10.3 A 1.45GHz 52-to-162GFLOPS/W Variable-Precision Floating-Point Fused Multiply-Add Unit with Certainty Tracking in 32nm CMOS

Himanshu Kaul, Mark Anders, Sanu Mathew, Steven Hsu,
Amit Agarwal, Farhana Sheikh, Ram Krishnamurthy, Shekhar Borkar

Intel, Hillsboro, OR

High-throughput floating-point computations are key building blocks of 3D graphics, signal processing and high-performance computing workloads [1,2]. Higher floating-point precisions offer improved accuracy at the expense of performance and energy efficiency, with variable-precision floating-point circuits providing run-time precision selection [3]. Real-time certainty tracking enables variable-precision circuits not only to operate at the higher energy efficiency of low-precision datapaths, but also to preserve high-precision accuracy. A variable-precision floating-point unit that performs fused multiply-adds (FMA) with single-cycle throughput while supporting operation in either 1-way single-precision (24b mantissa), 2-way 12b precision or 4-way 6b precision modes is fabricated in 32nm High-k/Metal-gate CMOS [4]. Simultaneous floating-point certainty tracking, preshifted addends, a combined rounding and negation incrementer, efficient reuse of mantissa datapath for multiple parallel lower precision calculations, robust ultra-low voltage circuits, and fine-grained clock gating enable nominal energy efficiency of 52GFLOPS/W (IEEE 32b single-precision, measured at 1.45GHz, 1.05V, 25°C) with a dense layout occupying 0.045mm^2 (Fig. 10.3.7) while achieving: (i) scalable performance up to 3.6GFLOPS (single-precision), 96mW measured at 1.2V; (ii) up to 4× higher throughput of 14.4GFLOPS with variable-precision, while maintaining single-precision accuracy; (iii) fast single-cycle precision reconfigurability; (iv) precision mode-dependent power consumption for up to 40% clock power reduction; (v) near-threshold single-precision operation measured at 300mV, 1.75MHz, 11μW; and, (vi) peak energy efficiency of 321GFLOPS/W (single-precision) and 1.2TFLOPS/W (6b precision) at 325mV, 25°C.

The 3-cycle latency FMA supports 24b, 2-way 12b, or 4-way 6b mantissas with corresponding single-precision exponents and denormals flushed to zero (Fig. 10.3.1). Certainty tracking circuits operate in parallel with exponent computation, calculating operand-dependent accuracy bounds that indicate the need for increased precision. These circuits enable improved performance and energy efficiency through parallelism when reduced mantissa precisions provide sufficient accuracy, while always achieving single-precision accuracy. A 5b certainty field (U) extends the floating point operands with an additional relative exponent indicating the number of accurate mantissa bits. Any unsuccessful computations, based on either the relative or absolute accuracy of an FMA result, are recomputed with increased precision as needed. Within the first cycle of the mantissa unit, a 72b addend alignment shifter operates in parallel to a 24b×24b unsigned radix-4 Booth-encoded multiplier for a 14% mantissa critical path reduction by avoiding the need for multiplier output alignment (Fig. 10.3.2). Together, a 3:2 compressor, 48b sparse-tree adder and 24b incrementer add the product to the addend in the second cycle, while the leading zero anticipator (LZA) estimates the leading zero count. The unnormalized mantissa width then reduces to 48b based on addend alignment. In the final cycle, the normalization left shifter further reduces the mantissa to 24b, followed by an increment operation to support all IEEE rounding modes. Computing unsigned mantissa results may require negation of the sum: an inversion followed by an increment; this increment operation is combined with the rounding incrementer for a 9% mantissa critical path reduction.

Circuit optimizations enable variable-precision support in the mantissa datapath while minimizing overhead. A unified partial product reduction tree supports multiple mantissa precisions by resetting off-diagonal partial products in 12b/6b modes, resulting in 41%/60% multiplier power reduction (Fig. 10.3.2). Likewise, the 48b adder resets carry signals at 24b/12b boundaries in low-precision modes. The 72b alignment right shifter also supports multiple independent 36b/18b shifts while distributing addends to align with corresponding adder and incrementer inputs (Fig. 10.3.3). Similarly, the 48b normalization shifter output is aligned to enable 2/4-way rounding in the variable-precision 24b incrementer, controlled by up to four independent increment signals (cin0-3) to produce four overflow carries (cout0-3). LZA circuits reset leading one detection and encoding signals at 24b/12b boundaries, enabling up to 4-way leading zero counts. Overall support for variable-precision increases mantissa power by 11% and area by 12% for 4× higher performance at lower precisions.

To realize performance gains at lower precisions while preserving single-precision accuracy, certainty tracking circuits operate in parallel with exponent circuits to calculate output certainty based on input certainties, FMA calculation and precision mode (Fig. 10.3.4). The floating-point calculation (O=A×B+C) determines the center of a range, while the certainty terms are grouped to obtain the output certainty (ΔO). An additional product recentering term ($\Delta A \times \Delta B = 2^{Ep-Ua-Ub}$) is added to the appropriate multiplier compressor tree column to reduce ΔO. The remaining certainty terms are summed by finding the largest with an adjustment to bound the smaller terms. First the minimum relative certainty of the multiplier inputs is found, followed by the minimum absolute certainty with the addend input, and finally the relative certainty is compared to precision-based rounding error. Since the output certainty is a power of 2, decrementing the certainty from this largest term by 1 or 2 accounts for the addition of all the smaller certainty terms. Compared to four single-precision FMAs, these circuits enable 2.3× area reduction and up to 2.7× power reduction for the same throughput.

The variable-precision FMA operates at 11.6GFLOPS (1FMA = 2FLOPs) throughput at 1.45GHz, measured at nominal 1.05V, 25°C when in 4-way 6b mantissa mode with 72mW total power and 440μW leakage component (Fig. 10.3.5). The corresponding performance for IEEE single-precision calculations is 2.9GFLOPS with a power consumption of 56mW without certainty tracking. The resulting energy efficiency is 52GFLOPS/W (19.4pJ/FLOP) for single-precision, scaling to 162GFLOPS/W (6.2pJ/FLOP) for 4-way 6b precision operations. Peak performance of 14.4GFLOPS is achieved in 4-way 6b mode by scaling the supply voltage to 1.2V for a power consumption of 121mW and efficiency of 119GFLOPS/W. Datapath circuits optimized for reliable ultra-low voltage operation enable robust functionality measured down to 300mV, enabling a peak efficiency of 1.2TFLOPS/W (7.3× improvement scaling from 1.05V) for 4-way 6b precision at 325mV with a power consumption of 21μW and performance of 25MFLOPS. Mode-dependent power savings are achieved by clock gating unused exponent and certainty circuits, reducing clock power by 40% from 4-way 6b precision with certainty tracking to single-precision with no certainty tracking (Fig. 10.3.6a). Overall, the 4× throughput increase in 6b mode results in a 3.1× - 3.7× energy efficiency improvement from 1.05V - 325mV. Algorithm-dependent energy efficiency gains are determined by successful computations at lower precisions; 6b mode is attempted first, with uncertain results recalculated in 12b mode, followed by single-precision mode when necessary (Fig. 10.3.6b). These energy efficiency gains, relative to only single-precision computation, improve as the percentage of successful 6b or 12b operations increases, to a maximum of 3.1×.

Acknowledgments:
The authors thank V. De, R. Forand, W.H. Wang, G. Taylor, J. Crawford, T. Piazza and W. Pinfold for encouragement and discussions. This research was, in part, funded by the U.S. Government under contract number HR0011-10-3-0007. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References:
[1] H.-J. Oh, et. al., "A Fully Pipelined Single-Precision Floating-Point Unit in the Synergistic Processor Element of a CELL Processor", *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 759-771, 2006.
[2] B. Curran, et. al., "4GHz+ Low-Latency Fixed-Point and Binary Floating-Point Execution Units for the POWER6 Processor", *ISSCC Dig. Tech. Papers*, pp. 436-437, 2006.
[3] M. J. Schulte, et. al., "A Family of Variable-Precision Interval Arithmetic Processors," *IEEE Trans. on Computers*, vol. 49, no. 5, pp. 387-397, May 2000.
[4] C.-H. Jan, et. al., "A 32nm SoC platform technology with 2nd generation high-k/metal gate transistors optimized for ultra low power, high performance, and high density product applications", *IEEE IEDM Dig. Tech. Papers*, pp. 1-4, 2009.

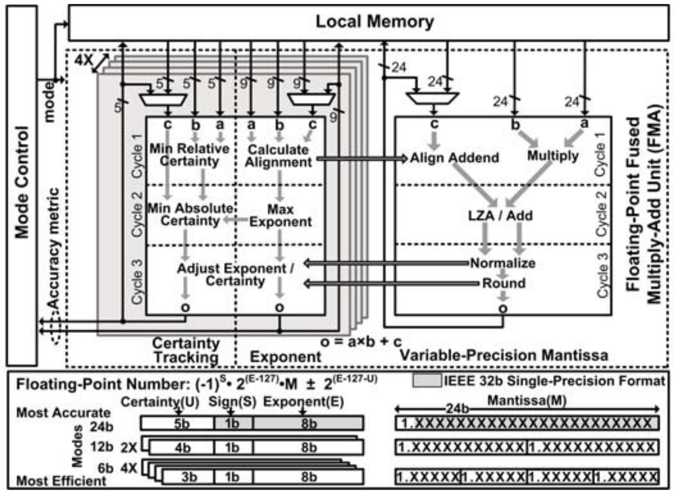

Local Memory
Mode Control
mode
Accuracy metric
4X
Cycle 1
Min Relative Certainty
Calculate Alignment
Cycle 2
Min Absolute Certainty
Max Exponent
Cycle 3
Adjust Exponent / Certainty
Align Addend
Multiply
LZA / Add
Normalize
Round
Floating-Point Fused Multiply-Add Unit (FMA)
Certainty Tracking
Exponent
o = a×b + c
Variable-Precision Mantissa
Floating-Point Number: $(-1)^S \cdot 2^{(E-127)} \cdot M \pm 2^{(E-127-U)}$
IEEE 32b Single-Precision Format
Certainty(U) Sign(S) Exponent(E)
Mantissa(M)
Most Accurate
Modes
24b
12b 2X
6b 4X
Most Efficient
5b 1b 8b
4b 1b 8b
3b 1b 8b
1.XXXXXXXXXXXXXXXXXXXXXXX
1.XXXXXXXXXXX1.XXXXXXXXXXX
1.XXXXX1.XXXXX1.XXXXX1.XXXXX

Figure 10.3.1: Overview of variable-precision floating-point fused multiply-add unit with certainty tracking.

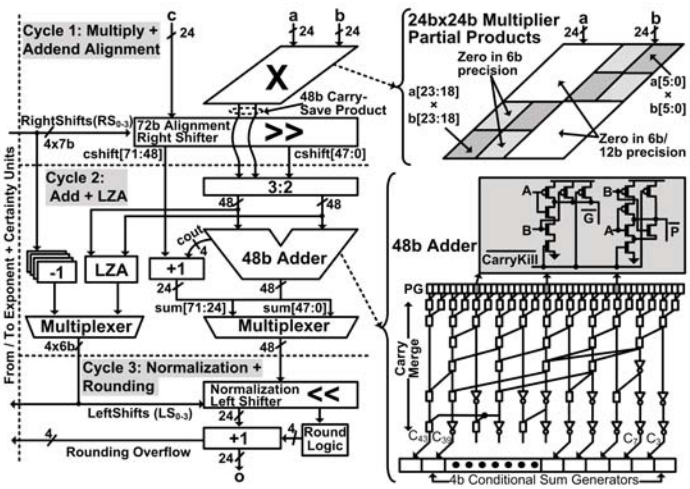


Figure 10.3.2: 3-cycle latency variable-precision mantissa datapath circuits.

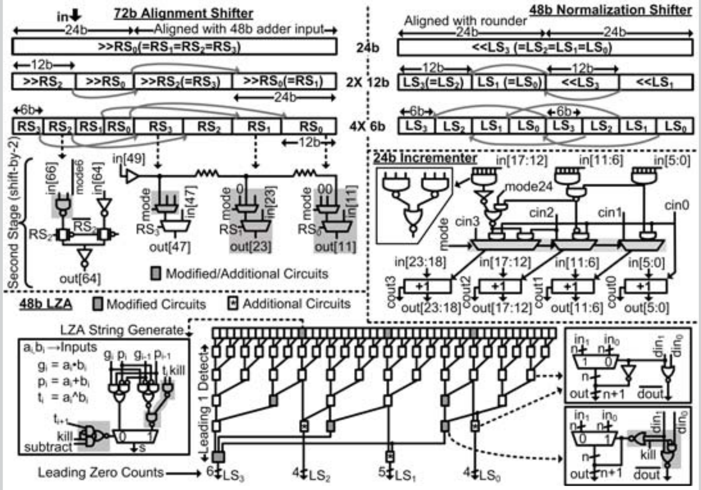


Figure 10.3.3: Variable-precision circuit optimizations for alignment and normalization shifters, incrementer, and LZA.

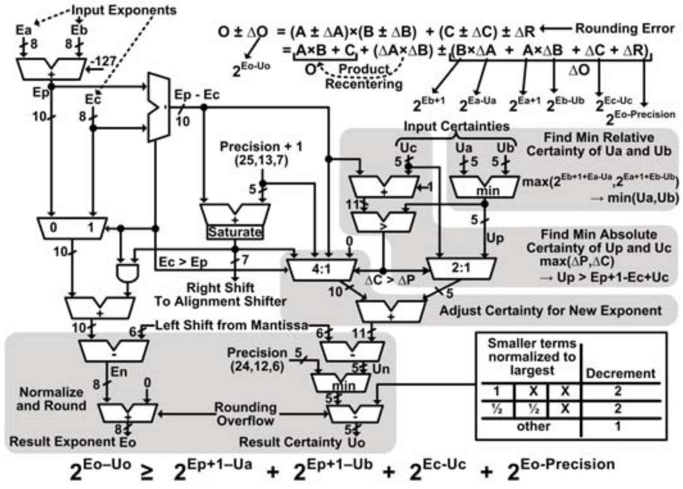

Input Exponents
Ea Eb
127
Ep Ec
Ep - Ec
$O \pm \Delta O = (A \pm \Delta A)\times(B \pm \Delta B) + (C \pm \Delta C) \pm \Delta R$
Rounding Error
$= A\times B + C + (\Delta A \times \Delta B) \pm (B\times\Delta A + A\times\Delta B + \Delta C + \Delta R)$
2^{Eo-Uo}
Product Recentering
ΔO
2^{Eb+1} 2^{Ea-Ua} 2^{Ea+1} 2^{Eb-Ub} 2^{Ec-Uc} $2^{Eo-Precision}$
Input Certainties
Uc Ua Ub
Find Min Relative Certainty of Ua and Ub
$\max(2^{Eb+1+Ea-Ua}, 2^{Ea+1+Eb-Ub}) \rightarrow \min(Ua,Ub)$
Precision + 1 (25,13,7)
Saturate
min
Up
Find Min Absolute Certainty of Up and Uc
$\max(\Delta P, \Delta C) \rightarrow Up > Ep+1-Ec+Uc$
Ec > Ep
4:1
2:1
ΔC > ΔP
Right Shift To Alignment Shifter
Adjust Certainty for New Exponent
Left Shift from Mantissa
En
Precision (24,12,6)
Un
Normalize and Round
Rounding Overflow
Result Exponent Eo
Result Certainty Uo
Smaller terms normalized to largest
Decrement
1 X X 2
½ ½ X 2
other 1
$2^{Eo-Uo} \geq 2^{Ep+1-Ua} + 2^{Ep+1-Ub} + 2^{Ec-Uc} + 2^{Eo-Precision}$

Figure 10.3.4: Exponent and certainty tracking circuits.

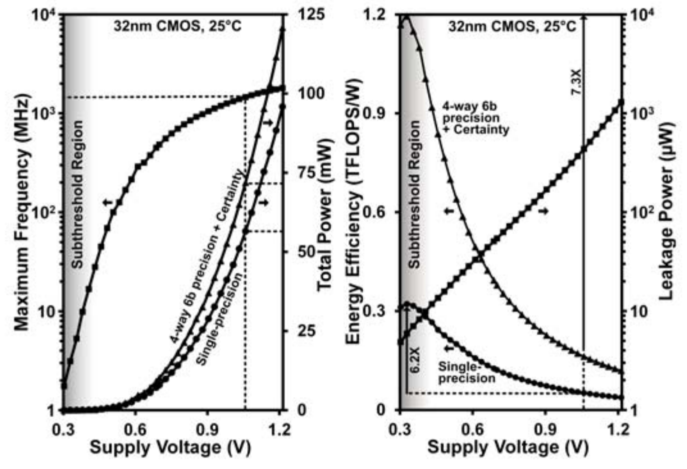


Figure 10.3.5: 32nm CMOS measured maximum frequency, total and leakage power, and energy efficiency for single-precision and 4-way 6b precision FMA operations.

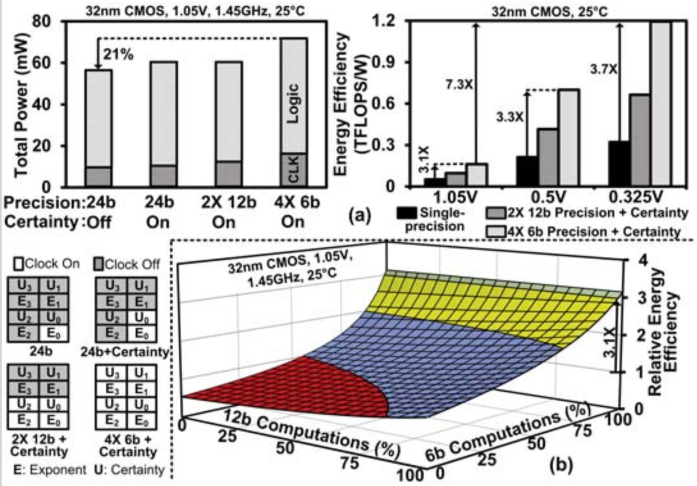
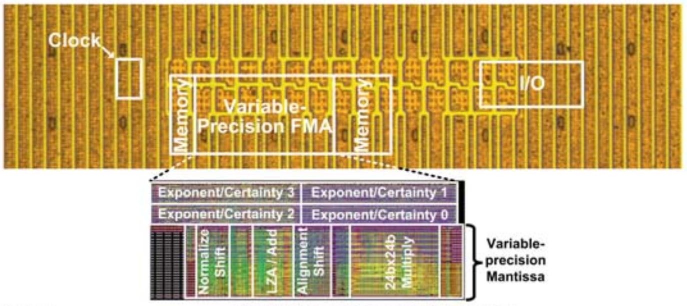


Figure 10.3.6: 32nm CMOS measurements of (a) mode-dependent power and energy efficiency, and (b) overall energy efficiency based on precision requirements.



Process	32nm High-k/Metal-gate CMOS, 9 Metal Cu
FMA area, Transistor count	0.045mm^2, 120K
Power (single-precision)	56mW at 1.45GHz, 1.05V, 25°C (nominal), 440µW active leakage
Variable-precision power	60mW (2-way 12b) or 72mW (4-way 6b) at 1.45GHz, 1.05V, 25°C
Nominal energy efficiency	52GFLOPS/W (single-precision) to 162GFLOPS/W (4-way 6b)
Peak performance	14.4GFLOPS, 121mW, 119GFLOPS/W (4-way 6b) at 1.2V, 25°C
Peak efficiency (ultra-low voltage)	1.2TFLOPS/W, 25MFLOPS, 21µW (4-way 6b) at 325mV, 25°C

Figure 10.3.7: 32nm CMOS variable-precision FMA die micrograph and measured performance summary.