For research article

Response to Reviewer 2 Comments

1. Summary

Thank you very much for taking the time to review this manuscript, for signing the review, and for the detailed and constructive comments. The seven points raised have helped us strengthen the motivation, the positioning of the contribution, and the experimental evaluation. Please find the detailed responses below. All the corresponding revisions are highlighted in green in the re-submitted manuscript.

2. Questions for General Evaluation

Questions for General Evaluation	Reviewer's Evaluation	Response and Revisions
Does the introduction provide sufficient background and include all relevant references?	Yes	A systematic summary of the drawbacks of existing FPGA floating-point implementations has been added (Comment 1), and a recent reference on FPGA-oriented leading-zero counters has been included (Comment 3).
Are all the cited references relevant to the research?	Yes	One additional, recent reference was added (Comment 3); all references remain relevant.
Is the research design appropriate?	Yes	The contribution is now explicitly separated into incremental and engineering parts (Comment 2).
Are the methods adequately described?	Yes	The 0.f-vs-1.f trade-off (Comment 4) and the area-equivalent weighting (Comment 5) are now described more precisely.

Continued on next page

Questions for General Evaluation	Reviewer's Evaluation	Response and Revisions
Are the results clearly presented?	Can be improved	SQNR and frequency-response results for the FIR filter were added (Comment 6), together with a quantitative mode-selection guideline (Comment 7).
Are the conclusions supported by the results?	Yes	The conclusions were updated to reflect the new FIR numerical results (Comment 6).

3. Point-by-point response to Comments and Suggestions for Authors

Comments 1: *The motivation section elaborates the drawbacks of fixed-point arithmetic for radar/communication DSP, but fails to systematically summarize the main drawbacks of existing FPGA floating-point implementations.*

Response 1: We agree. We have added to the introduction a systematic summary of the main drawbacks of existing FPGA floating-point implementations, organized along five recurring axes (lines 40 to 54): *High LUT and register utilization... Long critical paths through the exponent-comparison, alignment, normalization, and rounding stages... Elevated dynamic power... Limited portability, since many high-performance designs rely on HLS flows or on vendor-specific primitive instantiations... The additional logic required by full IEEE-754 special-value handling and by fixed single/double precision...* and we explicitly state that these are the drawbacks the present work targets.

Comments 2: *The authors claim the contribution lies in "a coherent portable unit set + unified ATP characterization", but do not distinguish incremental innovations (0.f encoding + configurable bit-widths) and engineering innovations (portable VHDL implementation + dual deployment modes).*

Response 2: We agree, and we have restructured the contribution paragraph of the introduction to separate the two categories explicitly (lines 94 to 101). The representation-level refinements are described as incremental (the explicit *0.f* significand encoding and the synthesis-time configurable widths), while the engineering contributions (architecture definition and single VHDL-2008 description with no HLS flow and no vendor primitive instantiation; dual DSP-inferred / LUT-only deployment of the same source; the fixed/float conversion units closing the datapath towards ADCs/DACs; and the uniform area-equivalent ATP characterization across two FPGA generations) are identified as the locus of the practical value of this work.

Comments 3: *In Subsection 2.4 (Discussion), the authors point out that the leading-one detector is a major resource overhead for FP adders, but no existing solutions or optimization ideas from published literature are reviewed.*

Response 3: We agree. We have extended the Discussion (Section 2.4) with a short review of leading-zero/leading-one optimization approaches (lines 263 to 271): the distinction between exact leading-zero detection and inexact leading-zero anticipation (LZA) with its one-bit error correction, recent FPGA-oriented leading-zero counter designs, and the leading-zero predictor/anticipator al-

ready used by the surveyed MAC works. We then clarify how the proposed adder addresses the overhead (lines 274 to 277): *. . . by restricting the priority encoder to the most significant half of the result vector. . . which halves the encoder width without loss of precision rather than introducing an inexact LZA and its associated correction logic.* A recent, FPGA-specific reference on leading-zero counter design has been added for this purpose.

Comments 4: *The transformation from standard 1.f to proposed 0.f significand is explained, but the quantitative analysis of precision loss and hardware gain is insufficient.*

Response 4: We have clarified the precision side quantitatively in Section 3.1 (lines 323 to 330). The key point is that the *0.f* encoding is *precision-lossless*: for the same number of stored bits it would lose exactly one bit of significand precision, but the single compensating bit restores bit-for-bit the same relative precision and the same $\frac{1}{2}$ ULP rounding bound as the equivalent *1.f* format, so the quantifiable cost is one bit of *storage density*, not accuracy: *. . . adding the single compensating bit restores bit-for-bit the same relative precision. . . The quantifiable cost is therefore one bit of storage density, not accuracy. . .* This is corroborated by the numerical validation (Section 4.7), where the units configured with $n_f = 24$ match IEEE-754 single precision. Regarding the *hardware-gain* side, since we do not currently have a functionally equivalent *1.f* re-implementation, we expanded the structural argument (which multiplexers and control are removed) and identified a direct, quantitative *1.f*-vs-*0.f* synthesis comparison as future work, to avoid reporting numbers we cannot yet substantiate (lines 312 to 321, in blue, also answering comment 2 of reviewer 1).

Comments 5: *The area-equivalent formula (Equation 6) uses an average weight of 1 DSP = 150 LUTs for two devices, but the rationality of the average weight is not verified.*

Response 5: We thank you the reviewer for noting that point. We have added a justification and a robustness analysis after Equation (6) (lines 611 to 624). First, we note that the exact arithmetic mean of the two device-specific weights (182 and 126) is 154, and that 150 is a rounded convention of the same order. More importantly, we show that the conclusions are robust to this choice: *. . . all comparisons against the Xilinx IP cores are made between designs that use the same number of DSP slices. . . so the DSP term* $150 \cdot DSPs$ *cancels in the numerator of any area difference and the relative ranking is therefore invariant to the chosen weight. . . the academic designs surveyed are LUT-only (zero DSPs), so their area-equivalent figures are completely independent of the DSP weight.* We also report a sensitivity check: recomputing with device-specific weights changes the proposed multiplier's area-equivalent figure by about 12% (UltraScale) and 16% (Artix-7), without altering any qualitative conclusion.

Comments 6: *Supplement SQNR and frequency response test results of the FIR filter, compare with fixed-point and standard IEEE-754 FP implementations to verify practical signal processing performance.*

Response 6: We agree, and we have added a new subsection (*Numerical Behaviour of the FIR Filter*, Section 5.1) with the requested results (lines 813 to 841, table 8 and figures 8 and 9). Using a bit-accurate model of the proposed units (reproducing the 0.f format, RNE rounding, saturation, and the non-fused chained MAC) that propagates the same per-operation behaviour validated on hardware at the unit level (Section 4.7), the 200-tap filter output was compared against a double-precision reference for three equal-width (32-bit) arithmetics: the proposed floating-point format, a 32-bit fixed-point (Q1.30), and IEEE-754 single precision. A new table reports the output SQNR versus input level, and two new figures show the SQNR over the dynamic range and the magnitude frequency response. The main findings, now stated in the manuscript, are: *At full scale the 32-bit fixed-point format attains the highest SQNR. . . the fixed-point SQNR degrades by approximately 1 dB per dB of input attenuation, whereas the floating-point SQNR stays essentially constant. . . The two formats cross over near* -50 *dBFS: below that level the proposed floating-point unit set outperforms the equal-width fixed-point implementation. . . the magnitude frequency response of the three arith-*

metics is visually indistinguishable from the double-precision reference. The proposed format and IEEE-754 single precision yield the same SQNR ($\approx 131.8\,\mathrm{dB}$), consistent with their identical 24-bit significand. A full head-to-head comparison against a Xilinx-IP-based realization of the complete filter system is kept as future work. These results are also stated in the conclusions section (lines 862 to 866 and 872 to 874).

Comments 7: *Analyze the applicable boundaries of LUT-only and DSP-inferred modes in combination with FPGA resource allocation, and give quantitative suggestions for mode selection.*

Response 7: We agree. We have added a quantitative mode-selection rule to the results summary (Section 4.3.5) (lines 751 to 764). It states that the DSP-inferred mode is preferable whenever DSP slices are available (best ATP, two DSP slices per multiply-class unit), while the LUT-only mode of the same source costs only *4–6% more area and 14–32% more ATP in the multiplier ($\approx 2\%$ area and 14% ATP in the MAC)* and should be selected for the $N - \lfloor DSP_{avail}/2 \rfloor$ *units that exceed the available DSP budget... or for all of them when DSP slices must be reserved for other functions, when targeting devices without hardened DSP macros, or when cross-vendor portability is required.*

4. Response to Comments on the Quality of English Language

Point 1: The reviewer indicated that the English is fine and does not require improvement.

Response 1: We thank the reviewer. The newly added text was written consistently with the existing style of the manuscript.

5. Additional clarifications

All revisions made in response to the comments above are highlighted in green in the re-submitted manuscript. We thank the reviewer again for the thorough and constructive feedback.