TETRIS: A Streaming Accelerator for Physics-Limited 3D Plane-Wave Ultrasound Imaging

Brendan L. West
University of Michigan
westbl@umich.edu

Jian Zhou
Arizona State University
jzhou50@asu.edu

Ronald G. Dreslinski
University of Michigan
rdreslin@umich.edu

J. Brian Fowlkes
University of Michigan
fowlkes@umich.edu

Oliver Kripfgans
University of Michigan
greentom@umich.edu

Chaitali Chakrabarti
Arizona State University
chaitali@asu.edu

Thomas F. Wenisch
University of Michigan
twenisch@umich.edu

ABSTRACT

High volume acquisition rates are imperative for medical ultrasound imaging applications, such as 3D elastography and 3D vector flow imaging. Unfortunately, despite recent algorithmic improvements, high-volume-rate imaging remains computationally infeasible on known platforms.

In this paper, we propose TETRIS, a novel hardware accelerator for ultrasound beamforming that enables volume acquisition rates up to the physics limits of acoustic propagation delay. Through algorithmic and hardware optimizations, we enable a streaming system design outclassing previously proposed accelerators in performance while lowering hardware complexity and storage requirements. For a representative imaging task, our proposed system generates physics-limited 13,020 volumes per second in a 2.5W power budget.

CCS CONCEPTS

• **Hardware** → **Emerging architectures**; *3D integrated circuits*;

KEYWORDS

Ultrasound, Beamforming, Accelerators, 3D Architectures

ACM Reference Format:

Brendan L. West, Jian Zhou, Ronald G. Dreslinski, J. Brian Fowlkes, Oliver Kripfgans, Chaitali Chakrabarti, and Thomas F. Wenisch. 2019. TETRIS: A Streaming Accelerator for Physics-Limited 3D Plane-Wave Ultrasound Imaging. In *Proceedings of The 56th Annual Design Automation Conference 2019 (DAC '19)*. ACM, New York, NY, USA, Article 4, 6 pages. <https://doi.org/10.1145/3316781.3317921>

1 INTRODUCTION

3D ultrasound is an imaging modality that has seen increased use in medical applications due to its low risk potential compared to X-ray and MRI [17]. Emerging 3D ultrasound applications in cardiac imaging [14], vector flow imaging [7, 23], and shear wave elastography [19] require extremely high volumetric image acquisition

rates to track high frequency motion. These applications also require large imaging apertures to capture sufficient resolution in a large region of interest. The combination of a large receive aperture (32×32 elements), large imaging volume ($32 \times 32 \times \sim 1680$ focal points), substantial precision requirement (10-12 bit ADCs), and enormous volume rate (up to 13,020 volumes / sec) pose a daunting computational requirement—straight-forward delay-and-sum beamforming is intractable even with a rack full of GPUs. The transducers' raw data rate is so high (~ 500 gigabits per second) that beamforming must be tightly coupled to the transducers. But, this tight coupling implies that beamforming must occur in the ultrasound probe handle, in close contact with the patient's skin. If the beamforming system dissipates more than $\sim 5W$, it could burn the patient [16].

Because beamforming (identifying the samples within each receive signal that correspond to each focal point) is the most computationally expensive aspect of 3D ultrasound [16], substantial work has been done on algorithms and computational architectures to make it tractable [24, 27]. Straight-forward time-delay beamforming calculates a time delay for the propagation path from each transmit element to each focal point to each receive element. For a target volume of size $M_x M_y M_z$ focal points and a receive array of $N_x N_y$ transducers, the number of required delays, and the associated computational complexity, is $M_x M_y M_z N_x N_y$. To reduce computational complexity, researchers have proposed separable beamforming algorithms [24–26], which approximate traditional beamforming by splitting it into two stages, beamforming consecutively along the two lateral image axes. The two-stage process reduces the 2D grid of contributing transducers for each voxel (multiplicative complexity) to two 1D arrays of contributing transducers (additive complexity), while sacrificing little image clarity.

Despite such algorithmic advances, however, naïve time-delay beamforming remains impractical, as delay calculations involve complex trigonometry. In 2D imaging, delays can be trivially stored in lookup tables [12], but the scale of 3D imaging has typically precluded such precomputation due to the vast number of delays required. Prior works [2, 3, 5, 6, 16, 18] have instead used specialized hardware to approximate time delays on the fly; but, all of these systems fall orders of magnitude short in achieving the 1,000+ volumes per second target for cardiac, vector flow imaging, and shear wave elastography applications.

Plane-wave ultrasound is a promising approach for achieving high volume rates, but can lack resolution and result in low signal-to-noise ratio (SNR) due to lack of transmit focusing [24]. Prior work [26] proposes a separable 3D plane-wave image compounding

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DAC '19, June 2–6, 2019, Las Vegas, NV, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6725-7/19/06...\$15.00

<https://doi.org/10.1145/3316781.3317921>

Figure 1: 3D plane-wave transmit and receive setup; overlapping beamforming apertures “step” across the receive aperture, each creating a single scanline.

scheme, which increases resolution and SNR by summing over multiple firing angles. The need to compound over multiple transmits further exacerbates the need for hardware to support extremely high volume acquisition rates, motivating our goal to design a hardware accelerator that can reach acoustic limits (i.e., volume acquisitions limited by physics rather than computation). However, the prior algorithm must store and iterate over the digitized transducer channel data, imposing high on-chip storage requirements ($\sim 800\text{KB}$ per cm of imaging depth). Any design that stores raw channel data on chip will suffer an inherent limit on imaging depth.

In this paper, we describe TETRIS, a new digital beamforming accelerator for high-volume-rate 3D plane-wave imaging, enabled by two key elements. First, TETRIS implements *delay compression*, a novel delay decomposition scheme that builds upon prior separable beamforming [24, 26] to exploit additional symmetries that arise in plane-wave imaging. Delay compression makes it feasible to store precomputed delays on-chip or stream them as needed by utilizing a specialized bit vector encoding, eliminating the most computationally intensive aspect of prior beamforming accelerators. Second, TETRIS incorporates a novel hardware functional unit, the TETRIS RESERVATION STATION (RS), which aligns time-delayed samples across receive channels for the beamsum operation. We name this unit after the popular 80s-era video game because its operation is conceptually similar to the game—as sample values “fall to the bottom” of a conceptual 2D grid, a delay-aligned horizontal row is “cleared” and sent for beamsum when the row is fully populated. The TETRIS RS removes the need for on-chip storage of the receive signal, enabling arbitrary imaging depth and facilitating volume rates that reach physics limits. In summary, we contribute:

- *Delay compression*, a two-part delay decomposition for separable 3D plane-wave ultrasound, which results in up to a $1024\times$ reduction in the number of delays computed for both the first and second beamforming stages.
- A *bit vector encoding* for compressed delays, which makes precomputed delays feasible.
- The TETRIS RS, a novel functional unit for single-pass, depth-agnostic, physics-limited volume reconstruction.
- TETRIS—a streaming plane-wave beamforming accelerator capable of physics-limited volume rates.

2 BACKGROUND

Plane-wave ultrasound imaging uses a flat-plane (unfocused) transmission scheme. A large 2D transducer array samples the reflected signals, with each receive *channel* digitizing the signal from one

transducer. The large receive array is divided into smaller, overlapping 2D sub-arrays called *beamforming apertures* (see Figure 1). The channels within each beamforming aperture contribute to a single *scanline*—an axial column in the final volumetric image. The receive array comprises $M_x M_y$ transducers, while the beamforming apertures include $N_x N_y$ transducers; with a depth of M_z , the resulting image is $M_x M_y M_z$ voxels.

Managing delay constants poses a significant challenge in many 3D ultrasound algorithms, and plane-wave is no different. For separable plane-wave imaging with angled compounding [26], each firing angle requires a combined $M_x M_z N_x N_y + M_x M_y M_z N_y$ delay constants for the two beamforming stages. The exact delay calculations for each focal point and firing angle are given in Equations (1) and (2). In these equations, (m_x, m_y, m_z) are the coordinates of each focal point, (n_x, n_y) are the coordinates of each receive transducer within the beamforming aperture. The plane’s firing angle is defined by α —the angle between the z axis and the normal vector, β —the angle between the x axis and the projection of the normal vector on the xy plane, and (x_0, y_0, z_0) —an arbitrary point on the plane at its origin. In this formulation, there is little symmetry between beamforming apertures due to the varying angle of the transmit plane. The transmit distance d_{tx} represents how far the plane must propagate to reach a point within the target volume, while d_{rx} represents the distance the reflection must travel to reach the transducers within the associated beamforming aperture.

$$d_{tx} = (m_x - x_0) \sin \alpha \cos \beta + (m_y - y_0) \sin \alpha \sin \beta + (m_z - z_0) \cos \alpha \quad (1)$$

$$d_{rx} = \sqrt{(m_x - n_x)^2 + (m_y - n_y)^2 + m_z^2} \quad (2)$$

The high number of required delay constants is one of the difficulties of 3D ultrasound in general, as it precludes combinations of large beamforming aperture, receive volume size, imaging depth, and volume formation rate. On-chip delay storage, or streaming them from off-chip storage, is prohibitive due to their sheer quantity, and the complexity of the trigonometric equations makes calculating them on the fly impractical due to power constraints.

3 DELAY COMPRESSION

Prior works [16, 24, 25] attempt to reduce on-chip delay storage by approximating the delays using a piece-wise function and an associated on-chip delay calculation unit for each channel. However, this approach overlooks an opportunity for delay sharing between beamforming apertures in plane-wave imaging. For example, in the approach of Yang et al. [26], some components of the distance calculation do not actually depend on the original position or angle of the transmitted plane-wave. In fact, the reflection distance equation is identical for every beamforming aperture, and the transmit distance differs only up to the point where the plane-wave touches the first focal point along each scanline (where m_z is 0); after this point, the calculation is identical for each scanline in the volume. This decomposition is visualized in Figure 2. The initial distance d_{tx1} is calculated by setting m_z to 0 in Equation 1, giving us the distance from the plane’s origin to the first focal point in each scanline. We then calculate d_{tx2} by setting x_0, y_0 , and z_0 equal to m_x, m_y , and 0 respectively, eliminating terms that depend on x and y :

$$d_{tx1} = (m_x - x_0) \sin \alpha \cos \beta + (m_y - y_0) \sin \alpha \sin \beta - z_0 \cos \alpha \quad (3)$$

Figure 2: X-Z slice of the image space showing a 2D component of delay compression. Unique d_{tx1} distances are shown in d_1 and d_2 , while a , b , and c represent the shared d_{tx2} distances from the plane to focal points m_z along each scanline.

Table 1: Parameters of the target 3D plane-wave system.

Property	Value
Speed of sound (tissue), m/s	1540
Pitch, μm	192.5
Transmit aperture size, transducers	128×96
Receive aperture size, transducers	32×32
Beamforming aperture size, transducers	32
Number of scanlines per stage	32×32
Stage 1 scanline output length (M_z), points	2,089
Stage 2 scanline output length (M_z), points	1,679
Image depth, cm	4
Center frequency, MHz	4
6 dB transducer bandwidth, MHz	2
ADC sampling rate, MHz	40

$$d_{tx2} = m_z \cos \alpha \quad (4)$$

By breaking the transmit distance into two components, d_{tx1} and d_{tx2} , we gain additional symmetry between beamforming apertures. Distance d_{tx1} measures from the plane’s origin to the “base” of each scanline (where m_z is 0), while d_{tx2} measures from that point to each focal point along the line. This symmetry allows us to share a single d_{tx2} calculation across all beamforming apertures.

Whereas delay compression can be applied to both separable and non-separable plane-wave beamforming algorithms, we focus on the separable variant, as it significantly reduces the computational—and therefore the power—requirements of a beamforming hardware accelerator. For the separable algorithm, delay compression reduces the number of delays to $M_x M_y + M_z N_x$ and $M_x M_y N_x + M_z N_y$ for the first and second stages, respectively. For the imaging parameters we evaluate (see Table 1), delay compression yields a savings of $1008\times$ and $636\times$, respectively, for the first and second stages. The significant reduction makes it feasible to store the delay constants on-chip or stream them from memory without exorbitant bandwidth requirements, which in turn enables new approaches to beamforming hardware design.

4 TETRIS HARDWARE DESIGN

We next present TETRIS, our 3D separable plane-wave beamforming accelerator, which uses delay compression and the TETRIS RESERVATION STATION (RS) to enable single-pass, depth-agnostic, physics-limited volume reconstruction in a fully streaming architecture.

Figure 3: Scaled-down depiction of TETRIS with one slice of eight pipelines (a-h). New samples arrive at a fixed rate, while the core of the pipeline is clocked $8\times$ faster, allowing samples to rotate across all pipelines via the barrel shift stage. Samples matching each focal point are selected according to the pre-computed bit-vector (a) in the select stage, and are then passed to the TETRIS RS (b) for temporary storage. When a row fills, it is erased and sent for apodization and accumulation modules to compute a single voxel.

Digital time-delay beamforming generally comprises five steps: (1) digitizing the input channels, (2) up-sampling received signals via interpolation, (3) selecting—through calculation of the round-trip delays—the samples corresponding to each voxel, (4) apodizing to improve focus and suppress side lobes, and (5) summing samples across channels to obtain voxel intensity. Time-delay beamforming is amenable to massive parallelization across both channels and voxels. Nevertheless, the load-store interface of conventional data parallel architectures, such as GPUs, is a poor match for beamforming. If channel data are mapped to memory, each volume requires more than 100M address calculation and load operations, rendering volume rates in excess of 1,000 per second impossible.

Prior 3D beamforming hardware [17] instead abstracts each incoming channel as a stream and selects samples that correspond to consecutive focal points along a scanline as the samples arrive from memory, without explicit loads or address calculation. Nevertheless, prior designs required large functional units to calculate the time-delay delta between consecutive focal points along a scanline. These delay estimation units dominate the hardware and limit performance to at most a few tens of volumes per second [25].

TETRIS achieves high volume rates within the tight power constraints of a hand-held system by exploiting delay compression to enable delay precomputation, eliminating the need for on-the-fly approximation. TETRIS comprises 1024 pipelines broken into 32 slices of 32 pipelines each, wherein each slice processes one row (or column) of transducer channels and each pipeline processes a single scanline; a simplified diagram depicting one slice with eight pipelines (labeled **a** through **h**) is shown in Figure 3. TETRIS exploits

that, under separable beamforming, only the receive channels in the same row and slice contribute to a corresponding row of scanlines, while channels in different rows/slices do not interact.

ADC Sampling, Interpolation, and Data Sharing. In each slice, incoming 12-bit samples arrive in each channel at 40MHz from the analog front-end (top of figure) and are up-sampled via $4\times$ linear interpolation to create 4-wide sample vectors. Samples from each vector potentially contribute to focal points in multiple scanlines, so all 32 of these 4-sample vectors must be processed by all 32 pipelines. The “barrel shift” stage rotates the sample vectors laterally across pipelines. This and subsequent stages are clocked $32\times$ faster (1.28GHz) than the ADC sampling rate (40MHz), so each sample vector visits each pipeline once. The figure shows an example state after six rotations, where vector **a** is in pipeline **c**.

Sample Selection using Bit Vectors. Each cycle, the “select” stage in each pipeline determines if any of the samples in the current vector contributes to focal points in its scanline. Delays are computed offline, with the d_{tx1} offsets being pre-loaded into each pipeline’s select stage. Shared delays (the sum of d_{tx2} and d_{rx}) are stored in a bit vector, where a one indicates that a sample corresponds to a focal point, while a zero indicates the sample should be discarded (i.e., it falls between focal points). The bit vector is stored off chip and is broadcast to each pipeline during the beamforming process. As incoming samples are presented to the select stage, the d_{tx1} offsets are decremented; all samples are discarded while the offset is non-zero. Once the offsets reach zero, the selection process begins using the shared bit vector. Figure 3(a) depicts this selection process for the **f** vector in pipeline **h**; bits that are set in the delay vector **s** indicate which corresponding samples should be selected.

TETRIS RESERVATION STATION (RS). Selected samples then flow into the TETRIS RS in each pipeline; the one for pipeline **h** is shown in Figure 3(b). The samples for each voxel still arrive offset in time, so a mechanism is needed to store them until summation can occur. As anywhere from zero to four samples may be selected each cycle (depending on depth), the temporary storage must allow for a variable number of writes per cycle without wasting space or incurring write port conflicts. The TETRIS RS is a 4-way banked register file logically arranged as a 2D grid, where columns correspond to input channels and rows are split round-robin across banks. There is a single write port that writes to up to four consecutive entries (i.e., across the four banks) in a single column each cycle, and a single read port that reads a single entry from the 2D array.

Conceptually, the unit acts as a 2D circular buffer. Arriving samples “fall down” the column corresponding to their receive channel to fill in the lowest-indexed empty cells (much like falling Tetris blocks). A tail pointer indicating the index for the next write is maintained for each column and a single column is written each cycle (up to 4 entries). Each row holds samples corresponding to a specific voxel, with a single head pointer that points to the current row (or voxel) being processed. As the bottom row of the array fills, samples are read out, one per cycle, and sent to the apodization unit. Once the entire row has been read, the head pointer is incremented, “dropping” the completed row (again, much like the game Tetris).

Apodization and Accumulation. As samples are read from the TETRIS RS, they are apodized to reduce side lobes and improve focus. Apodization involves multiplying each sample with a channel-specific weight. Apodized samples then pass to the accumulation

unit, which sums samples from a particular TETRIS RS row. Once an entire row has been accumulated, the final value is sent to either the second beamforming stage or to memory. Our approach maximizes data reuse and minimizes apodization and accumulation operations, which account for a large portion of the system’s power.

System Integration. We next describe how the TETRIS beamforming accelerator can be integrated into a complete hand-held 3D ultrasound system. Like prior designs [16, 26], we rely on 3D die stacking to enable integration of analog, digital, and memory components with sufficient interconnect bandwidth. The complete beamforming system requires a 4- or 5-layer die-stacked structure comprising a layer of 128×96 capacitive micromachined ultrasonic transducers (CMUTs) on the face of the ultrasound probe, with the ADCs and their associated amplifiers, one or two TETRIS beamforming accelerators, and a 12.5MB eDRAM layer that acts as a buffer for completed volumes and facilitates compounding.

Separable algorithms perform beamforming along the X- and Y-dimensions in sequence. TETRIS can be deployed in two different ways to address this requirement. In the first approach, the die stack includes a single TETRIS accelerator, which processes each volume twice, with intermediate partially beamformed results buffered in the eDRAM array. However, this approach halves the attainable volume acquisition rate. In the second approach, the system includes two TETRIS accelerators, rotated 90° with respect to each other to accommodate beamforming in both lateral dimensions. While consuming more power, this design can achieve a physics-limited volume rate while still meeting the target power budget.

5 METHODOLOGY & RESULTS

We first demonstrate why a hardware accelerator is needed for plane-wave ultrasound by estimating beamforming performance on existing computational platforms. We then validate the TETRIS design in two parts. First, we evaluate the power and area requirements of a TETRIS system in an industrial 16nm node, showing that we meet the 5W power budget deemed safe for contact with human skin. Second, we confirm the image quality of the hardware design against ideal beamforming using Field II [8–10]. Field II is a widely-used Matlab extension for modeling ultrasound physics.

5.1 Why a[nother] Specialized Accelerator?

To demonstrate why new approaches are needed to achieve physics—rather than computation—limited volume acquisition rates for 3D ultrasound, we briefly consider what it would take to perform beamforming on conventional computational platforms. The overall performance of the conventional platforms and the TETRIS system are compared in Table 3.

DSP. The TI C6678 is a Digital Signal Processor (DSP) that is specifically marketed for medical imaging applications [21]. A study performed with the C6678 found that a single DSP can output a 4096-point scanline for a 512-transducer array in $740\mu\text{s}$ when using a state-of-the-art algorithm and a 1GHz clock frequency [11]. We scale this result to estimate the time per scanline for our separable plane-wave algorithm. We estimate the C6678 requires $23.58\mu\text{s}$ for the first stage and $18.96\mu\text{s}$ for the second stage, resulting in a total of $42.54\mu\text{s}$ per scanline for our target system and imaging parameters (see Table 1). Using this back-of-the-envelope calculation, we

Table 2: 3D Plane-Wave System Power. Each column represents a different layer in the 3D stack.

	CMUTs (128 × 96)	ADCs (32 × 32)	TETRIS #1	TETRIS #2	eDRAM (12.5MB)	Total Power (W)
Power (mW)	300	143	782	782	325	2.3
Area (mm ²)	432	4.8	11.9	11.9	4.5	

Table 3: Hardware Performance - Volumes in 1 Second.

Bf. Ap. Size:	16	20	25	32
TETRIS (Full System)	13,020	13,020	13,020	13,020
Nvidia Titan XP (doubles)	120.8	103.3	91.5	75.1
Nvidia Titan XP (floats)	123.6	108.2	96.7	81.1
TI C6678 DSP [11]	45.9	36.7	29.4	23.0

estimate the C6678 will produce a single volume in 43.56ms. To achieve 13,020 volumes per second—where the round-trip time of the transmitted signal is the limiting factor—it would require the use of 568 DSPs, with a staggering 5.79kW power consumption [20]. The need for explicit memory addressing and load instructions makes a DSP-based solution untenable for such volume rates.

GPU. To evaluate performance on high-end Graphics Processing Units (GPUs), we implement 3D beamforming on an Nvidia Titan XP. We neglect data transfer and consider only computation. We implement our separable plane-wave algorithm with delay compression in CUDA to match the algorithm implemented by TETRIS as closely as possible, storing the pre-computed delays in an in-memory lookup table. To maximize efficiency, we use four separate GPU kernels: stage 1 interpolation, stage 1 beamforming, stage 2 interpolation, and stage 2 beamforming. In the interpolation kernels, each GPU thread performs 4x interpolation for a single sample, yielding $32 \times 32 \times 3077$ threads in the first stage and $32 \times 32 \times 2089$ threads in the second stage. The beamforming kernels compute a single voxel per thread, resulting in $32 \times 32 \times 2089$ threads in the first stage and $32 \times 32 \times 1679$ threads in the second stage.

The Titan XP calculates a single volume in 12.34ms using single-precision floating-point, yielding 81 volumes per second. However, to match the physics-limited volume rate of 13,020 per second still requires 161 Titan XPs for a total power consumption of nearly 34kW. TETRIS is drastically more efficient than the GPU because its streaming architecture does not require explicit address calculation and load instructions to access channel data and its datapath operates at 12-bit fixed-point precision.

Whereas there has been prior work on hardware-accelerator ultrasound beamforming [2, 5, 6, 12, 16, 24], none of the prior accelerators target volume acquisition rates within the same order of magnitude as TETRIS. We discuss the most related hardware acceleration work in Section 6.

5.2 Full-System Power

We estimate full-system power using a combination of hardware synthesis and previously published results. The area and power for each layer is shown in Table 2. Our design uses the same 128×96 transducer array as [16], with elements spaced exactly $\lambda/2$ apart, where λ is the transmit frequency (4MHz). While our receive aperture is only 32×32 elements, the larger transmit array avoids roll-off at the edges of the imaging volume when transmitting angled

Table 4: Anechoic Cyst CNR — 9-angle Compounding.

Implementation	Cyst #1	Cyst #2	Cyst #3
Ideal (doubles)	2.962	3.026	2.365
12-bit Fixed Point	2.936	3.038	2.354

planes. We estimate ADC power using a recently published implementation in 28nm technology [13], scaling the power linearly as we require only a 40MHz sample rate rather than the published 100MHz. We implement TETRIS in SystemVerilog and synthesize it using an industrial 16nm standard cell library and SRAM compiler. We report power for a 1.28GHz operating frequency, which is fast enough to achieve physics-limited volume rates for the full 32×32 receive aperture at any imaging depth; smaller apertures can be processed at lower frequencies (and lower power). We estimate the area and power requirements of the 12.5MB eDRAM based on reported results for a 22nm technology [4].

5.3 Image Quality

We verify TETRIS image quality using visual comparison of 2D volume slices and the Contrast-to-Noise Ratio (CNR) of simulated cyst phantoms. CNR is a standard metric for comparing ultrasound image quality [1, 15, 22] where the contrast resolution of the echoic (tissue) and anechoic (cyst) regions are compared under the effects of clutter and receiver noise (outside interference that produces image artifacts). The purpose of this validation is to confirm that TETRIS’s 12-bit data path does not degrade image quality relative to double-precision floating-point calculations.

Using Field II, we simulate a phantom with cysts at 13mm, 23mm, and 33mm, respectively. As shown in Table 4 and Figure 4, the fixed-point simulation produces images and CNR values indistinguishable from those using double-precision floating-point. These results demonstrate that the image quality degradation using our delay compression technique and hardware implementation is negligible.

6 RELATED WORK

Researchers have proposed a variety of algorithms [24, 26] and accelerator architectures [2, 5, 6, 16] to reduce the computational burden of the delay calculation. Hager et al. [2] designed a beamformer that fully samples a 100×100 element array and approximates the delay values, achieving 15 volumes per second. Sonic Millip3De [16] multiplexes their 128×96 array, sampling 1024 elements per firing, approximating the delay values using a piece-wise function and producing up to 32 volumes per second [25]. An approach [5, 6] that uses a look-up table and approximates the delay calculations to reduce complexity was proposed for another 100×100 array, and is estimated to achieve a rate of 15-30 volumes per second using a state-of-the-art FPGA. Whereas these systems demonstrate powerful beamforming capabilities at a variety of resolutions, they all

(a) Ideal (doubles)

(b) Fixed-Point (12-bit)

Figure 4: X-Z slices of cyst phantoms using separable plane-wave beamforming with delay compression, 9-angle coherent image compounding, and beamforming aperture size of 32×32 .

target medical applications with much lower volume rate requirements than TETRIS and therefore make architectural decisions that preclude the 1,000+ volumes per second target for cardiac, vector flow imaging, and shear wave elastography applications.

7 CONCLUSION

In this work, we have described TETRIS—the first beamforming accelerator for separable 3D plane-wave ultrasound capable of achieving physics-limited volume rates. TETRIS’s performance is enabled through a novel delay decomposition, paired with the TETRIS RS—a 2D register structure that sorts the incoming samples, avoiding the need to store the received signal and allowing for a streaming design. Through synthesis of our RTL-level design in a 16nm standard cell library, we show that TETRIS achieves physics-limited volume rates with a power consumption of only $\sim 780\text{mW}$. Furthermore, our system integration estimates indicate that TETRIS can be combined with the required supporting layers to form a full 3D ultrasound system, while remaining well within the tight 5W power constraint for safe contact with human skin.

ACKNOWLEDGMENTS

This work was supported in part by NSF CCF-1406739 and the Applications Driving Architectures (ADA) Research Center, a JUMP Center co-sponsored by SRC and DARPA.

REFERENCES

- [1] J. Dahl, G. Trahey, and G. Pinton. The effects of image degradation on ultrasound-guided hifu. In *Proc. of IEEE International Ultrasonics Symp.*, pages 809–812, Oct. 2010.
- [2] P. A. Hager, A. Bartolini, and L. Benini. Ekho: A 30.3w, 10k-channel fully digital integrated 3-d beamformer for medical ultrasound imaging achieving 298m focal points per second. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(5):1936–1949, May 2016.
- [3] P. A. Hager, P. Vogel, A. Bartolini, and L. Benini. Assessing the area/power/performance tradeoffs for an integrated fully-digital, large-scale 3d-ultrasound beamformer. In *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, pages 228–231, Oct 2014.
- [4] F. Hamzaoglu, U. Arslan, N. Bisnik, S. Ghosh, M. B. Lal, N. Lindert, M. Meterelliyozy, R. B. Osborne, J. Park, S. Tomishima, Y. Wang, and K. Zhang. A 1 gb 2 ghz 128 gb/s bandwidth embedded dram in 22 nm tri-gate cmos technology. *IEEE Journal of Solid-State Circuits*, 50(1):150–157, Jan 2015.
- [5] A. Ibrahim, P. Hager, A. Bartolini, F. Angiolini, M. Ardit, L. Benini, and G. D. Micheli. Tackling the bottleneck of delay tables in 3d ultrasound imaging. In *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1683–1688, March 2015.
- [6] A. Ibrahim, P. A. Hager, A. Bartolini, F. Angiolini, M. Ardit, J. Thiran, L. Benini, and G. D. Micheli. Efficient sample delay calculation for 2-d and 3-d ultrasound imaging. *IEEE Transactions on Biomedical Circuits and Systems*, 11(4):815–831, Aug 2017.
- [7] J. Jensen, S. Nikolov, A. Yu, and D. Garcia. Ultrasound vector flow imaging—part ii: Parallel systems. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 63(11):1722–1732, Nov 2016.
- [8] J. A. Jensen. FIELD: A program for simulating ultrasound systems. *10th Nordicbaltic Conference on Biomedical Imaging, Vol. 4, Supplement 1, Part 1:351–353*, pages 351–353, 1996.
- [9] J. A. Jensen. Simulation of advanced ultrasound systems using field ii. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 636–639 Vol. 1, April 2004.
- [10] J. A. Jensen and N. B. Svendsen. Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers. *IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control*, 39(2):262–267, Mar. 1992.
- [11] K. Karadayi, C. Lee, and Y. Kim. Software-based ultrasound beamforming on multi-core dsps. Technical report, University of Washington, March 2011. <http://www.ti.com/lit/wp/sprabo0/sprabo0.pdf>.
- [12] B. Lam, M. Price, and A. P. Chandrakasan. An asic for energy-scalable, low-power digital ultrasound beamforming. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, pages 57–62, Oct 2016.
- [13] C. Liu, M. Huang, and Y. Tu. A 12 bit 100 ms/s sar-assisted digital-slope adc. *IEEE Journal of Solid-State Circuits*, 51(12):2941–2950, Dec 2016.
- [14] G. Matrone, A. S. Savoia, G. Caliano, and G. Magenes. Ultrasound plane-wave imaging with delay multiply and sum beamforming and coherent compounding. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3223–3226, Aug 2016.
- [15] M. S. Patterson and F. S. Foster. The improvement and quantitative assessment of b-mode images produced by an annular array/cone hybrid. *Ultrasonic Imaging*, 5(3):195–213, 1983. PMID: 6356553.
- [16] R. Sampson, M. Yang, S. Wei, C. Chakrabarti, and T. F. Wenisch. Sonic millip3de: A massively parallel 3d-stacked accelerator for 3d ultrasound. In *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, pages 318–329, Feb 2013.
- [17] R. Sampson, M. Yang, S. Wei, C. Chakrabarti, and T. F. Wenisch. Sonic millip3de with dynamic receive focusing and apodization optimization. In *2013 IEEE International Ultrasonics Symposium (IUS)*, pages 557–560, July 2013.
- [18] W. Simon, A. C. Yüzügüler, A. Ibrahim, F. Angiolini, M. Ardit, J.-P. Thiran, and G. D. Micheli. Single-fpga, scalable, low-power, and high-quality 3d ultrasound beamformer. In *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–2, Aug 2016.
- [19] M. Tanter and M. Fink. Ultrafast imaging in biomedical ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 61(1):102–119, January 2014.
- [20] Texas Instruments. C6678 power consumption model (rev. d), Feb. 2014. <http://www.ti.com/lit/zip/sprm545>.
- [21] Texas Instruments. Tms320c6678 datasheet (rev. e), Mar. 2014. <http://www.ti.com/lit/ds/symlink/tms320c6678.pdf>.
- [22] K. F. Üstüner and G. L. Holley. Ultrasound imaging system performance assessment. American Association of Physicists in Medicine Annu. Meeting, 2003.
- [23] S. Wei, M. Yang, J. Zhou, R. Sampson, O. D. Kripfgans, J. B. Fowlkes, T. F. Wenisch, and C. Chakrabarti. Low-cost 3-d flow estimation of blood with clutter. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 64(5):772–784, May 2017.
- [24] M. Yang, R. Sampson, S. Wei, T. F. Wenisch, and C. Chakrabarti. High frame rate 3-d ultrasound imaging using separable beamforming. *J. Signal Process. Syst.*, 78(1):73–84, Jan. 2015.
- [25] M. Yang, R. Sampson, S. Wei, T. F. Wenisch, and C. Chakrabarti. Separable beamforming for 3-d medical ultrasound imaging. *IEEE Transactions on Signal Processing*, 63(2):279–290, Jan 2015.
- [26] M. Yang, R. Sampson, S. Wei, T. F. Wenisch, B. Fowlkes, O. Kripfgans, and C. Chakrabarti. High volume rate, high resolution 3d plane wave imaging. In *2014 IEEE International Ultrasonics Symposium*, pages 1253–1256, Sept 2014.
- [27] M. Yang, R. Sampson, T. F. Wenisch, and C. Chakrabarti. Separable beamforming for 3-d synthetic aperture ultrasound imaging. In *SiPS 2013 Proceedings*, pages 207–212, Oct 2013.