Reducing Latency, Power, and Gate Count with the Tensilica Floating-Point FMA

David Chen, Cadence

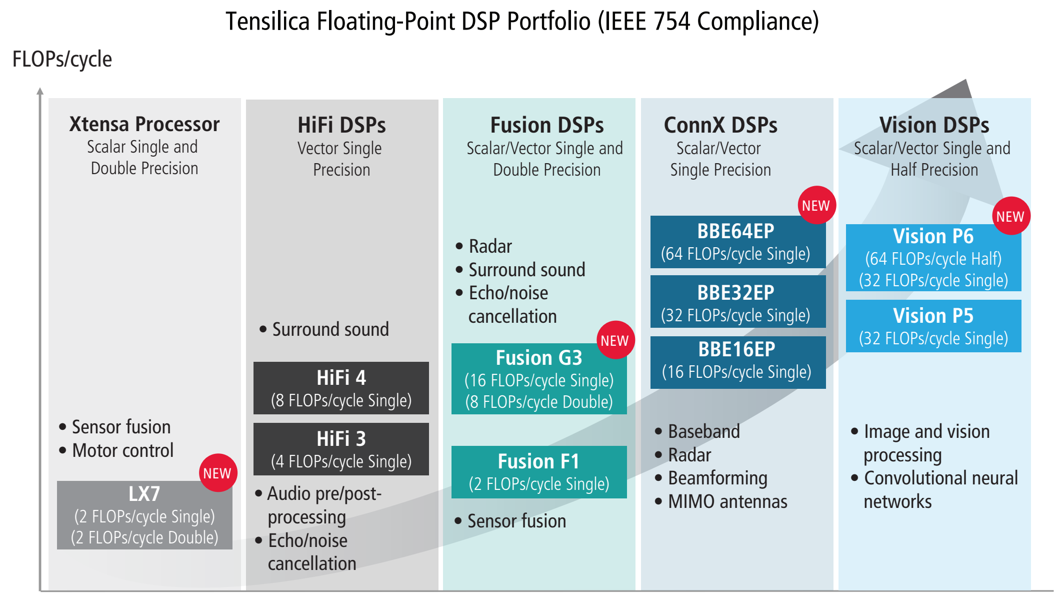
Today's digital signal processing applications, which have greater demand for real-time computing, require more dynamic range for number representation and computation accuracy. Floating-point arithmetic units better meet these needs than fixed-point arithmetic units. To overcome some of the drawbacks of floating-point arithmetic units, such as latency, power, and gate count, Cadence has developed the innovative Tensilica® floating-point fused multiply-add (FMA) design.

Contents

Introduction

Today's digital signal processing applications such as radar, echo cancellation, and image processing are demanding more dynamic range and computation accuracy. Floating-point arithmetic units offer better precision, higher dynamic range, and shorter development cycles when compared to fixed-point arithmetic units. Minimizing the design's time to market is more important than ever. Algorithm developers use MATLAB to develop and test their ideas, which are mostly floating-point arithmetic based. However, digital signal processor (DSP) programmers port the algorithms into fixed-precision arithmetic units since floating-point arithmetic units are considerably larger, slower, and more power-hungry than the fixed-point arithmetic units. This is not a trivial effort as the programmers must verify the results, including the error rate (accuracy) of fixed-point and floating-point algorithms. Furthermore, usually fixed-point software codes require more cycles than floating-point versions on many algorithms. For example, using the Cadence® Tensilica BBE32EP DSP, a 4x4 matrix Cholesky decomposition in fixed-point takes 18 cycles, while in floating-point it takes 15 cycles. As this example illustrates, it makes sense to keep using floating-point computation units when greater dynamic range and accuracy are required for an application. To overcome some of the drawbacks of floating-point arithmetic units, Cadence has developed an innovative patent-granted design.

Cadence offers various DSPs to meet the needs of a variety of applications, such as audio/voice processing, 5G LTE, and image processing with low-end and high-end computing needs. Most of these DSPs have floating-point options (vectored) ranging from half-precision (16-bit) to double-precision (64-bit), as shown in Figure 1. The Tensilica DSP floating-point unit is a highly innovative, patent-granted (US Patent # 9,519,458) design that overcomes issues associated with the floating-point arithmetic unit.



Note: FLOPs = Floating-point ADD, SUB, or MUL operations

Figure 1 The Tensilica Floating-Point DSP Portfolio

What Is an FMA?

A fused-multiply-add (FMA) is a floating-point multiply-add operation performed in a single step with single rounding. Compared to executing floating-point multiply and addition separately, an FMA can improve the accuracy of many computations that involve the accumulation of products. Therefore, many DSPs that include a floating-point instruction set also include FMA instructions. An FMA is useful for various digital signal processing codes such as filters (FIR, IIR, etc.) and FFTs. Also, an FMA is useful for software implementations of floating-point division and square root computations.

However, a conventional floating-point FMA design for performance-oriented DSPs (with higher clock frequency) does not meet the low-latency, lower power, and low-gate-count cost requirements of computation-intensive applications, such as radar. Low latency is desired to enable the best performance of DSP kernel codes to reduce loop unrolling and a data-dependency interlock stall. Costs (chip size and power) are crucial. Moreover, due to high computation demands, DSPs need to provide vectored floating-point units to increase computation throughputs. Therefore, a more cost-effective FMA design is in high demand.

Tensilica FMA vs. Conventional FMA

To address this demand, Cadence has invented a new Tensilica FMA design, which enables shorter latency without degrading clock frequency. Because its latency is shorter, the gate cost and power dissipation are both reduced compared to a conventional design, thus addressing all three requirements of computation-intensive applications.

Figure 2 shows a conventional floating-point FMA that takes six notable steps from multiplication to rounding. As this is a long sequence of computation steps, to achieve a higher clock frequency, designers usually insert pipeline stages (flip-flops). The Tensilica FMA design (Figure 3) removes the 2’s complementor step, which is required to create a positive significand/mantissa as part of the floating-point value (the IEEE 754 standard). Cadence eliminated this step by inventing a new computation flow efficiency. Since the Tensilica FMA achieves the same clock frequency as the conventional FMA, its addition/subtraction step can have a greater timing budget, and thus may utilize cheaper adders. As a result, the Tensilica FMA design achieves a much smaller gate cost than a conventional design.

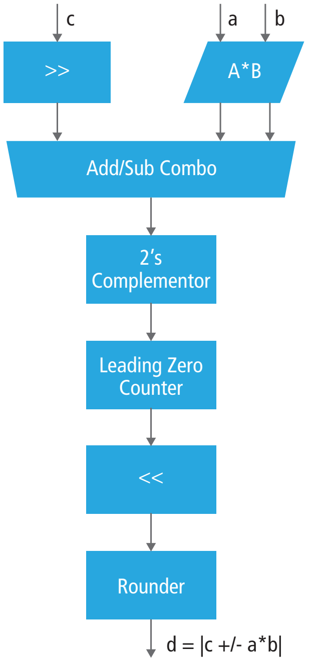


Figure 2: Conventional FMA block diagram

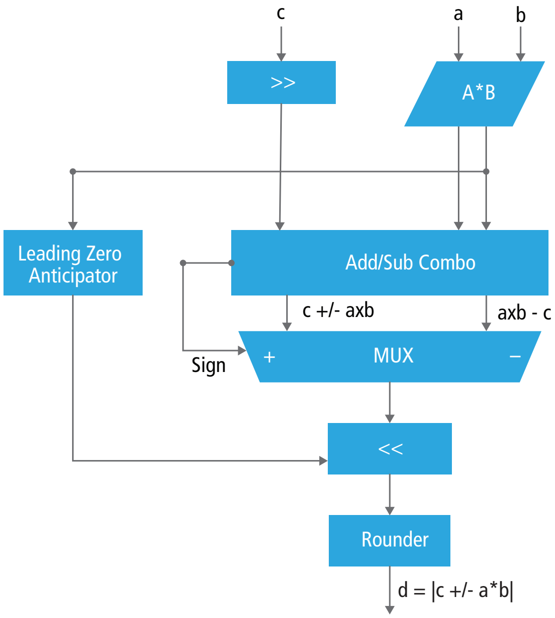


Figure 3: Tensilica FMA block diagram

Example Results

Figure 4 shows example results from the Tensilica FMA design compared to the conventional FMA design (both using double-precision FMA (DPFMA) and single-precision FMA (SPFMA). The FMA logic is synthesized with various target frequencies. The horizontal axis represents delay from the input to the output and the vertical axis represents cost (area). Thus, at a given delay, a lower cost value is better, and at a given cost (area), a lower delay value is better. This example compares the Tensilica FMA from the Tensilica Xtensa® LX7 processor with a previous FMA from the Xtensa LX6 processor (similar to the conventional FMA design shown in Figure 2).

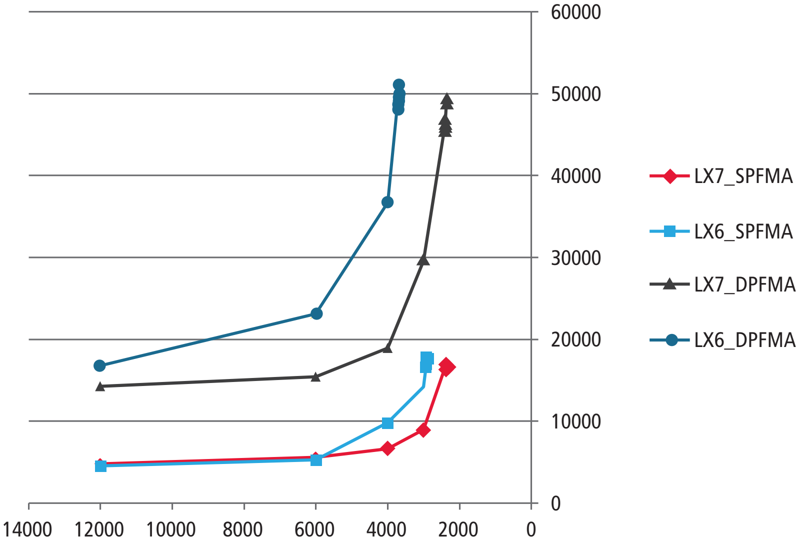


Figure 4. Delay versus area example using the Tensilica FMA

As seen in Figure 3, when the target clock frequency becomes higher (lower delay), the Tensilica FMA design reduces the area cost significantly. This is more obvious for double-precision as the computation requires a larger multiplier and adders. Since power dissipation is roughly in proportion to the gate cost, the Tensilica FMA results in a lower power dissipation.

Applications

We use the Tensilica FMA design with different objectives. For the Tensilica Fusion F1 DSP, more often used with a relatively lower clock frequency (with a specific CMOS process and library), the FMA has a two-cycle latency option. Nonetheless, the Tensilica FMA can achieve a higher MHz target than the conventional FMA. In comparison, the FMA for the Tensilica Vision DSPs, Baseband DSPs, and Fusion G3 DSP, with their higher clock frequencies, has a four-cycle latency. Since these DSPs are wide SIMD machines of 128-bit to 512-bit width, the reduction of FMA area and power dissipation is important, especially for battery-operated applications.

Conclusion

For many years, Cadence has been innovating the floating-point unit to address the demand of massive floating-point computation capability while lowering silicon cost and power dissipation. Our patented Tensilica FMA design achieves these demanding and challenging goals efficiently and effectively and has been successfully deployed by top semiconductor companies. We continue to innovate our floating-point offering for demanding digital signal processing needs.

Additional Information

For more information on the unique abilities and features of Cadence Tensilica processors, see http://ip.cadence.com/ipportfolio/tensilica-ip.

Cadence Design Systems enables global electronic design innovation and plays an essential role in the creation of today's electronics. Customers use Cadence software, hardware, IP, and expertise to design and verify today's mobile, cloud and connectivity applications. www.cadence.com