WHITE PAPER

Gene Frantz,
TI Principal Fellow,
Texas Instruments

TEXAS INSTRUMENTS

Where will floating point take us?

Introduction

Floating point has been the center of debate in the signal-processing arena for as long as I have been part of this industry. My first introduction to signal processing was in an era where mini-computers were the mainstay of research, and every mini-computer had an array processor attached to it. Thus began the debate.

The debate was simple: one of performance versus accuracy. While floating point, particularly double-precision floating point, would guarantee that the accuracy requirements of the system would be retained, it reduced the raw performance of the system.

There are generally three different accuracy requirements of a signal-processing system: data accuracy, coefficient accuracy and internal accuracy. Each data set has a different combination of these three.

When TI introduced its first digital signal processor (DSP), we focused on raw performance and used sophisticated tricks to overcome issues introduced into the required accuracy. The TMS32010 DSP was a 16-bit fixed-point machine. Generally, the industry followed this path of fixed-point machines ranging from 16 to 24 bits of data accuracy, double that and add a little for internal accuracy.

The TMS320C30 DSP

After two generations of fixed-point signal processors with 16-bit data words, we began to consider the possibility of a floating-point architecture for our third-generation device. It was an immediate success in the market. The difficulty was determining why it was successful. Was it because it was a) floating point, b) 32 bits or c) high-level language friendly? Even with its great success, it did not alter the even greater success of the previous generations of fixed-point DSPs.

The reason for the success of the floating-point device was added convenience, but performance was sacrificed as a result. The world of signal processing was performance driven, and sacrificing any amount of raw performance was unacceptable. Although, there was a mathematics niche discovered for floating points within the DSP community, it was more for any end system that required a division. This design requirement gave the floating-point device an advantage, as the designer did not have to worry about the “divide by zero” scenario.

This advantage was particularly helpful in systems performing matrix math, such as military applications and medical imaging. In each of these cases, the lack of performance could easily be corrected by multi-processing solutions, because system cost was not the highest priority of the design. In the world of floating-point DSPs, this created a second generation of floating-point devices which were designed for multi-processing designs. For Texas Instruments, this new device was the TMS320C40 DSP.

So where did that leave fixed-point architecture? For tangible DSP applications where performance was king, the fixed-point architecture continued to thrive.

Return to performance

After the introduction of the TMS320C40 DSP, we were mostly done with our experiment with floating-point signal processors. In fact, as we introduced higher-performance processors to the industry, we noticed medical imaging designs moving back to our fixed-point DSPs for the additional raw performance. That is why fixed point became our focus for a decade.

During this resurgence of the fixed-point DSP, we took some time to visit our floating-point DSP customers to learn exactly why our hardcore floating-point customers weren’t buying based on the performance philosophy we were selling. This was certainly an eye opener.

The big surprise

Not only did floating point give the advantage of ease-of-use and protection against the "divide by zero" concern, it was virtually on par with our fixed-point, high-performance devices. Furthermore, the significant price placed on floating-point devices equated to higher profit margins based on the lower volume production levels. Floating point was alive and well.

The value of faster time-to-market and ease-of-design using high-level programming languages and operating systems and higher comfort levels with algorithms, kept floating point alive – in spite of the sacrifice of performance and higher price.

What is it that makes floating point the obvious choice in many markets? Let's begin with the original concept of this paper – the three accuracy requirements of a data set. Table 1 below gives examples of several markets.

Table 1.

Market	Data Accuracy	Data Dynamic Range	Coefficient Accuracy	Coefficient Dynamic Range	Internal Accuracy	Internal Dynamic Range
Telecom	6	13	12	16	32	32
Audio	24		53		32	
Video/Imaging	6	28	12	12	16	26
Radar	1	8	8	8	12	12

Today's floating-point processors are designed to handle two different data types: the single-precision floating point and the double-precision floating point. These two data types cover all the various data accuracies necessary for the markets classically driven by digital signal processing. However, focusing on floating point is done with sacrifice to performance when compared to a typical 16-bit fixed-point machine.

This was the secret of the battle between fixed- and floating-point devices for the last decade. A closer look revealed that our fixed-point devices, which were 16-bit devices, were slowly evolving to our 32-bit floating-point devices. With both fixed- and floating-point systems becoming 32-bit systems, we were driven to bring the two systems to parity.

Once we were aware our DSP architectures were 32 bit, we began to separate the concept of architecture from data set sizes, which then led us to the future world of signal processing.

Giving fixed point more performance

Before looking at the future of floating-point systems, let me take a few paragraphs to note what has been done in the fixed-point world to significantly improve raw performance. This will help capture the vision of the future of floating point.

Once we evolved our processors to 32 bits, we saw there were actually two variables affecting performance: Size of the processor, (a 32-bit machine is slower than a 16-bit machine) and size of the hardware multiplier. In a fixed-point processor, the multiplier is typically a 16 × 16-bit multiplier. Depending on the data, however, this can easily be broken into four 8 × 8 multipliers. If the data set allows, (i.e., telecom) for 8-bit data, then four multiples can occur in the same multiplier as one 16-bit multiplier. This equates to a significant increase in performance, given the right data set.

Audio

Over the history of digital audio, the processors used in designs have evolved from a 24-bit, fixed-point processor to an extended, single-precision floating-point processor (32-bit mantissa, rather than 24-bit single-precision floating point) to a double-precision, floating-point processor (53-bit mantissa). In each of these steps, the quality of audio was improved. Surprisingly, it was not the data that was needed to extend accuracy, but rather the coefficient accuracy.

In a digital audio system, the band-pass filters are implemented using bi-quad filters. Bi-quad filters are a popular way of implementing an Infinite Impulse Response (IIR) filter. However, there is a significant weakness in a bi-quad filter when the center frequency of the filter is a small percentage of the sample rate (i.e., at low frequencies). At these points, the poles in the filter are very near the unit circle and need to be very accurate. In floating-point systems, the least accurate point is at "1," which in this case, the poles of the filter are very near. In simple terms, the more bits in the mantissa (accuracy), the better the filter functions. With each of the advances described above, more bits were available to accurately describe the poles (24 bits to 32 bits to 53 bits). This became even more important as the system requirements moved to higher sample rates (48 KHz to 96 KHz, etc.) and to lower desired frequency cut-off points (20 Hz going to 10 Hz, etc.).

In our study of "why floating point", we concluded that there were two significant markets: 1) those markets needing ease-of-use and short time-to-market and 2) audio. Of those two, audio was the interesting market to pursue where floating point was more important than raw performance.

What does this mean?

I have spent a bit of time in audio-specific design, so I'd like to share my vision of where I believe we will take floating point in the future. My comments on how we increased performance on fixed-point processors by fitting the math system to the data set help to support this concept. Let me demonstrate this with some research currently underway at the University of Southern California (USC).

USC has been conducting research for the last decade on artificial vision. TI has been collaborating with them as they endeavor to return sight to people who have gone blind as a result of Retinitis Pigmentosa and Macular Degeneration. The latest collaboration has been on a camera small enough to be implanted in the eye. This began our discussion on what would be the best data representation of the human eye. It seems that the human eye has a dynamic range of about 24 to 26 bits with an accuracy need of 6 to 8 bits. Asked how one would use a 16-bit number to best represent the eye, the research team came up with a 16-bit floating-point number with 5 bits of exponent and 11 bits of mantissa.

Ultimately, this system would allow for a dynamic range of up to 42 bits while keeping 11 bits of accuracy. Not surprisingly, this matches the half-precision definition of the Institute of Electrical and Electronics Engineering (IEEE) floating-point standard. An important side note: 16-bit floating point should give an order of magnitude more raw performance than a 32-bit floating-point system and 30 percent higher raw performance than its equivalent 16-bit fixed-point system.

This is an example of a data set that fits a particular math system well. We have talked about two of them so far in this paper:

- Audio, which matches well with a 32- and 64-bit floating-point system (single and double precision).
- Video and imaging, which matches well with a 16-bit floating-point system.

However, there are other data sets and other math systems to consider. With the flexibility we are incorporating into our signal-processing architectures, we will be prepared to adopt many other floating-point systems.

Floating-point does not need to take on the same concepts as it did when we first introduced it to DSP, which means 16-bit data is fixed point and 32-bit data is floating point. Just as we now have double-precision, single-precision and half-precision IEEE formats (64, 32 and 16 bits respectively), why not add a quarter-precision (perhaps 8-bits exponent and 1-bit implied mantissa), or 12-bit floating point (4-bits exponent and 8-bits mantissa), and so on. The question of what data sets best fit these different math systems might be answered in the following recommendations in Table 2:

Table 2.

Precision	Data Sets
Double	Analog, high-performance audio
Single	Audio
Half	Video and imaging
12 bit	Speech, telecom
Quarter	3-D imaging (e.g., ToF)

Conclusion

In the history of signal processing, we started with analog. Analog was replaced with computers and array processors with double-precision floating point. Array processors were replaced by 16- to 24-bit fixed-point signal processors. The fixed-point signal processors were replaced by single-precision, floating-point processors. The single-precision, floating-point processor was replaced by either a return to 16-bit fixed-point processors for raw performance or an extended-/double-precision floating-point processor for niche opportunities. TI prides itself in offering the right processor for the right design, and now, we have it all. Plus, we can create new variations of our architecture, as we maintain compatibility and match our math system with the data set of interest. In doing so, we can effectively optimize the system's trade off of performance, price and power dissipation.

IMPORTANT NOTICE

Texas Instruments Incorporated and its subsidiaries (TI) reserve the right to make corrections, modifications, enhancements, improvements, and other changes to its products and services at any time and to discontinue any product or service without notice. Customers should obtain the latest relevant information before placing orders and should verify that such information is current and complete. All products are sold subject to TI's terms and conditions of sale supplied at the time of order acknowledgment.

TI warrants performance of its hardware products to the specifications applicable at the time of sale in accordance with TI's standard warranty. Testing and other quality control techniques are used to the extent TI deems necessary to support this warranty. Except where mandated by government requirements, testing of all parameters of each product is not necessarily performed.

TI assumes no liability for applications assistance or customer product design. Customers are responsible for their products and applications using TI components. To minimize the risks associated with customer products and applications, customers should provide adequate design and operating safeguards.

TI does not warrant or represent that any license, either express or implied, is granted under any TI patent right, copyright, mask work right, or other TI intellectual property right relating to any combination, machine, or process in which TI products or services are used. Information published by TI regarding third-party products or services does not constitute a license from TI to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property of the third party, or a license from TI under the patents or other intellectual property of TI.

Reproduction of TI information in TI data books or data sheets is permissible only if reproduction is without alteration and is accompanied by all associated warranties, conditions, limitations, and notices. Reproduction of this information with alteration is an unfair and deceptive business practice. TI is not responsible or liable for such altered documentation. Information of third parties may be subject to additional restrictions.

Resale of TI products or services with statements different from or beyond the parameters stated by TI for that product or service voids all express and any implied warranties for the associated TI product or service and is an unfair and deceptive business practice. TI is not responsible or liable for any such statements.

TI products are not authorized for use in safety-critical applications (such as life support) where a failure of the TI product would reasonably be expected to cause severe personal injury or death, unless officers of the parties have executed an agreement specifically governing such use. Buyers represent that they have all necessary expertise in the safety and regulatory ramifications of their applications, and acknowledge and agree that they are solely responsible for all legal, regulatory and safety-related requirements concerning their products and any use of TI products in such safety-critical applications, notwithstanding any applications-related information or support that may be provided by TI. Further, Buyers must fully indemnify TI and its representatives against any damages arising out of the use of TI products in such safety-critical applications.

TI products are neither designed nor intended for use in military/aerospace applications or environments unless the TI products are specifically designated by TI as military-grade or "enhanced plastic." Only products designated by TI as military-grade meet military specifications. Buyers acknowledge and agree that any such use of TI products which TI has not designated as military-grade is solely at the Buyer's risk, and that they are solely responsible for compliance with all legal and regulatory requirements in connection with such use.

TI products are neither designed nor intended for use in automotive applications or environments unless the specific TI products are designated by TI as compliant with ISO/TS 16949 requirements. Buyers acknowledge and agree that, if they use any non-designated products in automotive applications, TI will not be responsible for any failure to meet such requirements.

Following are URLs where you can obtain information on other Texas Instruments products and application solutions:

Products		**Applications**	
Amplifiers	amplifier.ti.com	Audio	www.ti.com/audio
Data Converters	dataconverter.ti.com	Automotive	www.ti.com/automotive
DLP® Products	www.dlp.com	Communications and Telecom	www.ti.com/communications
DSP	dsp.ti.com	Computers and Peripherals	www.ti.com/computers
Clocks and Timers	www.ti.com/clocks	Consumer Electronics	www.ti.com/consumer-apps
Interface	interface.ti.com	Energy	www.ti.com/energy
Logic	logic.ti.com	Industrial	www.ti.com/industrial
Power Mgmt	power.ti.com	Medical	www.ti.com/medical
Microcontrollers	microcontroller.ti.com	Security	www.ti.com/security
RFID	www.ti-rfid.com	Space, Avionics & Defense	www.ti.com/space-avionics-defense
RF/IF and ZigBee® Solutions	www.ti.com/lprf	Video and Imaging	www.ti.com/video
		Wireless	www.ti.com/wireless-apps

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265