Tunable Floating Point for High Quality Audio Systems: The Sound of Numbers

G.C. Cardarilli, L. Di Nunzio, R. Fazzolari, R. La Cesa, A. Nannarelli, M. Re

Technical University, Denmark and Univ. Roma “Tor Vergata”, Italy

The Reason

The purpose of this work is to show the advantages of implementing digital signal processing for high quality audio applications in custom floating-point.

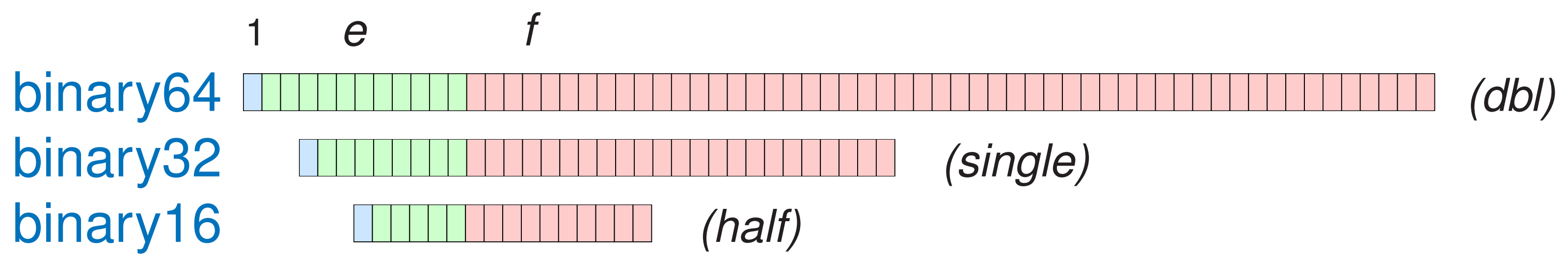
We consider the trade-offs dynamic range vs. precision (i.e., quantization) by comparing standard floating-point (namely, binary32) to custom floating-point.

Moreover, by resorting to Tunable Floating-Point (TFP) hardware units, the dynamic range and the precision can be changed depending on the requirements in different parts of the algorithm.

Results show that 16-bit floating-point formats can give a good compromise between quality and energy efficiency.

Binary Floating-Point Formats

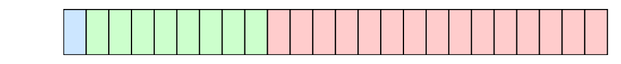
IEEE 754 Standard



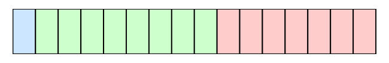
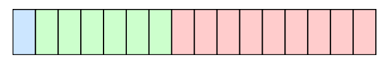
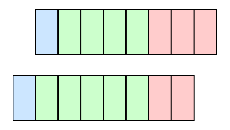
Sign: 1 bit. Exponent: e bits (Bias = $2^{e-1}-1$).

Significand: $m = 1+f$ bits, normalized $1.0 \leq 1.F < 2.0$

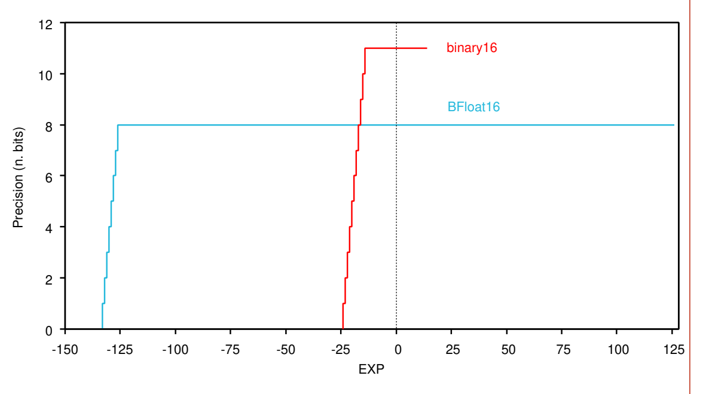
Pixar’s 24-bit Format

PXR24  (storage=24, $e = 8$, $m = 16$)

Formats introduced for Deep Learning

- Google’s BFloat16 (storage=16, $e = 8$, $m = 8$)
  
- IBM’s DLFloat16 (storage=16, $e = 6$, $m = 10$)
  
- *Nvidia’s FP8* two 8-bit formats
  
  (storage=8, $e = 4$, $m = 4$)
  (storage=8, $e = 5$, $m = 3$)
- *Tesla’s CFloat8* same 8-bit formats with custom bias

Precision and Dyn. Range



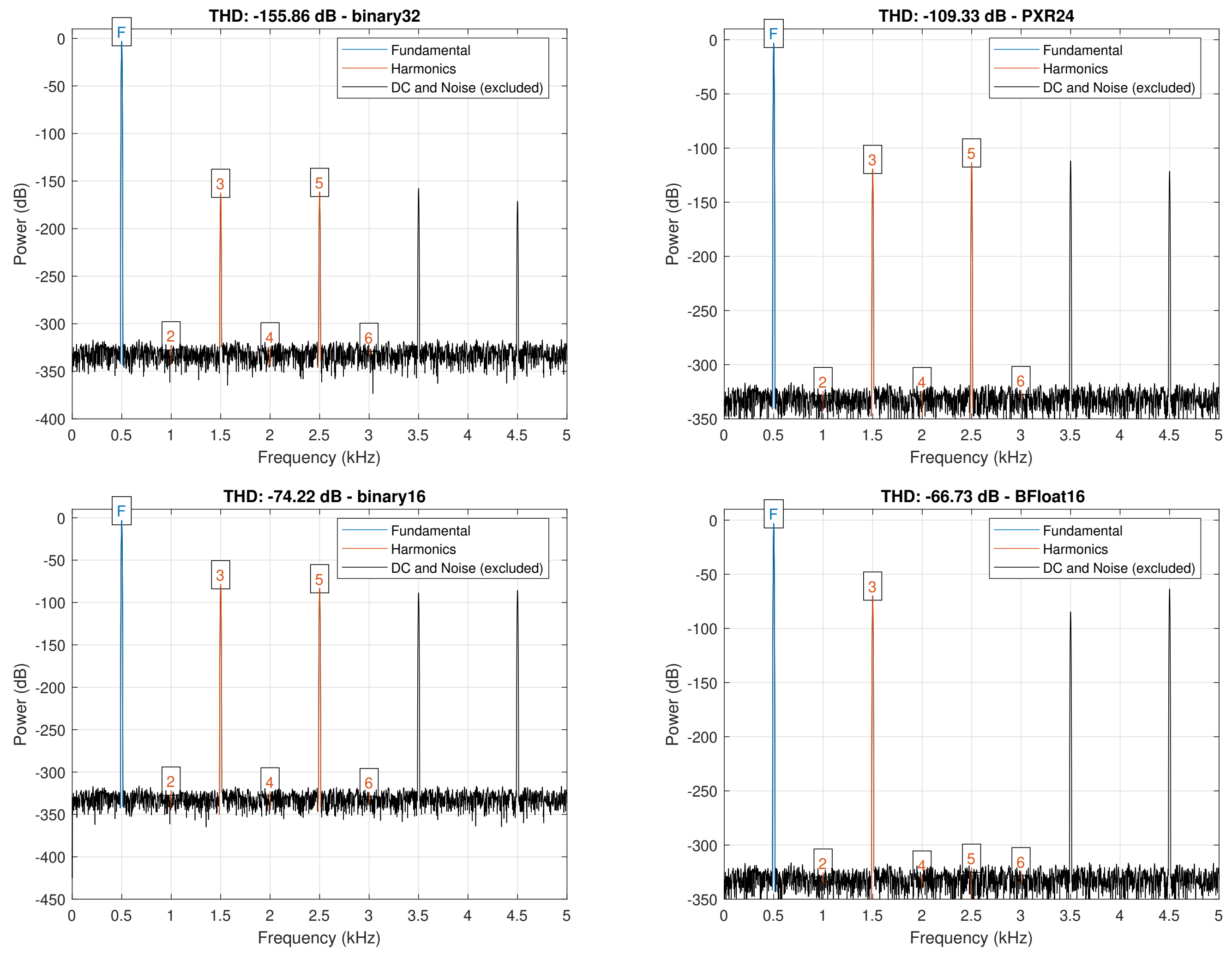
Limited storage formats must compromise between

- precision n. of bits in m
- dyn. range n. of bits in e

The Experiments

1. Total Harmonic Distortion (THD)

THDs and spectra of the different FLP representations for a 500 Hz sinusoidal signal (Kaiser window with $\beta = 38$)



Different FLP formats introduce even and odd order harmonics related to the precision of the representation.

2. Audio Track played through an IIR Filter

An audio track (sample rate 48 KHz, about 9 s.) is filtered by a Butterworth IIR biquad filter ($F_{cut-off} = 10KHz$).

Errors (max. and average) with respect to binary32 for other FP-formats

Format	m	e	ϵ_{max}		ϵ_{ave}	
binary32	24	8	-		-	
PXR24	16	8	1.41×10^{-5}	$< 2^{-16}$	6.93×10^{-7}	$< 2^{-20}$
binary16	11	5	2.98×10^{-4}	$< 2^{-11}$	2.23×10^{-5}	$< 2^{-15}$
BFloat16	8	8	3.02×10^{-3}	$< 2^{-8}$	2.18×10^{-4}	$< 2^{-12}$
FP8 formats	4	4*	3.91×10^{-2}	$< 2^{-4}$	3.43×10^{-3}	$< 2^{-8}$
	3	5	3.91×10^{-2}	$< 2^{-4}$	2.97×10^{-3}	$< 2^{-8}$

* With $e = 4$ about 18% of results are flushed to 0 $\Rightarrow \epsilon_{ave}$ increases.

3. Psychoacoustics Tests for several Audio Tracks

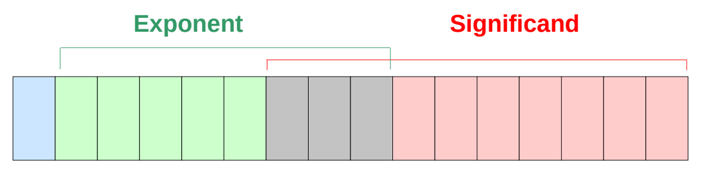
Tests are underway and results will be presented in the final paper.

Tunable Floating-Point Units

TFP32 is 32-bit storage FP-format with adjustable significand and exponent fields bit-width

- significand $m = [3, 24]$ bits (including hidden bit)
- exponent $e = [4, 8]$ bits
- several rounding modes

TFP16 is 16-bit storage TFP-format



- significand fraction $f = [7, 10] \rightarrow m = 1+f$ bits
- exponent $e = [5, 8]$ bits. Customizable bias.

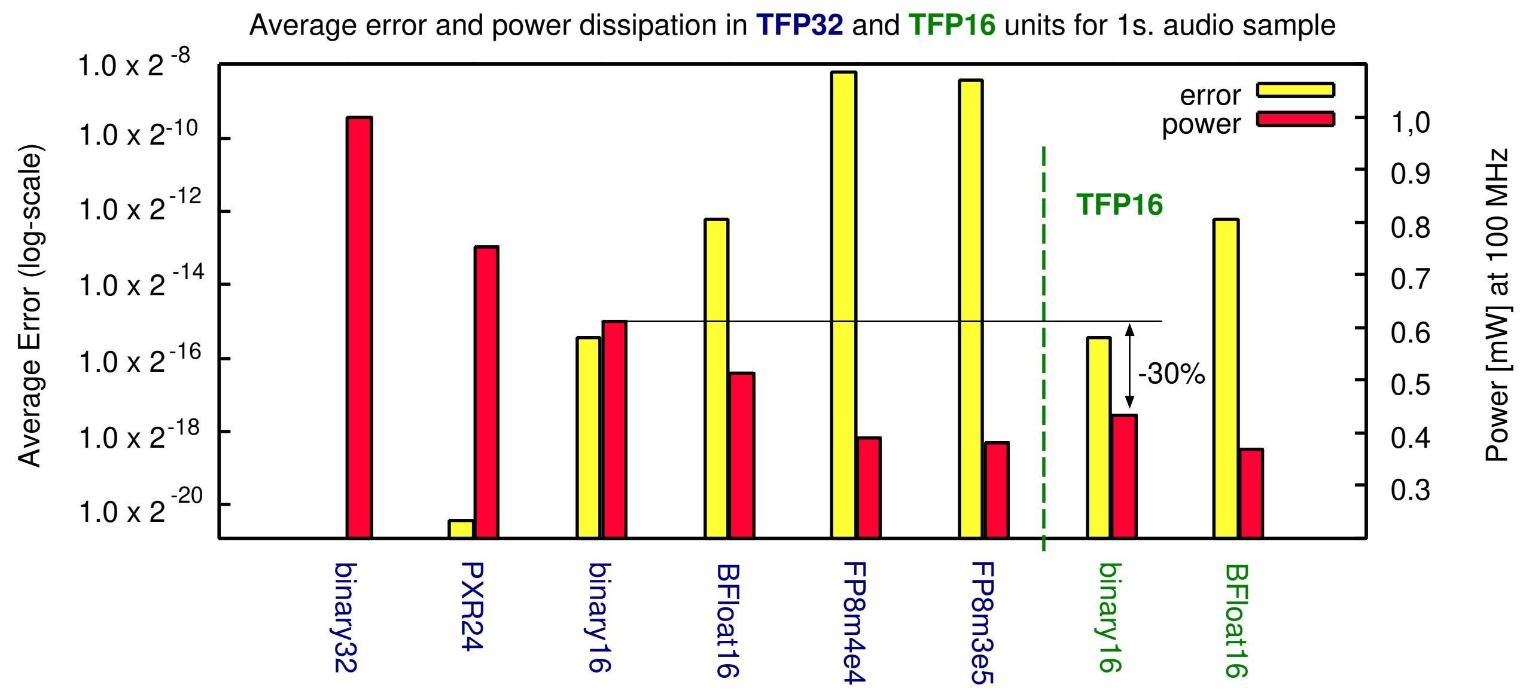
Hardware

Pipelined units implemented in STM 45 nm library of standard cell.

Units		# stages	T_{clk} [ns]	f_{max} [MHz]	Area unit*	Area total*	Area ratio
TFP32	ADD MUL	2 2	1.5	667	6,080 10,250	16,330	1.00
TFP16	ADD MUL/DIV	2 2	1.5	667	1,960 4,020	5,980	0.37

* Area is given in [μm^2]. Area NAND-2 $\simeq 1.06\ \mu m^2$.

Average error and average power dissipation for TFP formats



Summary

- Best error/power trade-off is binary16, followed by BFloat16.
- TFP16 unit is about 30% more power efficient for 16-bit formats.
- The error grows large for FP8 formats ⇒ *“distortion”*.
  The power savings (about 25% vs. BFloat16 in TFP32 unit) are probably not worth the larger errors.

G. C. Cardarilli (`g.cardarilli@uniroma2.it`) 57th Asilomar Conference on Signals, Systems, and Computers Oct. 29 – Nov 1, 2023. Asilomar Conference Grounds, California, USA