Sonic Millip3De: An Architecture for Handheld 3D Ultrasound

Sonic Millip3De, a system architecture and accelerator for 3D ultrasound beamforming, has a three-layer die-stacked design that combines a hardware-friendly approach to the ultrasound imaging algorithm with a custom beamforming accelerator. The system achieves high-quality 3D ultrasound imaging within a full-system power of 15 W in 45-nm semiconductor technology. Sonic Millip3De is projected to achieve the target 5-W power budget by the 16-nm technology node.

Richard Sampson
University of Michigan

Ming Yang
Siyuan Wei
Chaitali Chakrabarti
Arizona State University

Thomas F. Wenisch
University of Michigan

Much as every medical professional listens beneath the skin with a stethoscope today, we foresee a time when handheld medical imaging will become as ubiquitous—"peering under the skin" using a handheld imaging device. Mobile medical imaging is advancing rapidly to reduce the footprint of bulky, often room-sized machines to compact handheld devices. In the last five years, research has demonstrated that by combining the increasing capabilities of mobile processors with intelligent system design, portable and even handheld imaging devices are not only possible, but commercially viable. In particular, ultrasound imaging has proven to be an especially successful candidate for high portability due to its safety and low transmit power, with commercial handheld 2D ultrasound devices marketed and being used in hospitals today. Newly developed portable imaging devices have not only led to demonstrated improvements in patient health,[1] they have also enabled new applications for handheld ultrasound, such as disaster relief care[2] and battlefield triage.[3] However, despite the increasing capabilities of handheld ultrasound devices, these systems remain unable to produce the high-quality real-time 3D images that are possible with their nonhandheld counterparts.

In recent years, many hospitals have been transitioning to 3D ultrasound imaging when mobility is not required because it provides numerous benefits over 2D, including increased technician productivity, greater volumetric measurement accuracy, and more readily interpreted images. 3D imaging can also enable advanced diagnostic capabilities, such as tissue sonoelastography through high-velocity 3D motion tracking and accurate blood-flow measurements via 3D Doppler. Creating a handheld 3D system could enable hospital-quality ultrasound imaging in nearly any setting, greatly expanding the way ultrasound is used today. However, 3D ultrasound comes with many challenges that are compounded when implementing a system in a handheld form factor. The sheer amount of data that must be sensed, transferred, and computed is nearly 5,000 times more than in a 2D system. At the same time, the massive data rate (as high as 6 terabits/second) of the

Published by the IEEE Computer Society

0272-1732/14/$31.00 © 2014 IEEE

received echo signals is so high that the data cannot easily be transferred off chip for image formation; current 3D systems typically transfer data for only a fraction of receive channels, sacrificing image quality or aperture size. In addition to the extreme computational requirements, power is of the utmost importance, not only to ensure adequate battery life, but more importantly because the device is in direct contact with the patient's skin, placing tight constraints on safe operating temperature.

For safe operation, a handheld ultrasound system must operate within roughly a 5-W power budget. Implementing a handheld 3D system with commercially available digital signal processor (DSP) or graphics-accelerator chips using conventional beamforming algorithms designed for software is simply infeasible. Our analysis indicates that it would take 700 ultrasound DSP chips with a total power budget of 7.1 kW to meet typical 3D imaging computational demands at just 1 frame per second (fps). To enable such demanding computation on such a low power budget, a complete rethink of both the algorithm and architecture is required.

In this article, we introduce Sonic Millip3De,[4,5] a hardware accelerator that combines a new approach to the ultrasound imaging algorithm better suited to hardware with modern computer architecture techniques to achieve high-quality 3D ultrasound imaging within a full-system power of 15 W in 45-nm semiconductor technology. Under anticipated scaling trends, we project that Sonic Millip3De will achieve our target 5-W power budget by the 16-nm technology node.

We present this work both to make progress on realizing the promise of handheld medical imaging and as a case study for application-specific accelerator design. Our work also illustrates the unique benefits of 3D die stacking in heterogeneous systems and motivates moving beyond the limitations of the conventional von Neumann architecture in certain applications.

Synthetic aperture ultrasound

Synthetic aperture ultrasound imaging is performed by sending high-frequency pulses (typically 1 to 15 MHz) into a medium and constructing an image from the reflected pulse signals. The process comprises three stages: transmit, receive, and beamsum. Transmission and reception are both done using an array of transducers that are electrically stimulated to produce the outgoing signal and generate current when they vibrate from the returning echo. After all echo data is received, the beamsum process (the computation-intensive stage) combines the data into a partial image. The partial image corresponds to echoes from a single transmission. Several transmissions from different locations on the transducer array are needed to produce high-quality images, so several iterations of transmit, receive, and beamsum are typically necessary to construct a single complete frame.

Each transmission is a pulsed signal conceptually originating from a single location in the array, shown in Figure 1a. To improve signal strength, multiple transducers can fire together in a pattern to emulate a single virtual source located behind the transducer array.[6] The pulse expands into the medium radially, and as it encounters interfaces between materials of differing density, the signal is partially transmitted and partially reflected, as shown in Figure 1b. The returning echoes cause the transducers to vibrate, generating a current signal that is digitized and stored in a memory array associated with each transducer. Each position within these memory arrays corresponds to a different round-trip time from the emitting transducer to the receiving transducer. Because transducers cannot distinguish the direction of an incoming echo, each array element contains the superimposed echoes from all locations in the imaging volume with equal round-trip times (that is, an arc in the imaging volume). Because of the geometry of the problem, the round-trip arcs are different for each transducer, resulting in different superpositions at each receiver. The beamsum operation sums the echo intensity observed by all transducers for the arcs intersecting a particular focal point (that is, a location in the imaging volume), yielding a strong signal when the focal point lies on an echoic boundary. Combining transmissions from multiple source locations allows further focusing.

A typical beamsum pipeline first transforms the raw signal received from each

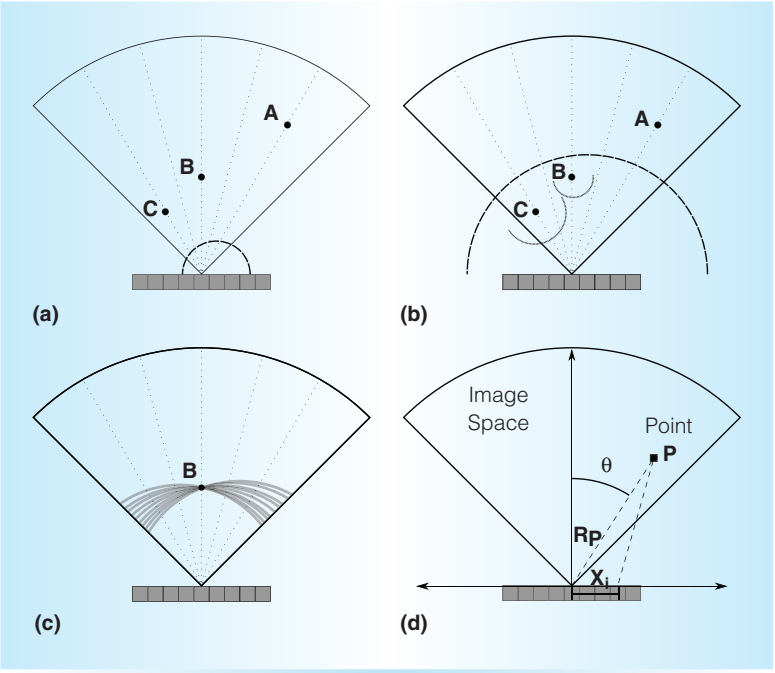

Figure 1. Synthetic aperture ultrasound. Pulse leaving transmit transducer (a). Echo pulses reflecting from points *B* and *C*. All transducers in array (or subaperture) will receive the echo data, but at different times due to different round trip distances (b). All of the reconstructed data for point *B* from each of the transducers added together. By adding thousands of "views" together, crisp points become focused (c). Variables used in calculating round-trip distance, d_p, for the *i*th transducer and point *P* in Equation 1 (d).

transducer to enhance signal quality. The signal is upsampled using an interpolation filter to generate additional data points between the received samples. This process enhances resolution without the power and storage overheads of increasing the data sampling rate of the analog front end. Then, so-called apodization scaling factors are applied to the interpolated data to place greater weight on receivers near the origin of the transmission, because these signals are more accurate owing to their lower angle of incidence.

Once the data has been preprocessed ("transformed"), the beamsum operation can begin. In essence, this entails calculating the round-trip delay between the emitting transducer and all receiving transducers through each focal point, converting these delays into indices in each transducer's received signal array, retrieving the corresponding data, and summing these values. Figure 1c illustrates this process. These partial images are then summed over multiple transmissions. Finally, a demodulation operation is applied to remove the ultrasound carrier signal.

The delay calculation (identifying the right index within each receive array) is the most computationally intensive aspect of beamsum, because it must be completed for every {focal point, transmit transducer, receive transducer} trio. Because the transmit signal propagates radially, the image space is described by a grid of scanlines that radiate at a constant angular increment from the center of the transducer array into the image volume. Focal points are located at even spacing along each scanline, in effect creating a spherical coordinate system. However, the transducers are laid out in a grid-based Cartesian coordinate system, requiring a fairly complex law of cosines calculation to compute round-trip distances via Equation 1:

$$d_P = \frac{1}{c}\left(R_p + \sqrt{R_p^2 + x_i^2 - 2x_i R_p \sin\theta}\right) \quad (1)$$

In this equation, d_p is the round-trip delay from the center transducer to point P to transducer i, c is the speed of sound in tissue (1,540 m/s), R_p is the radial distance of point P from the center of the transducer, θ is the angular distance of point P from the line normal to the center transducer, and x_i is the distance of transducer i from the center. Figure 1d shows variables as they correspond to the system geometry. This formula requires extensive evaluation of both trigonometric functions and square roots; hence, many 2D ultrasound systems precalculate all delays and store them in a lookup table (LUT).[7] However, a typical 3D system requires roughly 250 billion unique delay values, making a LUT implementation impractical. Instead, delays are calculated as needed.[8]

Redesigning the ultrasound algorithm for hardware acceleration

A key innovation of Sonic Millip3De is to codesign hardware with a new beamforming algorithm better suited to hardware acceleration. Our main algorithmic insight is to replace the expensive exact delay calculation of Equation 1 with an iterative, piecewise quadratic approximation, which can be

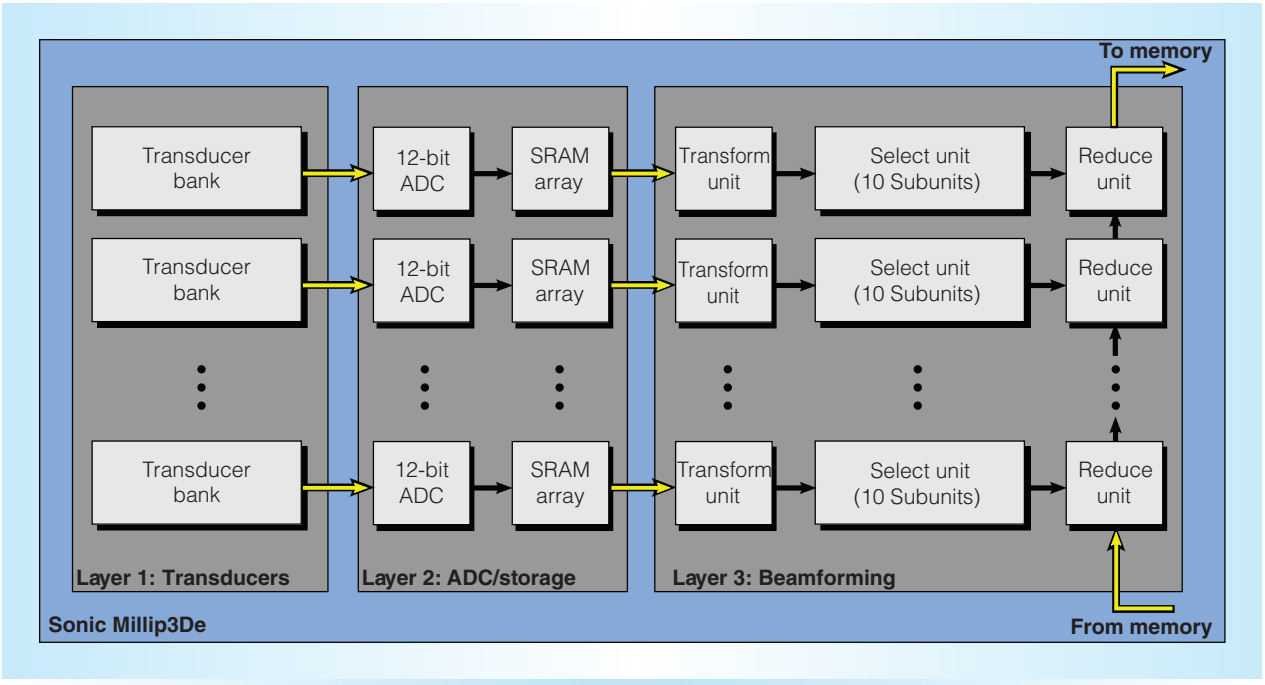

Figure 2. Sonic Millip3De hardware overview. Layer 1 comprises 120 × 88 transducers grouped into banks with one transducer per bank in each subaperture. Analog transducer outputs from each bank are multiplexed and routed over through-silicon vias (TSVs) to Layer 2, comprising 1,024 analog-to-digital converter (ADC) units operating at 40 MHz and static RAM (SRAM) arrays to store incoming samples. The stored data is passed via face-to-face links to Layer 3 for processing in the three stages of the 1,024-unit beamsum accelerator. The "transform" stage upsamples the signal to 160 MHz. The 10 units in the "select" stage map signal data from the receive time domain to the image space domain in parallel for 10 scanlines. The "reduce" stage combines previously stored data from memory with the incoming signal from all 1,024 beamsum nodes over a unidirectional pipelined interconnect, and the resulting updated image is written back to memory.

computed efficiently using only add operations. The algorithm's iterative nature lends itself to an efficient data streaming model, allowing the proposed hardware to exploit locality and eliminate inefficient address calculation and memory-access operations that are a bottleneck in conventional implementations. Our early analysis shows that the delta function between adjacent focal-point delays on a scanline forms a smooth curve and indices can be approximated accurately (with error similar to that introduced by interpolation) over short intervals with quadratic approximations. We replace these exact delta curves with a per-transducer precomputed piecewise quadratic approximation constrained to allow an index error of at most 3 (corresponding to at most a 30-μm error between the estimated and exact focal point), thus resulting in negligible blur.

Using offline image quality analysis, we have determined that, for a target imaging depth of 8 cm, we can meet the error constraints with only three piece-wise sections. Each section requires precalculating three coefficients and a section cut-off, achieving a 250-times storage reduction relative to an exhaustive lookup table. Through careful pipelining of the beamforming process, the constants can be efficiently streamed from off-chip memory, limiting storage requirements within the beamforming accelerator.

Sonic Millip3De

The Sonic Millip3De system hardware (shown in Figure 2) is divided into three distinct silicon die layers—transducers and analog components, analog-to-digital converters (ADCs) and storage, and beamforming computation—which are connected vertically using through-silicon vias (TSVs). The 3D-stacked chip connects to separate LPDDR2 (low-power double data rate 2) memory. All

of these components are integrated directly into the ultrasound scanhead, the wand-like device held against the patient's skin to obtain ultrasound images, allowing for a complete handheld device. These components comprise an ultrasound system's front end, capable of generating raw, volumetric images. A separate back end for viewing and postprocessing might be implemented in a tablet or PC.

Using a 3D die-stacked design provides several architectural benefits. First, it is possible to stack dies manufactured in different technologies. Hence, the transducer layer can be manufactured in a cost-effective process for the analog circuitry, higher voltages, and large geometry of ultrasonic transducers, while the beamforming accelerator can exploit the latest digital logic process technology. Second, stacking allows far more TSV links between dies than conventional chip pins, resolving the bandwidth bottleneck that plagues existing 3D systems where the probe and computation units are connected via cable. Third, TSV connections replace long wires that would otherwise be required in such a massively parallel system, reducing interconnect power requirements. Finally, stacking provides the potential for design modularity, where the same beamforming accelerator die could be stacked with alternative transducer arrays designed for different imaging applications.

The top die layer comprises a 120 × 88 grid of capacitive micromachined ultrasonic transducers (CMUTs) with $\lambda/2$ spacing. The area between the transducers is used for additional analog components and routing to the TSV interface. Our system uses a sliding "subaperture" technique where, for each transmit, only a 32 × 32 subgrid of transducers receive. The full 120 × 88 aperture is sampled over multiple transmissions, reducing hardware (and power) requirements at the cost of more transmissions per frame. The transducers are grouped into banks such that only one transducer per bank receives data in any subaperture. With this banking design, only a single beamforming channel is necessary for each of the 1,024 banks rather than each of 10,560 transducers.

The second layer comprises 1,024 (12-bit) ADCs and static RAM (SRAM) arrays, which each correspond to the 1,024 transducer banks of the analog transducer layer. The ADCs are sampled at 40 MHz, storing the digital output into corresponding 6-Kbyte SRAM arrays. The SRAMs are clocked at 1 GHz and connect vertically to a corresponding computational unit on the beamforming accelerator layer, requiring a total of 24,000 face-to-face TSVs for data and address signals.

The final layer is the most complex of the three, comprising the beamforming accelerator processing units, a unidirectional pipelined interconnect, and a control processor (M-class ARM core) that interfaces to the LPDDR2 off-chip memory.

Beamforming accelerator design

The beamforming accelerator comprises 1,024 independent channels, each divided into three conceptual stages: transform, select, and reduce. Each of these stages performs a separate operation to convert the digitized receive samples into beamformed focal-point intensities. Although the transform-select-reduce conceptual framework is particularly well suited for ultrasound beamforming, this design paradigm could also be applicable to other problems with similar dataflow.

Transform

The transform unit operates on all of the receive data, performing a 4-times linear interpolation on the raw receive signals. After upsampling, a constant apodization is applied providing a weight based on transducer position as previously described.

Select

The select unit remaps data from the receive time domain to the image space domain using the algorithm described previously. The select unit is split into 10 subunits that concurrently operate on neighboring scanlines. These subunits each iterate over the same incoming datastream from a corresponding second-layer SRAM array in a synchronized fashion, reducing the number of times data must be read from the SRAM by a factor of 10. Figure 3 shows a block diagram of a single subunit.

Data is streamed simultaneously into the input first-in, first-out (FIFO) buffer of each

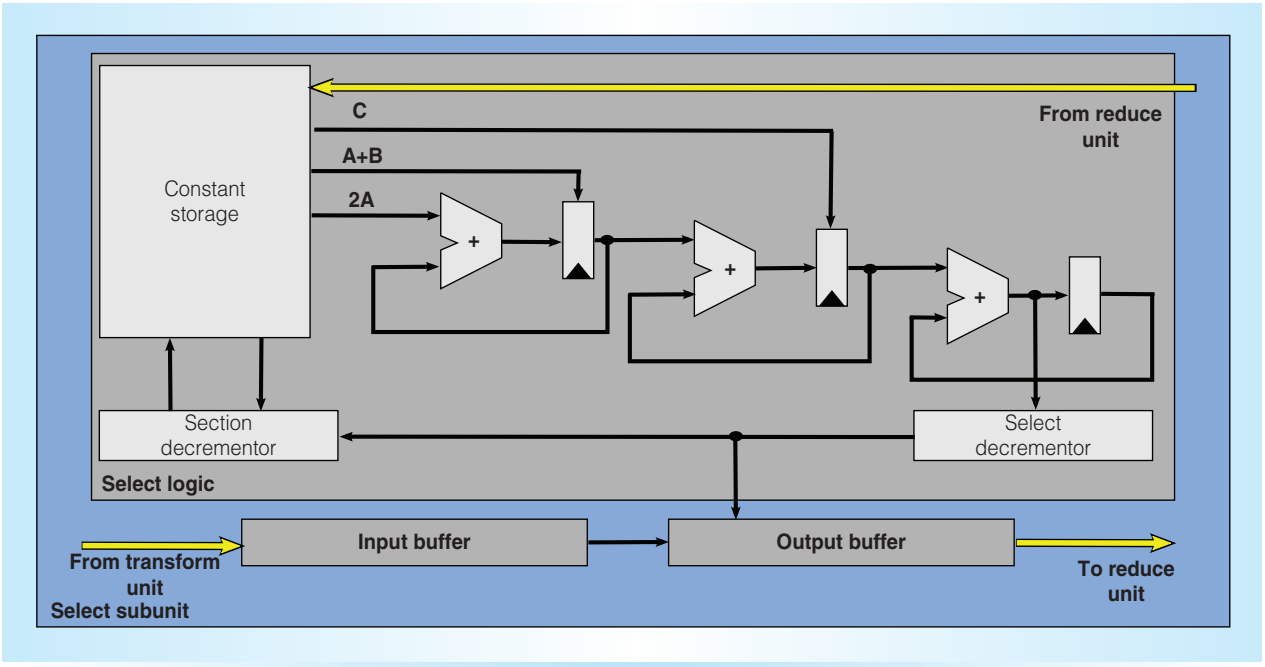

Figure 3. Select unit microarchitecture. Select units map upsampled echo data from the receive time domain to image focal points. Sample data arrives from the transform unit at the input buffer, and each sample is either discarded or copied to the output buffer as determined by our piecewise quadratic approximation algorithm. The constant storage holds the precomputed constants and boundary for each approximation section. The adder chain calculates the next delta index value to determine how far ahead the hardware needs to iterate to find the next focal point, with the final adder accumulating fractional bits from previous additions. The select decrementor is initialized with the integer component of the adder chain. In each cycle, the head of the input buffer is copied to the output if the decrementor is zero, or discarded if it is nonzero. The section decrementor tracks when to advance to the next piece-wise section.

select subunit. As the subunits each drain their input buffers, new data is streamed in from SRAM. On each clock cycle, a sample is popped from the head of the input buffer, and the select logic determines whether the data corresponds to the next focal point on the scanline. If the sample is selected, it is copied to the output buffer. Otherwise, the sample is discarded. Whenever the output buffer fills, its contents are sent to the reduce stage.

The select logic implements the piecewise quadratic delay calculation described previously. The logic calculates the delta—that is, the number of samples to discard—between two consecutive focal points. The unit comprises constant storage that is preloaded with the piecewise quadratic constants required to process a particular set of scanlines, a decrementor that determines when to change sections, a series of adders to generate the delta value, and finally a decrementor that counts down in step with the input buffer and determines which data should be selected. After initialization, the subunit generates the first delta value ($n = 0$) to determine how much to advance the input. This delta value is then loaded into the select decrementor. Once the select decrementor reaches 0, the current data is "selected" and written to the output buffer. A new delta value ($n = 1$) is calculated and the process continues until the entire scanline has been generated. Because of the iterative nature of the calculation, deltas can be calculated efficiently using the adder chain shown in Figure 3. Using this design approach, we can change what is typically an address-calculation and load-intensive software loop to a streaming design, greatly improving efficiency.

Reduce

The final stage is the reduce unit, which ties the 1,024 channels together via a pipelined network. Each reduce unit corresponds to a single node on the network and adds the

Table 1. 3D ultrasound system parameters.

Parameter	Value
Total transmits per frame	96
Total transducers	10,560
Receive transducers per subframe	1,024
SRAM size per receive transducer	4,096 × 12 bits
Focal points per scanline	4,096
Image depth	10 cm
Image total angular width	$\pi/6$
Sampling frequency	40 MHz
Interpolation factor	4×
Interpolated sampling frequency	160 MHz
Speed of sound (tissue)	1540 m/s
Target frame rate	1 frame/s

Table 2. CNR values for ideal system and Sonic Millip3De (SM3D). Values correspond to cysts shown in Figure 4.

Left column of cysts in Figure 4		Right column of cysts in Figure 4	
Ideal	**SM3D**	**Ideal**	**SM3D**
3.59	3.58	1.93	1.85
3.18	3.21	1.51	1.41
2.68	2.67	1.94	1.85
1.61	1.62	2.10	2.01
1.10	1.18	2.39	2.30
0.33	0.39	2.43	2.34

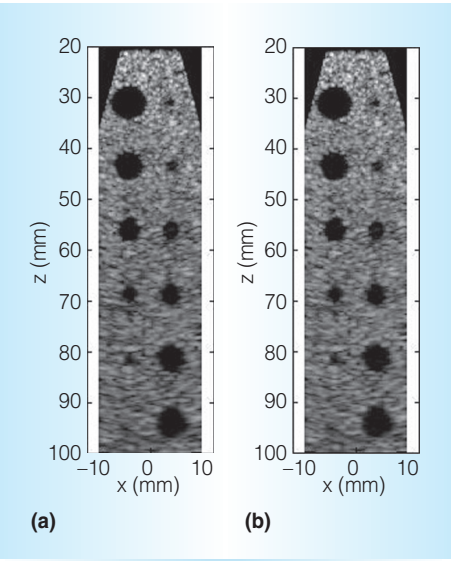

Figure 4. Image quality comparison. *X* to *Z* (horizontal) slice through a series of cysts from a 3D simulation using Field II,[9,10] generated with double-precision floating-point and exact delay index calculation (a). The same slice generated via our delay algorithm, fixed-point precision, and dynamic focus (b). Table 2 gives the contrast-to-noise ratios (CNRs) for both.

data received from the preceding node with the data from the local select unit before sending the summed result to the next node on the network.

Methodology and results

We evaluate our system in terms of both image quality and system power. Because Sonic Millip3De is intended for diagnostic medical imaging, it is critical that it generates high-quality images, comparable to existing devices. Hence, the goal of our image quality evaluation is to confirm that the approximation techniques used to reduce power do not unacceptably degrade image quality. We contrast images generated according to our method against an ideal system without power constraints or approximations.

In our image quality analysis, we simulate cysts in tissue using Field II,[9,10] varying cyst size with depth and covering a range of 8 cm (2- to 10-cm depth). Table 1 shows the relevant ultrasound system parameters. We generate 3D images using both our system (iterative delay calculation and fixed-point adders) as well as an ideal system (full delay calculation and double-precision floating-point arithmetic). Figure 4 shows a 2D slice for both images. We quantitatively compare image quality using contrast-to-noise ratios (CNRs) for each cyst, shown in Table 2. Overall, Sonic Millip3De's image quality is nearly indistinguishable from the ideal case, providing high image quality and validating our algorithm design.

We analyze the full system power of Sonic Millip3De using a register transfer level design targeting a 45-nm standard cell library and Spice models of the global interconnect. Using results from synthesis (SRAM, beamformer, interconnect) and published values (transducers, analog-to-digital converters, memory interface, DRAM), we determine that the design requires 14.6 W in current 45-nm technology, falling a bit short of the ambitious 5-W power target. However, under current

scaling trends, we project that the design will meet the target 5-W budget by the 16-nm technology node.

Computer architecture is continuing to move toward heterogeneous designs, with accelerated processing units for multimedia and graphics already commonplace today. As the heterogeneous space grows, so will the need for more advanced accelerators. With Sonic Millip3De, we have targeted a specific application (handheld 3D ultrasound) that has incredible potential impact and whose unique form of computation is not well suited to existing designs. We believe that focusing on such problems will help future accelerator design move forward.

Furthermore, a key take-away from our process is the importance of codesigning algorithms and hardware when high-efficiency gains are required. Sonic Millip3De achieves orders-of-magnitude greater energy efficiency over stock ultrasound designs, which would not have been possible with just a simple hardware solution. The fundamental reworking of the algorithm itself enabled the streaming dataflow that lies at the heart of our efficiency gains, and was a critical component in our hardware design. However, because our algorithmic modifications introduce approximations, they necessitated a domain-specific evaluation to ensure that result quality was not compromised. Additionally, emerging architectural techniques such as 3D die stacking helped us create a design that simply could not have existed previously, and they show great potential for new and unique heterogeneous systems. MICRO

Acknowledgments

This work was partially supported by NSF CCF-0815457, CSR-0910699, and grants from ARM Inc. The authors thank J. Brian Fowlkes, Oliver Kripfgans, and Paul Carson for feedback and assistance with image quality analysis and Ron Dreslinski for assistance with Spice.

References

1. D. Weinreb and J. Stahl, *The Introduction of a Portable Head/Neck CT Scanner May Be Associated with an 86% Increase in the Predicted Percent of Acute Stroke Patients Treatable with Thrombolytic Therapy*, Radiological Soc. of North America, 2008.
2. M. Shorter and D.J. Macias, "Portable Handheld Ultrasound in Austere Environments: Use in the Haiti Disaster," *Prehospital Disaster and Medicine*, vol. 27, no. 2, 2012, pp. 172-177.
3. J.A. Nations and R.F. Browning, "Battlefield Applications for Handheld Ultrasound," *Ultrasound Quarterly*, vol. 27, no. 3, 2011, pp. 171-176.
4. R. Sampson et al., "Sonic Millip3De: A Massively Parallel 3D Stacked Accelerator for 3D Ultrasound," *Proc. 19th IEEE Int'l Symp. High-Performance Computer Architecture*, 2013, pp. 318-329.
5. R. Sampson et al., "Sonic Millip3De with Dynamic Receive Focusing and Apodization Optimization," *Proc. IEEE Ultrasonics, Ferroelectrics, and Frequency Control Soc. Symp.*, 2013, pp. 557-560.
6. C. Passmann and H. Eermert, "A 100-MHz Ultrasound Imaging System for Dermatologic and Ophthalmologic Diagnostics," *IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control*, vol. 43, no. 4, 1996, pp. 545-552.
7. K. Karadayi, C. Lee, and Y. Kim, *Software-Based Ultrasound Beamforming on Multicore DSPs*, tech. report, Univ. of Washington, March 2011.
8. F. Zhang et al., "Parallelization and Performance of 3D Ultrasound Imaging Beamforming Algorithms on Modern Clusters," *Proc. 16th Int'l Conf. Supercomputing*, 2002, pp. 294-304.
9. J. Jensen, "Field: A Program for Simulating Ultrasound Systems," *Proc. 10th Nordic-Baltic Conf. Biomedical Imaging*, 1996, pp. 351-353.
10. J.A. Jensen and N.B. Svendsen, "Calculation of Pressure Fields from Arbitrarily Shaped, Apodized, and Excited Ultrasound Transducers," *IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control*, vol. 39, no. 2, 1992, pp. 262-267.

Richard Sampson is a PhD candidate in the Department of Electrical Engineering and Computer Science at the University of

Michigan. His research interests include hardware-system and accelerator design for imaging and computer vision applications. Sampson has an MS in computer science and engineering from the University of Michigan.

Ming Yang is a PhD student in the School of Electrical, Computer and Energy Engineering at Arizona State University. His research focuses on the development of algorithms and low-power hardware for a handheld 3D ultrasound imaging device. Yang has an MS in electrical engineering from Beijing University of Posts and Telecommunications.

Siyuan Wei is a PhD student in the School of Electrical, Computer and Energy Engineering at Arizona State University. His research focuses on algorithm-architecture codesign of ultrasound imaging systems, especially those based on Doppler imaging. Wei has a BE in electrical engineering from Huazhong University of Technology and Science.

Chaitali Chakrabarti is a professor in the School of Electrical, Computer and Energy Engineering at Arizona State University. Her research interests include low-power system design, VLSI algorithms, and architectures for signal processing systems. Chakrabarti has a PhD in electrical engineering from the University of Maryland. She is a fellow of IEEE.

Thomas F. Wenisch is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. His research focuses on computer architecture, particularly multiprocessor and multicore systems, multicore programmability, smartphone architecture, data center architecture, and performance evaluation methodology. Wenisch has a PhD in electrical and computer engineering from Carnegie Mellon University. He is a member of IEEE and the ACM.

Direct questions and comments about this article to Thomas F. Wenisch, Computer Science & Engineering Department, 2260 Hayward St., Ann Arbor, MI 48109; twenisch@umich.edu.