values or unpredictable behavior. The variables that are scrutinized in this case are the read and write pointers. In order to know if the FIFO is empty or full, a comparison of the two pointers must be done before determining if a read or write is valid. Therefore, the read and write pointers must cross between clock domains safely. Using the double register method no longer works for more than one bit because the meta-stable states grow exponentially as the number of bits increases. The only way to safely cross the clock domain in this case is to ensure that only one bit of the read and write counters changes at any given time. This is accomplished using gray codes, a counting method designed specifically so that

only one bit changes for each increment. The details regarding gray codes, and conversion to and from standard binary are included in [31].

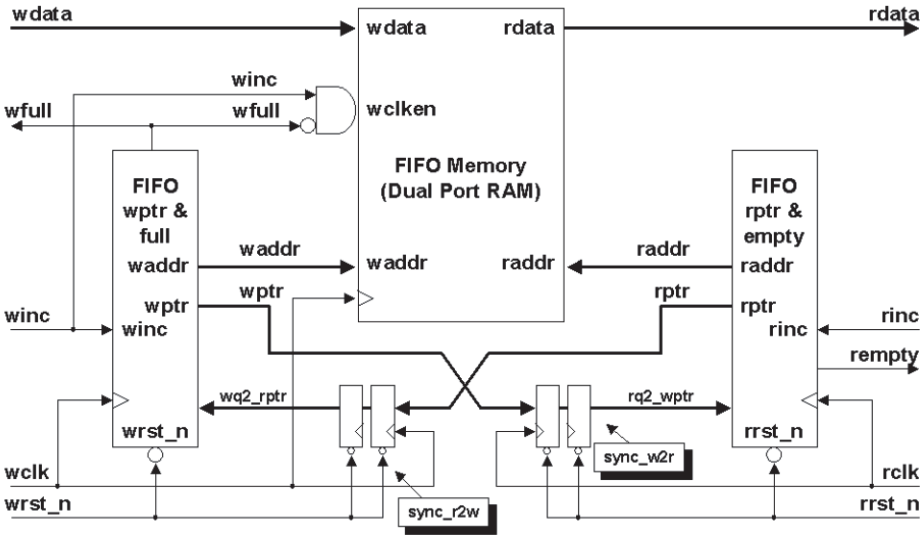


Figure 5.12: Standard asynchronous FIFO design [31].

Six total FIFOs are used to control the flow of received and transmitted data. Each FIFO may have a different depth and width, but all are instantiated from the same functional RTL. Clock domains and data streams are separated using the various FIFOs, as described in Table 5.4.

Table 5.4: DAQ-DSP FIFO Specification

FIFO Direction	Depth	Write Clock	Read Clock	Purpose
RX	8	SPI	DSP	DSP Instruction
	8	SPI	Master	Control Instruction
TX	8	Master	SPI	Control Readback
	8	DSP	SPI	DSP Readback
	32	DSP	SPI	DSP Data
	256	DSP	SPI	DSP Buffer

On the transmit (TX) side, only one FIFO is connected to the output at a given time. The connected output is selected using bit 28 (OUT). The readback FIFOs and DSP buffer

FIFO are included solely for debugging purposes. During standard operation, the DSP data FIFO should be connected to the output to read out event packets.

5.4 Placement and Synthesis Results

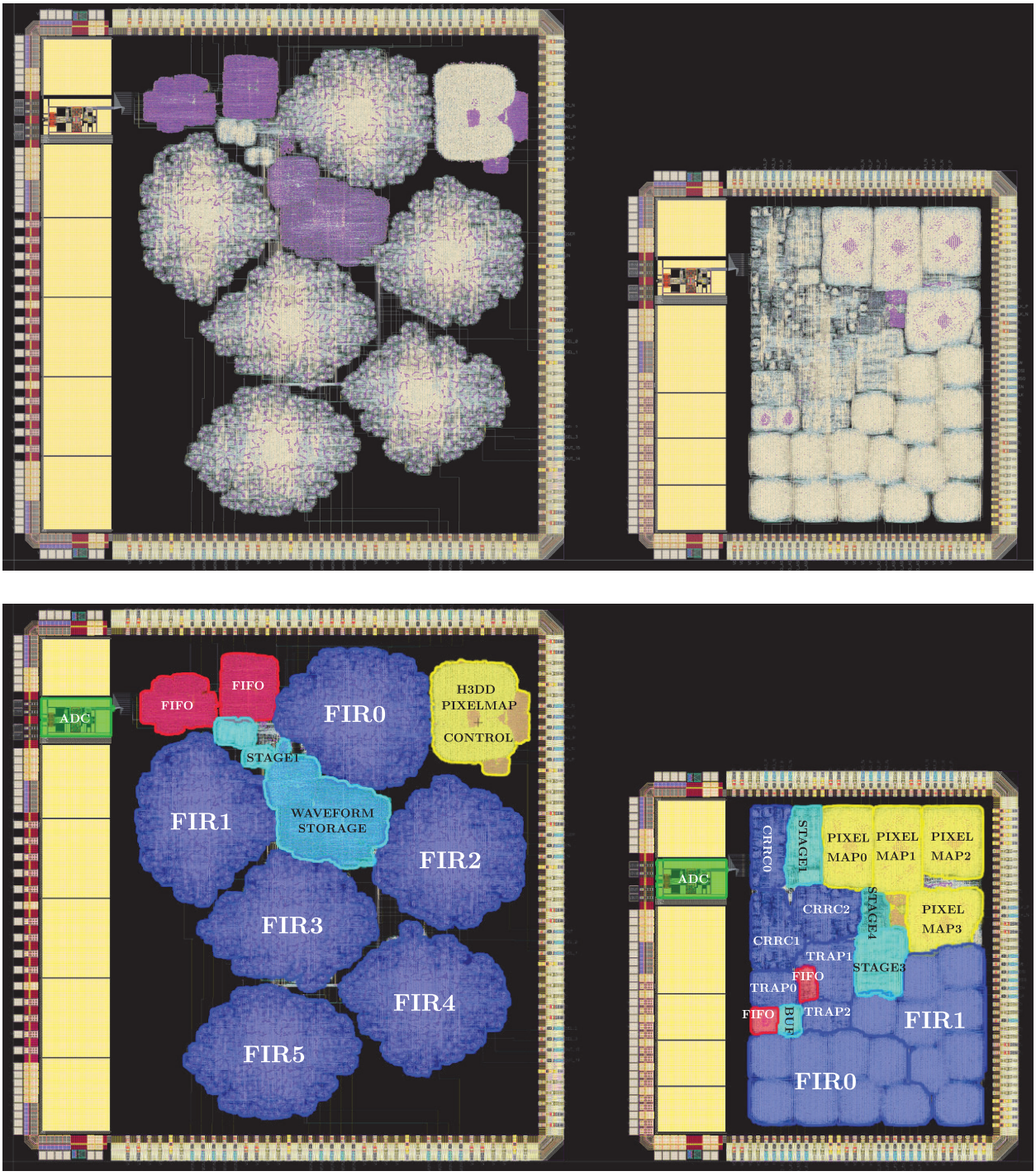
The resulting 4th revision ASIC is shown – and compared to the 3rd revision – in Figure 5.13. In the top row, the final layout of the DAQ-DSPv3 is illustrated to take 4.2×4.2 mm^2, while the final layout of the DAQ-DSPv4 is shown to take only 3×3 mm^2, a reduction of nearly 50%. The 3rd revision clearly had room to be compressed to an extent, but certain design choices were also necessary to achieve the areal reduction. Also of note is the minimum timing slack of 3.342 ns on a 100 MHz clock path. Such a margin ensures that no timing related issues will occur.

The bottom half of Figure 5.13 shows a roughly annotated breakdown of the area usage in each revision. In dark blue, the filtering blocks are highlighted. While the FIR functionality is consistent between revisions, the base RTL structure was altered in order to break each FIR into several sub-blocks. As is evident in the 4th revision layout, this allows a more optimal placement. The FIR and IIR areas may also be compared. By eye, it is clear that each individual IIR CRRC or IIR trapezoidal filter requires far less area than one FIR filter. The FIR filter requires approximately $539,000$ μm^2, and the IIR filters each require approximately $169,000$ μm^2 a reduction of 69%.

In yellow, the H3DD Core pixelmap and control blocks are highlighted. The critical difference here is that in the 4th revision, 4 pixelmaps are required to interface with 4 H3DD-UMv4 ASIC's, necessitating 4 times the area.

The cyan blocks are related to the DSP Core non-filtering blocks. One key contrast here is that the waveform storage – known just as stage 3 in the 4th revision – is approximately half the size. This relates to unnecessary waveform storage which was included between stages 1 and 2 in revisions 1-3. In the 4th revision, stage 1 requires more area than in the 3rd revision due to the inclusion of one IIR core used to execute the exponential deconvolution. The revision 3 stage 1 used simple fixed point calculations for the deconvolution, resulting in unacceptable rounding errors. In revision 4, the FP-based IIR core described in Section 5.2.2.3 is used.

Finally, the red blocks correspond to the FIFO's used. Since area was not a primary concern in the 3rd revision, the FIFO's are relatively larger. The FIFO size was optimized in revision 4 to use only the necessary depth.



(a) Left: DAQ-DSPv3. Right: DAQ-DSPv4.

Figure 5.13: ASIC layout annotation and comparison of revision 3 and revision 4.

CHAPTER 6

DAQ-DSP ASICv4 - Measurements

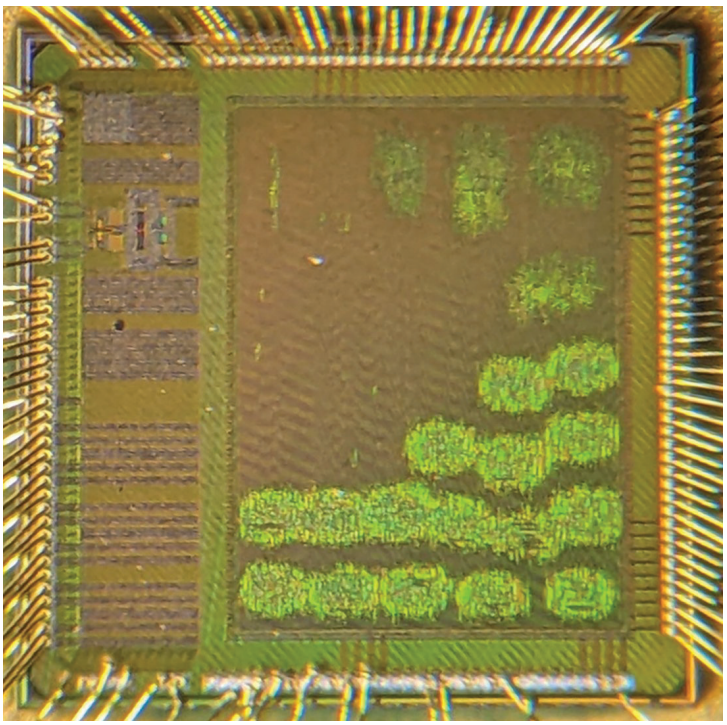


Figure 6.1: DAQ-DSPv4 sample die microscopic image.

6.1 Motivation

The DAQ-DSPv4 was taped out on March 7th 2024. Between the tape-out and arrival of the chip on June 5th, a compact system was designed as a prototype to demonstrate the advantages of the ASIC. For revisions 2 to 4 of the ASIC, simple test PCBs were also used

to provide debugging capabilities. These boards typically included headers to allow driving or monitoring of all ASIC pins. Due to the simplistic nature of these boards, they are not described in detail here. The compact system detailed in this chapter is a more aggressive design aimed at making the most compact, low noise, and low power system possible.

6.2 Compact System Design

6.2.1 Functional Block Diagram

Figure 6.2 illustrates the major components of the compact test system. Four CZT crystal and H3DD-UM ASIC modules are connected to the motherboard. The DAQ-DSPv4 ASIC is designed to interface with four H3DD-UM ASICs, but when a trigger occurs, all four must be readout in sequence. To accommodate this structure, only one ADC and one ADC driver is required. Since only one H3DD-UM ASIC will be read out at a time, all analog outputs are linked together by 0 Ω resistors. All communication between the DAQ-DSPv4 ASIC and the H3DD-UM ASICs occurs through a digital bus. To bring data out of the system, a Cypress FX3 USB chip is connected via a SPI bus to the DAQ-DSPv4 ASIC. Communication is established by connecting the USB port to a computer.

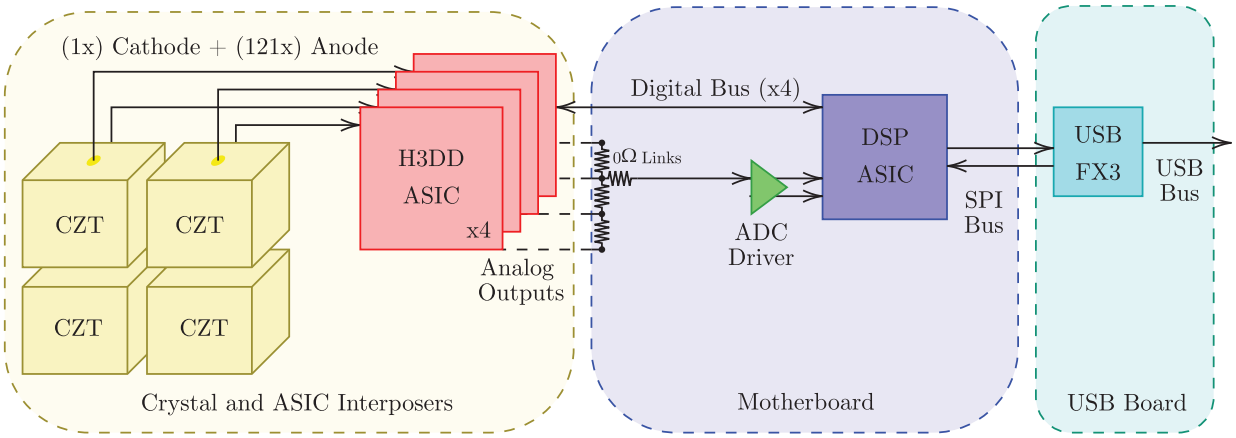


Figure 6.2: DAQ-DSPv4 compact system functional diagram.

6.2.2 Power Distribution Strategy

The system is designed to receive power from a single port which can range from 5 V to 17 V. One option is to supply power from a wall adapter at 12 V, but in future revisions the design may be altered to be powered directly from the USB port. The power requirements are shown in Table 6.1. Towards low noise performance, several regulators are used to keep the power domains separate. Even among the analog supplies, separate 1.2 V regulators are used.

Table 6.1: Compact System Power Requirements

Chip	Purpose	Typ. Voltage (V)
H3DD-UM ASIC	Preamplifier Supply	1.2
	Analog Supply	1.2
DAQ-DSP ASIC	Digital I/O	1.2
	Digital Core	1.0
	ADC Supply	1.2
	ADC Positive Reference	1.2
	ADC Common Mode	0.6
	ADC Negative Reference	0
ADC Driver	Supply	2.8

The full power distribution scheme is shown in Figure 6.3. Notably, two sets of 1.2 V regulators are used for the H3DD-UM ASICs. This method helps to distribute the heat dissipation across the PCB, and also provides the best current return paths. On one branch, a 2.8 V regulator is used due to the constraint of the ADC driver, and on the other branch a 1.8 V regulator is used to provide a more gradual power drop. A switching regulator is used for the input (3 V) regulator, and all others are linear drop-off (LDO) regulators.

6.2.3 PCB Design

As shown in Figure 6.2, the total measurement system consists of 3 components. The motherboard and USB board are the two main PCBs designed specifically to mount and demonstrate performance of the DAQ-DSPv4 ASIC.

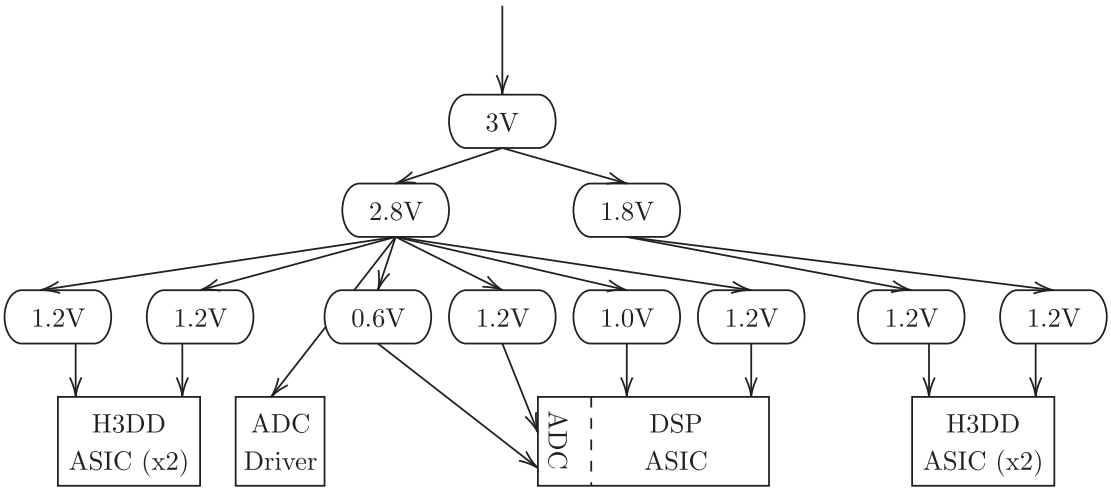


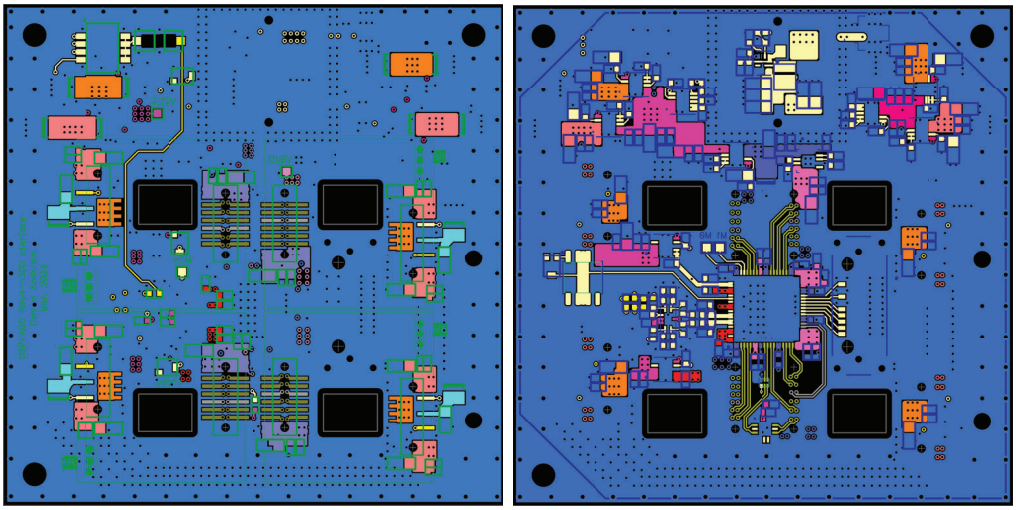
Figure 6.3: DAQ-DSPv4 compact system power distribution.

6.2.3.1 Motherboard Design

Figure 6.4 shows the top and bottom layers of the 6-layer motherboard. The top layer of the board is shown in Figure 6.4a. The important features are the 4 mounting patterns for 4 H3DD-UMv4 ASICs. For the most part, high-speed signals are kept away from the top layer to prevent any interference with the CZT modules and front-end ASICs that will be mounted. One debug analog trace is routed on the top layer, but in typical operation it may be disconnected. Each front-end landing pattern also includes a board cutout to bring in a cold finger for heat dissipation.

The bottom layer of the board is reserved for power distribution and connections to the DAQ-DSPv4 ASIC, as shown in Figure 6.4b. The DAQ-DSPv4 ASIC is mounted centrally with the various H3DD-UM I/O signals taking short traces to the respective pins. The upper section of the board is used to bring in power to various LDOs used for regulation. On the left-hand side, all analog traces are handled. The H3DD-UM front-end analog outputs are connected through 0 Ω connections as discussed in Section 6.2.1. On the right-hand side, all high-speed signals are brought in through a specialized inter-board connector, and routed through as small of traces as possible in order to reduce interference. The inter-board connector is fully shielded, and half of the pins are used for grounding to provide clean return paths.

The internal layers are omitted, but are stacked as follows:



(a) Compact system board top (b) Compact system board bottom

Figure 6.4: DAQ-DSPv4 compact system motherboard PCB design.

2. Solid ground plane
3. Digital power routing, and analog traces
4. Analog power routing, and digital traces
5. Solid ground plane

On layer 3, the digital power routing is done through central fills underneath the digital board section, and the analog traces are routed along the board externals away from the digital signals. Similarly, on layer 4, the analog power is confined to the board sides, and the digital traces are all underneath the central digital section of the board.

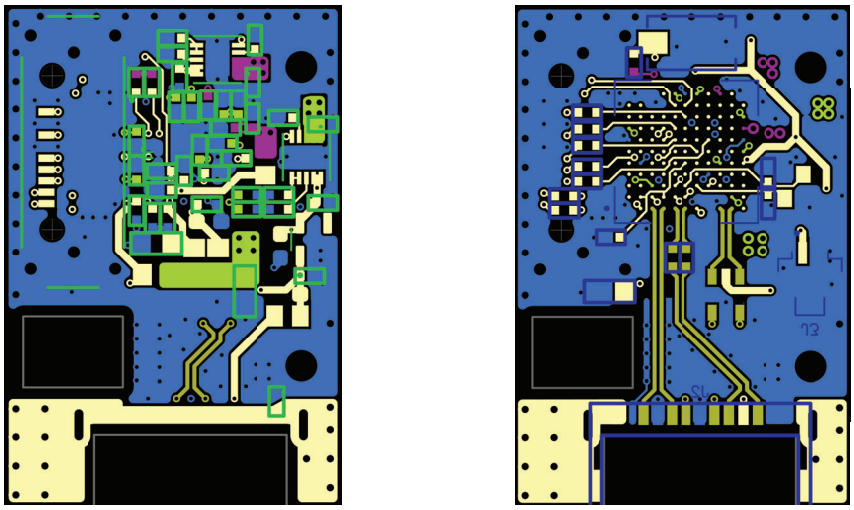
The motherboard takes a total area of 65×65 mm^2, of which the H3DD-UM landing patterns require approximately 46.5×46.5 mm^2. The landing patterns are the bare minimum necessary for the board to support 4 H3DD-UMv4 modules, so the areal efficiency can be considered as follows:

$$\frac{46.5 \times 46.5 \text{ mm}^2}{65 \times 65 \text{ mm}^2} \approx 51\%$$

In other words, it is theoretically possible to reduce the area by another $\approx 51\%$ if the power routing is further optimized. For now, an area of 65×65 mm^2 is considered satisfactory.

6.2.3.2 USB Design

The USB design top layer and bottom layer are shown in Figure 6.5. In this case, only 4 layers were required for the USB board. The top layer, which will face towards the motherboard when mounted, contains two power regulators to generate 1.2 V and 1.8 V and all required bypass capacitors. The bottom side, which will face away from the motherboard, contains the Cypress EZ-USB FX3 chip which handles communications to the host system. The USB signals are routed in through the minimal trace length to reduce interference. Similarly, the high-speed signals which interface with the DAQ-DSPv4 ASIC are brought out directly on this layer as much as possible.



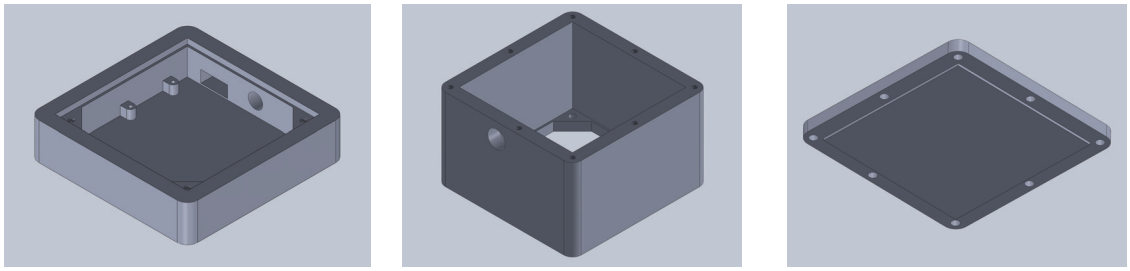
(a) USB board top (b) USB board bottom

Figure 6.5: DAQ-DSPv4 compact system USB PCB design.

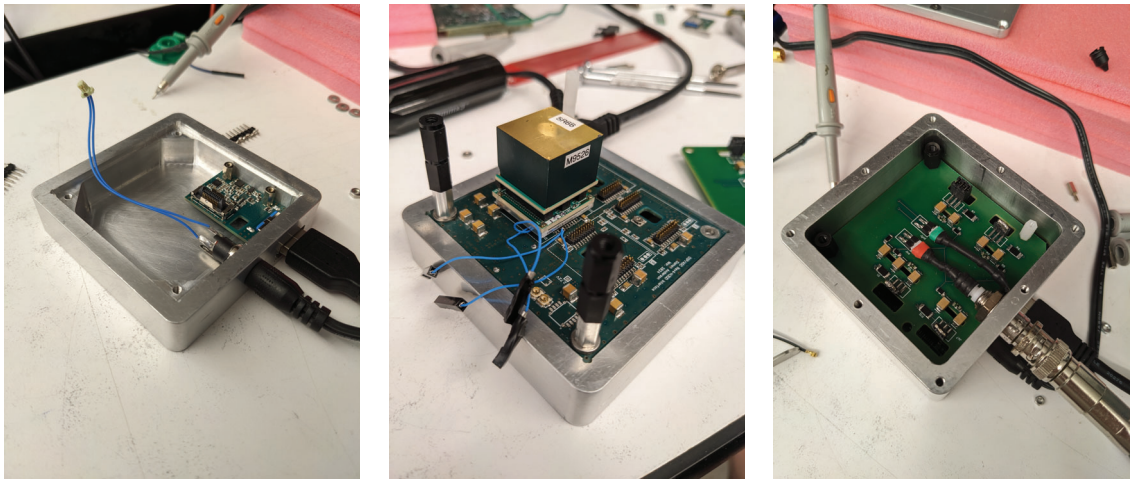
On this board, the two internal layers are power and ground, as is typical. The area of this board is 36.5×24.0 mm^2, a subset of the motherboard. The additional space required by this board comes in the vertical direction. In future board revisions, if possible, the USB chip may be mounted directly on the motherboard, although the constraints imposed on the area would be extremely difficult to work with.

6.2.4 Enclosure Design and Assembly

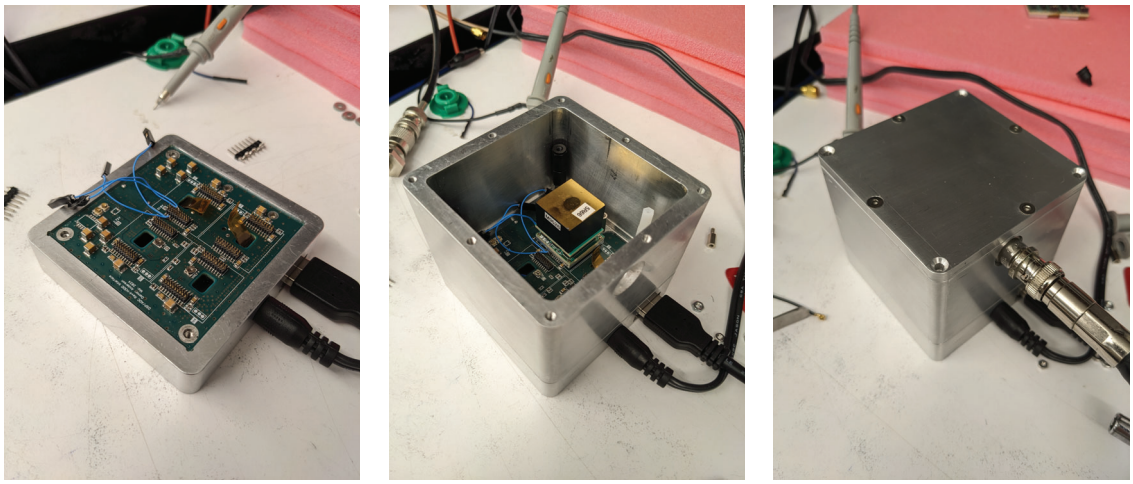
To accompany the system boards, a physical enclosure was designed and manufactured to demonstrate the commercial potential, and to allow the system to be brought safely to -3000 V bias. The physical enclosure consists of 3 sections to facilitate debugging methods.



(a) Bottom section (b) Middle section (c) Top section



(d) USB mounted (e) One CZT mounted (f) High voltage board



(g) Motherboard mounted (h) Middle section attached (i) Fully assembled

Figure 6.6: Design and assembly of the 3-part enclosure used to house the DAQ-DSPv4 compact test system.

As shown in Figure 6.6, the enclosure is divided into a bottom, middle, and top section. The bottom section houses both the USB board and motherboard, as shown in Figures 6.6d and 6.6g, respectively. The bottom section contains two ports – power and USB. The USB board directly slides into the USB port, and is fixed in place by two mounting screws. The motherboard is then connected to the power header, and inserted above the USB board to be flush with the enclosure section.

Next, the middle section is attached and all mounting screws are fastened. Four CZT and H3DD-UMv4 ASIC modules can be inserted as shown in Figure 6.6h. For debugging purposes, the order of the previous steps may be reversed so that all board components are exposed for oscilloscope probing while the CZT module is mounted.

Figure 6.6f shows the enclosure after the high voltage board is attached and screwed into place. The high voltage board is not discussed in detail since it is a basic design. The high voltage is brought in through an SHV cable, and passed through a series of RC filters before being applied to each CZT cathode. Finally, the enclosure top section is screwed into place to provide a full Faraday cage for standard operation.

6.3 Compact System Measurements

6.3.1 Power Consumption

The DAQ-DSP ASIC power is measured based on the current draw from a standard power supply. The results and comparison to revision 3 are summarized in Table 6.2. The "power down" mode shown is entered by disabling all internal clocks.

Table 6.2: DAQ-DSP Power Consumption Summary

	Power Consumption (mW)		
	Revision 3	**Revision 4**	**Revision 4 (Power Down)**
Core (1.0V)	58	30	22
I/O (3.3V/1.2V/1.2V)	30	5	0
Total	**88**	**35**	**22**

The voltages annotated in the power column indicate the supplies used for the core cells and I/O cells in each of the respective measurements. In particular, while revision 3 and revision 4 both use 1.0 V for the core, the revision 4 systems are designed to use 1.2 V for the I/O as opposed to the 3.3 V used in revision 3. Notably, the power has been reduced by ≈60% from revision 3 to revision 4. This is in part due to the shift from FIR to IIR filters which consume lower power. Various reductions in area (i.e. waveform storage changes) also help to reduce the power consumed. The ADC is estimated to draw around 1 mW. The 4th revision design also includes the option to turn off clocks, which allows the power draw to be reduced to only 22 mW.

It is worth noting that each H3DD-UMv4 ASIC draws approximately 300 mA at 1.2 V, or ≈360 mW. For the foreseeable future, the order of magnitude for the H3DD-UM power draw will not change, which means that overall system power will be on the order of magnitude of 4×360 mW = 1440 mW. Since 35 mW is far lower than the total power consumed by the system, it can be considered negligible, although further power reductions are always favorable.

6.3.2 DSP Core Verification

The DSP Core is verified using three methods:

1. Using the SPI read-back method outlined in Section 5.3.3
2. Using the DSP buffer described in Section 5.2.5
3. Using the standard event readout from the chip with a test pulse

The SPI read-back method proved to be a successful upgrade to the ASIC in revision 4. The DAQ software was rewritten to include the protocol detailed in Section 5.3.3, and the parameters from all DSP components were proven to be programmed correctly. Figure 6.7 shows the current DAQ software page used to download the DSP settings. When the user clicks "Download" they will be instantly notified if any of the DSP settings failed to download. Thus far, there are no issues with downloading and reading out the DSP parameters.

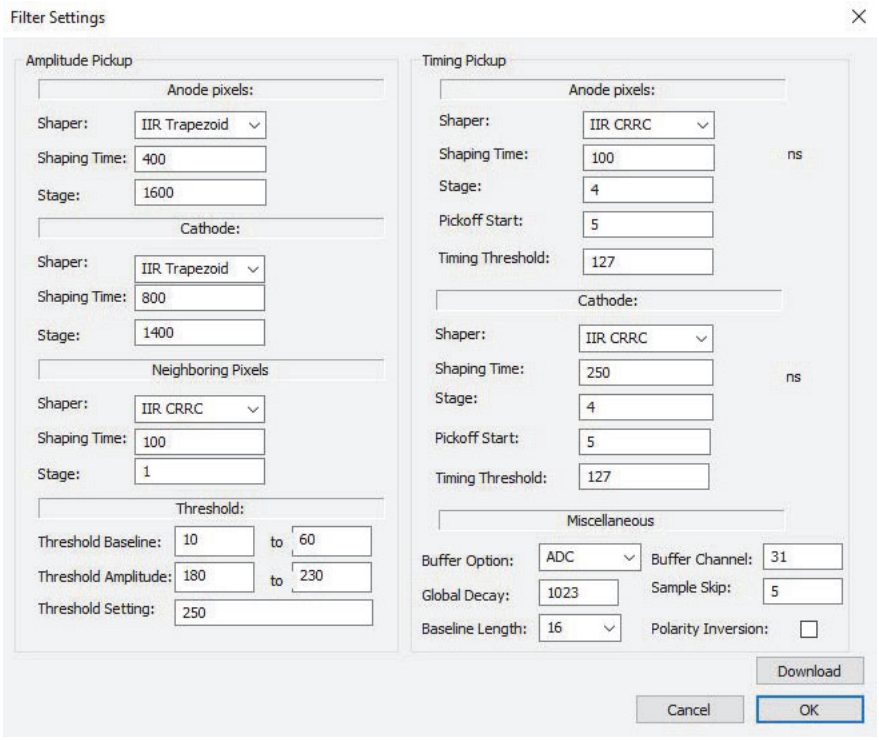
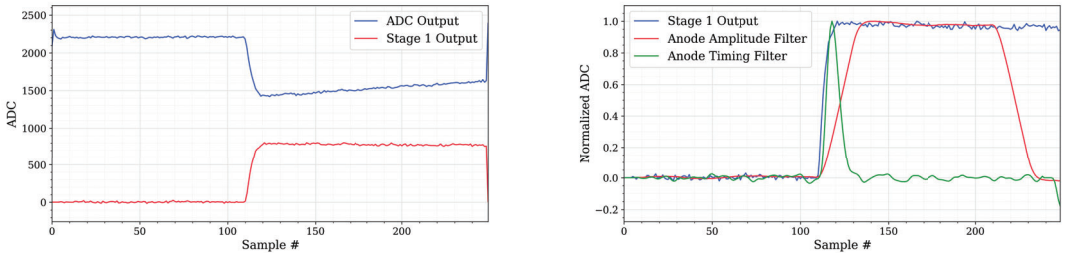


Figure 6.7: DAQ software filter dialog used to program the DAQ-DSPv4 ASIC.

The DSP buffer adds the capability to interrogate the output of each step within the DSP Core. So far, the readout buffer has provided valuable results. Figure 6.8 shows an example

of real waveforms read out from the DAQ-DSPv4 using the buffer. This step proves that, after the parameters are correctly programmed, the filters are performing as expected. Aside from a minor issue in stage 1 – which will be detailed in Section 6.3.3 – the buffer outputs have shown the expected DSP Core functionality.



(a) ADC and stage 1 waveform

(b) Stage 1 and anode waveforms

Figure 6.8: Examples of waveforms read out from the DAQ-DSPv4 buffer.

Finally, the standard DSP readout method is used to plot the amplitude and timing from a test pulse input. Due to an error in the ASIC configuration method, it is difficult to rapidly scan the test pulse with the current system; however, the filtered amplitude reacts to the change in test pulse as expected. Results shown in Section 6.3.4 further verify that the DSP Core works as expected.

6.3.3 Known Bugs

Two major known bugs were found during the revision 4 evaluation. While both issues were resolved to finally show a spectrum, these bugs added some complications and will be revised in the final design cycle. The first of the two design issues involved programming the H3DD-UMv4 ASIC. Two differential signals are used to control the mode of the H3DD-UMv4: READ and ENA. During the configuration process, the READ and ENA signal pairs must both be set to VDD in order to properly program the front-end ASIC. However, one of the major changes made to the H3DD Core firmware during the 4th revision design cycle was to expand the control capability from a single ASIC to 4 ASICs. This modification requires the control signals to be reworked so that each ASIC can be controlled properly. During the rewrite, the READ and ENA signals were set to the wrong signal in the configuration function, and the result is that the two signals are set to ground instead of 1.2 V. Luckily, the configuration signals have no strict timing requirements, so one solution is to pull those

communication lines high externally when configuration is desired. This proved to be a successful solution, and new PCB designs were implemented to allow the digital control of the READ and ENA signals.

The second major bug relates to the baseline calculation of the DSP Core. As shown in Figure 6.8a, the stage 1 output retains a small baseline offset even after the rolling average is completed. This issue is due to the number of bits used to store the baseline. With an excess baseline offset, the exponential decay deconvolution will be incorrect and cause degradation. However, this can be resolved by setting the deconvolution parameters – the $\frac{d}{1024}$ factor described in Section 5.2.1 – to ≈ 1 so that no deconvolution takes place. Without deconvolution the energy resolution will be worse than expected, but the result is still passable to generate a spectrum. The best energy resolution will only be achieved with complete baseline subtraction and deconvolution used during the digital processing.

6.3.4 Spectrum Performance

Using the compact test system, a 15 mm thick CZT crystal was biased to -3000 V to evaluate the spectral performance of the end-to-end H3DD-UMv4 and DAQ-DSPv4 system using the 662 keV photopeak from a Cs-137 check source. In addition, the count-rate capability of the system is evaluated using one CZT module. Here, it is important to show a count-rate and resolution improvement from the 3rd revision DAQ-DSP ASIC.

The first calibrated Cs-137 spectrum is shown in Figure 6.9, the filter settings are the same standard ones summarized in Table 1.1, and the measured counts are summarized in Table 6.3. The spectrum was measured using a single detector over a duration of 45 minutes with a total count-rate of 2992.4 cps. The spectrum was measured using the 860 keV dynamic range of the H3DD-UMv4 ASIC, and trigger-only readout. The overall single-pixel resolution and non-anode side resolutions were measured to be 0.62% and 0.55%, respectively. The best expected resolution for the particular detector measured is $\approx 0.4\%$, so the result still falls short of the expectation. The performance degradation may be attributed to multiple unresolved issues. Mainly, the cathode signal has not yet been completely stable during the measurement due to high crystal leakage in the compact system which has not been debugged. The DSP related issues outlined in Section 6.3.3 also contribute to the resolution degradation, although the result is not expected to be significantly worse due to these minor issues. Furthermore, the functionality verification methods outlined in Section 6.3.2 prove that the DSP is working as expected. System debugging will continue with the arrival of redesigned PCBs which allow for the H3DD configuration cycle to run digitally.

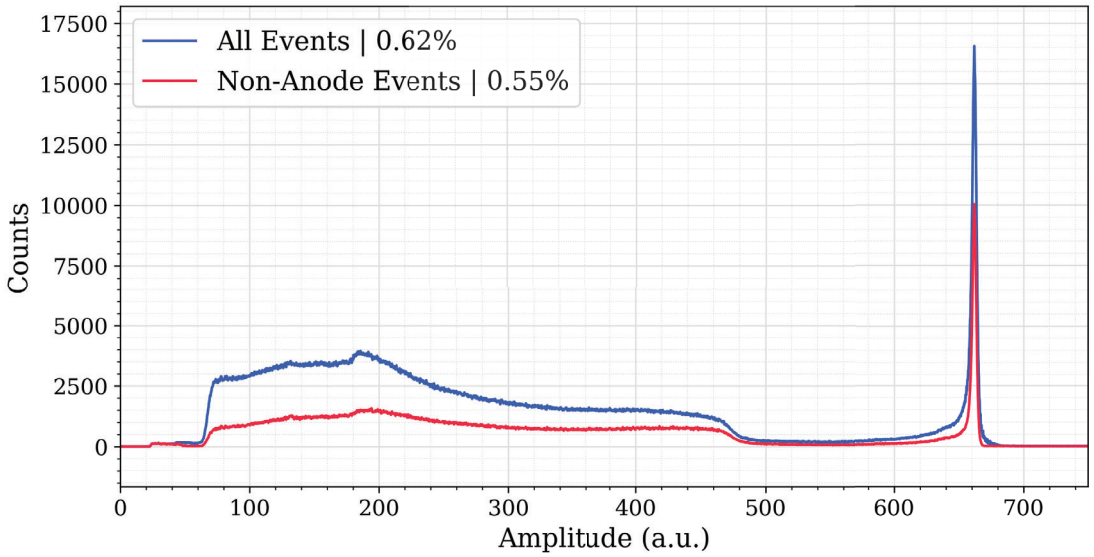


Figure 6.9: Calibrated 1-pixel Cs-137 spectrum from detector 5R74.

Two additional spectra are shown in Figure 6.10 to demonstrate the timing pick-off and neighbor waveform processing capabilities. Although the CRRC filtering functionality can be directly confirmed using the buffer output, the timing pick-off calculation can only be verified using the event readout. The standard timing spectrum shape shown in Figure 6.10 is generated by calculating the difference in cathode and anode timings read out from the DAQ-DSPv4 ASIC for each event. The spectrum shows the characteristic shape that reflects the depth distribution of gamma interactions, with the right-hand side corresponding to the cathode-side where most interactions take place. The sub-pixel spectrum shown in Figure 6.10 is generated by operating the H3DD-UM ASIC in neighbor readout mode, and then performing the standard sub-pixel calculation [34] with the received neighbor amplitudes. The sub-pixel spectra also roughly reflect the expected shape; however, the sub-pixel resolution is clearly lower than expected, as indicated by the lack of sharp rising edges that would correspond to the physical pixel edges. This is an indication that the simple neighbor processing methods implemented on the DAQ-DSPv4 ASIC are not satisfactory and should be re-examined for the final design cycle. At the very least, the timing and sub-pixel spectra show that the DAQ-DSPv4 functions are working as intended for this revision.

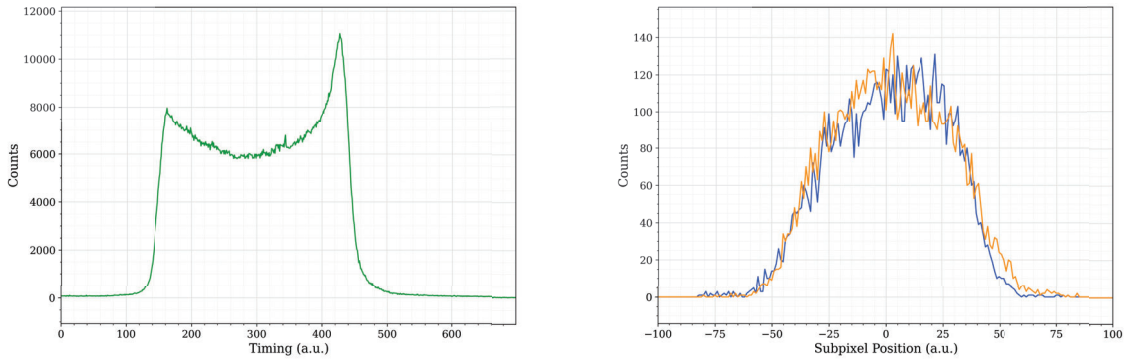


Figure 6.10: Left: Timing spectrum measured using the DAQ-DSPv4 system to illustrate timing pick-off functionality. Right: Sub-pixel spectrum measured using the DAQ-DSPv4 system to illustrate neighbor readout and filtering capability.

Table 6.3: Measured Count-rates from the DAQ-DSPv4 ASIC

	Counts	**Count Rate (1/s)**
1-Pixel	5,928,095	2195.6
2-Pixel	1,755,072	650.0
3-Pixel	353,893	131.1
4-Pixel	42,294	15.7
Total	8,079,354	2992.4

6.4 Conclusions and Future Revisions

The DAQ-DSPv4 was designed, fabricated, and tested as the final cycle of the original DAQ-DSP project contract. The aim of the project was to design and implement a chip which controls the H3DD-UM front-end ASIC, converts the output samples from analog to digital, and processes the digital samples using standard filtering techniques to yield the amplitude and timing results.

The 4th revision design cycle consisted of several major changes including the redesign of all chip-boundary FIFOs to handle asynchronous clock-domain-crossings, the complete rewrite of the DSP Core to use majority IIR filters, and the transition from single-ASIC control to four-ASIC control. Additional quality of life improvements include the DSP readout buffer, the SPI read-back functionality, improved timing slack, and the decrease in chip area from 4.2×4.2 mm^2 to 3×3 mm^2. Moreover, the chip power reduced by a factor

of ≈ 2 due to the filtering modifications.

The chip proved capable of measuring high count-rate and high resolution calibrated Cs-137 662 keV spectra despite two major bugs still remaining. A total count-rate of 2992.4 cps was measured using a single CZT module, and a single-pixel resolution of 0.62% was achieved. Although the resolution is yet to match the best expected performance, it is most likely that this is due to system design issues – such as cathode noise – rather than digital noise incorporated in the filtering process or sampling from the ADC. The DAQ-DSPv4 is considered a successful chip by delivering this result.

One final design cycle – the 5th revision – has been proposed and will be executed to make a version of the ASIC which includes the best possible throughput and also removes any of the remaining bugs. The goals of the final revision are summarized as follows:

- **Bug Fixes:** Primarily, the H3DD-UMv4 ASIC configuration issue must be removed, as well as all other bugs detailed in Section 6.3.3.
- **Independent H3DD-UMv4 Control:** The final revision will allow each of the interfacing H3DD-UMv4 ASICs to be fully independent as opposed to the current scheme of reading out each ASIC when a trigger occurs. To enable this, 4 ADCs will be included, and the DSP operation speed will be increased to 100 MHz.
- **USB IP Inclusion:** The option to bring a USB 2.0 IP on chip to improve the system compactness will be evaluated.
- **SRAM Optimization:** In the current implementation, the H3DD pixelmap RAM (shown in Figure 5.13a) is not optimized vendor RAM. For the final revision, these elements will be replaced with vendor RAM to save more space on chip.
- **Filter Optimizations:** Filter improvements will continue to be improved starting with the merger of some anode and cathode filters. For example, both anode and cathode amplitude filters are typically trapezoidal filters, but it is never the case that the two will be used at the same time since it is initially known whether the signal will be a cathode based on the channel. Thus, it is a waste to include two separate filters for the anode cathode. Instead, only two sets of parameters will be included to save space. Optimizations such as this will be explored for the final redesign.

Even still, further technological advances can be targeted in the same vein of digital ASIC design. The current design node – TSMC 65nm MS RF GP – is a relatively old node. Use

of newer nodes may enable lower power and lower area designs. Computationally complex ASICs involving the integration of a basic CPU may be proposed to enable end-to-end processing for various needs in an extremely compact form factor. Ultimately, it is desirable to enable systems in which the majority of the volume is taken by active detection volume while offering the user the standard readout capabilities of 3D-CZT.

CHAPTER 7

Conclusions and Outlook

Two primary studies have been conducted relating to advancing 3D-CZT signals processing techniques. The viability of 3D-CZT for coincidence applications was investigated by measuring the limits of timing resolution for current research systems. To determine the time between coincidence events, the cathode drift start was determined using a simple linear fitting procedure, as described in Chapter 2. Although timing resolutions as low as 10 ns [18] have been measured for CZT using more advanced techniques, the practical linear fitting technique proposed here was able to achieve a best resolution of 36.3 ns. However, it was shown that the timing resolution is approximately limited by the cathode signal noise, and as a result timing resolution improvements must be possible by improving system noise. This study is a building block in understanding the fundamental limits of 3D-CZT in coincidence measurements. Towards fully understanding the best coincidence set-up for CZT, the cathode noise dependencies and sensitivity limits must be further explored. Practical experiments using variable high voltage bias, and variable CZT thicknesses will provide insight into the magnitude of cathode noise reduction that is possible. The sensitivity will be determined not only by the detector intrinsic efficiency, but also by the electronic readout scheme used. The complexity associated with such an evaluation motivates another full investigation in the future.

The first successful measurements from a back-end ASIC – the DAQ-DSP ASIC – in use with a CZT module biased to -3000 V have been reported. The DAQ-DSP ASIC project aimed to prove that the signals processing, data acquisition, and analog-to-digital conversion components related to electronic readout for 3D-CZT could be reduced to a single chip, thus improving the power and footprint of the system. The fourth revision is the key result from the initial project, demonstrating power consumption on the order of 35 mW, using a die area of 3×3 mm^2, and reporting a best Cs-137 energy resolution of $\approx 0.62\%$. The energy resolution is further expected to reach the best measured resolution of 0.4% from the CZT

crystal in use after improvements are made to the test setup. The result shows that digital ASICs are not only viable, but a strong solution towards making a compact and low power system. Moreover, the technology node used to fabricate the DAQ-DSP ASIC is relatively old, suggesting that further improvements to the area and power are readily available. The DAQ-DSP ASIC, in that sense, is only a first step. One final revision of the DAQ-DSP – the 5th revision – will be completed to remove all bugs associated with the 4th revision, and to instantiate improvements discussed in Chapter 6. From there, ASIC commercialization and the further development of more advanced ASICs are to come.

BIBLIOGRAPHY

[1] M400, Custom Integrable Detector Module. Accessed: 2024-07-21.

[2] Medipix4. Accessed: 2024-07-21.

[3] EV509 series Preamplifiers for Radiation Detectors. Accessed: 2024-07-21.

[4] 2006, Proportional Counter Preamplifier. Accessed: 2024-07-21.

[5] Biograph Vision PET/CT Scanner. Accessed: 2024-07-15.

[6] Ieee standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, 2008.

[7] Sara Abraham. *Capability Demonstration of 3D CdZnTe Gamma-Ray Detectors in Extreme Environments*. PhD thesis, University of Michigan, 2023.

[8] Zhuo Chen. *Quantitative Measurement and Development of Back-end Processing System-on-Chip for the Pixelated CdZnTe Detector*. PhD thesis, University of Michigan, 2022.

[9] Gianluigi de Geronimo, George Iakovidis, Sorin Martoiu, and Venetios Polychronakos. The vmm3a asic. *IEEE Transactions on Nuclear Science*, 69(4):976–985, 2022.

[10] R.H. Dennard, F.H. Gaensslen, Hwa-Nien Yu, V.L. Rideout, E. Bassous, and A.R. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[11] Z. He, W. Li, G.F. Knoll, D.K. Wehe, J. Berry, and C.M. Stahle. 3-d position sensitive cdznte gamma-ray spectrometers. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 422(1):173–178, 1999.

[12] Zhong He. Review of the shockley–ramo theorem and its application in semiconductor gamma-ray detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 463(1):250–267, 2001.

[13] William R. Kaye. *Energy and Position Reconstruction in Pixelated CdZnTe Detectors*. PhD thesis, University of Michigan, 2022.

[14] Glenn F. Knoll. *Radiation Detection and Measurement, Fourth Edition*. John Wiley & Sons, Inc., 2010.

[15] Yong Lim and Michael P. Flynn. A 1 mw 71.5 db sndr 50 ms/s 13 bit fully differential ring amplifier based sar-assisted pipeline adc. *IEEE Journal of Solid-State Circuits*, 50(12):2901–2911, 2015.

[16] X. Llopart, J. Alozy, R. Ballabriga, M. Campbell, R. Casanova, V. Gromov, E.H.M. Heijne, T. Poikela, E. Santin, V. Sriskaran, L. Tlustos, and A. Vitkovskiy. Timepix4, a large area pixel detector readout chip which can be tiled on 4 sides providing sub-200 ps timestamp binning. *Journal of Instrumentation*, 17(01):C01044, jan 2022.

[17] Paul N. Luke. Single-polarity charge sensing in ionization detectors using coplanar electrodes. *Appl. Phys. Lett. 65 2884-2886*, 1994.

[18] L.J. Meng and Z. He. Exploring the limiting timing resolution for large volume czt detectors with waveform analysis. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 550(1):435–445, 2005.

[19] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, 38(8), 1965.

[20] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing, Third Edition*. Pearson Higher Education, Inc., 2010.

[21] Matthew Petryk. *Algorithms and Electronics for Processing Data from Pixelated Semiconductor Gamma-Ray Detectors*. PhD thesis, University of Michigan, 2023.

[22] F. Piro, P. Allport, I. Asensi, I. Berdalovic, D. Bortoletto, C. Buttar, R. Cardella, E. Charbon, F. Dachs, V. Dao, D. Dobrijevic, M. Dyndal, L. Flores, P. Freeman, A. Gabrielli, L. Gonella, T. Kugathasan, M. LeBlanc, K. Oyulmaz, H. Pernegger, P. Riedler, M. van Rijnbach, H. Sandaker, A. Sharma, C. Solans, W. Snoeys, T. Suligoj, J. Torres, and S. Worm. A 1-μw radiation-hard front-end in a 0.18-μm cmos process for the malta2 monolithic sensor. *IEEE Transactions on Nuclear Science*, 69(6):1299–1309, 2022.

[23] Simon Ramo. Currents induced by electron motion. *Proc. Ire.*, 27:584–585, 1939.

[24] Harold Rothfuss, Larry Byars, Michael E. Casey, Maurizio Conti, Lars Eriksson, and Christian Michel. Energy resolution and absolute detection efficiency for lso crystals: A comparison between monte carlo simulation and experimental data. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 580(2):1087–1092, 2007. Imaging 2006.

[25] Syed Adeel Ali Shah, Hubert Kroha, Marcello De Matteis, Andrea Baschirotto, and Robert Richter. 65 nm cmos 8 mv/fc, 14.6 ns rising time analog front-end for atlas muon drift tubes detectors. In *2023 18th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, pages 97–100, 2023.

[26] William Shockley. Currents to conductors induced by a moving point charge. *J. Appl. Phys. 9*, 635, 1938.

[27] Seungheun Song, Taewook Kang, Seungjong Lee, and Michael P. Flynn. A 150-ms/s fully dynamic sar-assisted pipeline adc using a floating ring amplifier and gain-enhancing miller negative-c. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–2, 2023.

[28] S. Tang, F. Bonini, M. Begel, M. Benoit, H. Chen, V. Filimonov, F. Lucca, D. Matakias, W. Qian, D. Sankey, and H. Xu. Prototype design of global common module for atlas experiment's phase-ii upgrade. *IEEE Transactions on Nuclear Science*, 70(9):2248–2255, 2023.

[29] Joyce van Sluis, Johan de Jong, Jenny Schaar, Walter Noordzij, Paul van Snick, Rudi Dierckx, Ronald Borra, Antoon Willemsen, and Ronald Boellaard. Performance characteristics of the digital biograph vision pet/ct system. *Journal of Nuclear Medicine*, 60(7):1031–1036, 2019.

[30] Stefaan Vandenberghe, Pawel Moskal, and Joel S Karp. State of the art in total body pet. *EJNMMI Phys, 7(1), 35*, 2020.

[31] Expert Verilog and Clifford Cummings. Simulation and synthesis techniques for asynchronous fifo design. 01 2002.

[32] Jiawei Xia. *Interaction Reconstruction in Digital 3-D CdZnTe Under Various Circumstances*. PhD thesis, University of Michigan, 2019.

[33] Yuefeng Zhu. *Digital Signal Processing Methods for Pixelated 3-D Position Sensitive Room-Temperature Semiconductor Detectors*. PhD thesis, University of Michigan, 2012.

[34] Yuefeng Zhu, Stephen E. Anderson, and Zhong He. Sub-pixel position sensing for pixelated, 3-d position sensitive, wide band-gap, semiconductor, gamma-ray detectors. *IEEE Transactions on Nuclear Science*, 58(3):1400–1409, 2011.