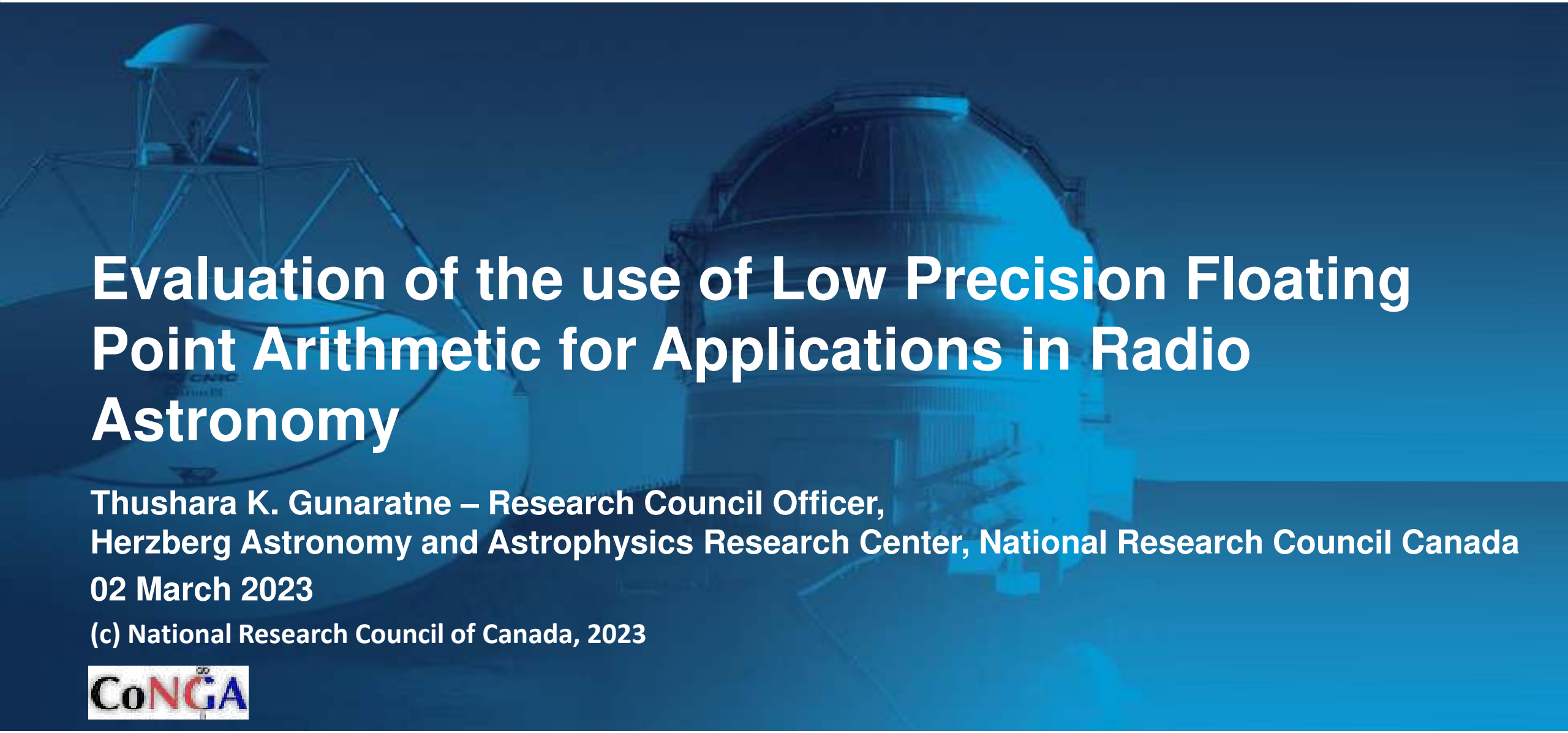


Evaluation of the use of Low Precision Floating Point Arithmetic for Applications in Radio Astronomy

Thushara K. Gunaratne – Research Council Officer,
Herzberg Astronomy and Astrophysics Research Center, National Research Council Canada

02 March 2023

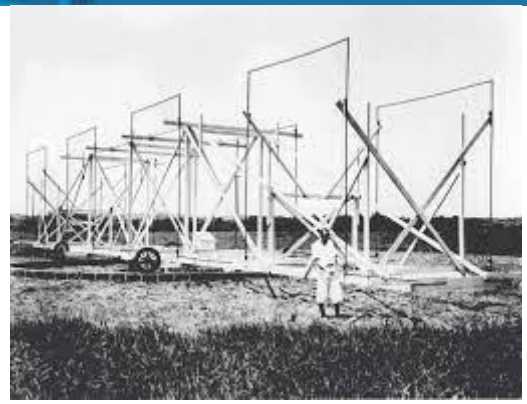
CoNGA

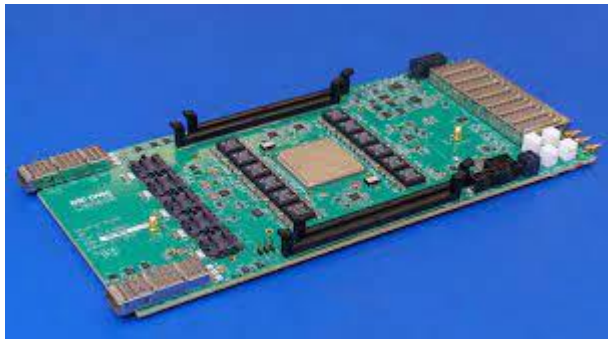
Who We Are….

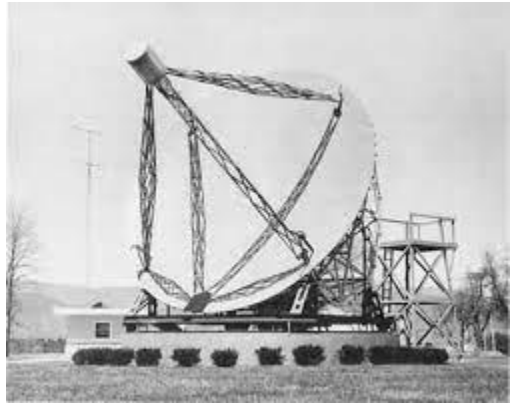
- Herzberg Astronomy and Astrophysics (HAA) Research Center is part of the National Research Council (NRC) Canada
- Two locations
  - Dominion Astrophysical Observatory (DAO) near Victoria, BC
  - Dominion Radio Astrophysical Observatory (DRAO) near Penticton, BC
- Involved in the design/ implementation/observation of many international optical/radio astronomical observatories
- Employs about 40 radio/optical astronomers and 40 engineers and more

Outline

- Historical use of digital-arithmetic in radio astronomy
- The case for *better performance* with floating point arithmetic
- Case Study: The Correlator Beamformer for the Square Kilometre Array Phase-1
  - Key processing modules
  - Considered formats
  - Key performance indecies
  - Simulation results
- Conclusion







Historical use of Digital-Arithmetic in Radio Astronomy

- In the beginning the digital correlators used 1-bit (i.e. two-level) representation!
- The ability to integrate for a *longer duration* with minimum interferences facilitated the imaging process for strong sources.
- The *efficiency* achievable with 1-bit quantization is ~67%.
- Corrections have to be applied to properly quantify the signal strengths from different sources.
- With the advancement of digital electronics the number of bits in the samples grew….
  - 3 - 4 bits in the late 70's
  - 5 - 8 bits in early 2000's
  - 8 – 12 bits, state of the art

TALON DX card : supports (9 + 9b) Correlations

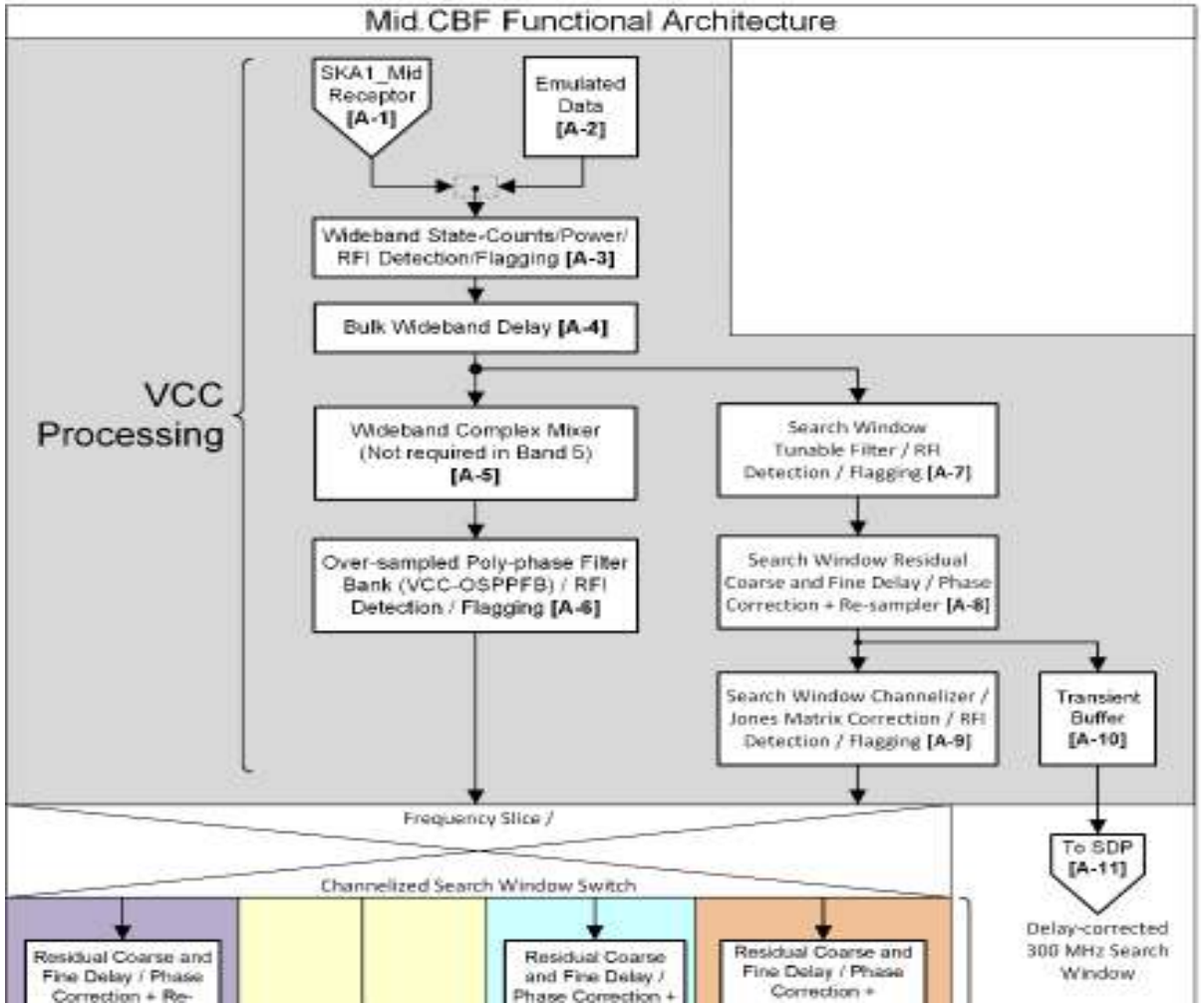
The Case for Higher Overall Performance with Floating Point Arithmetic

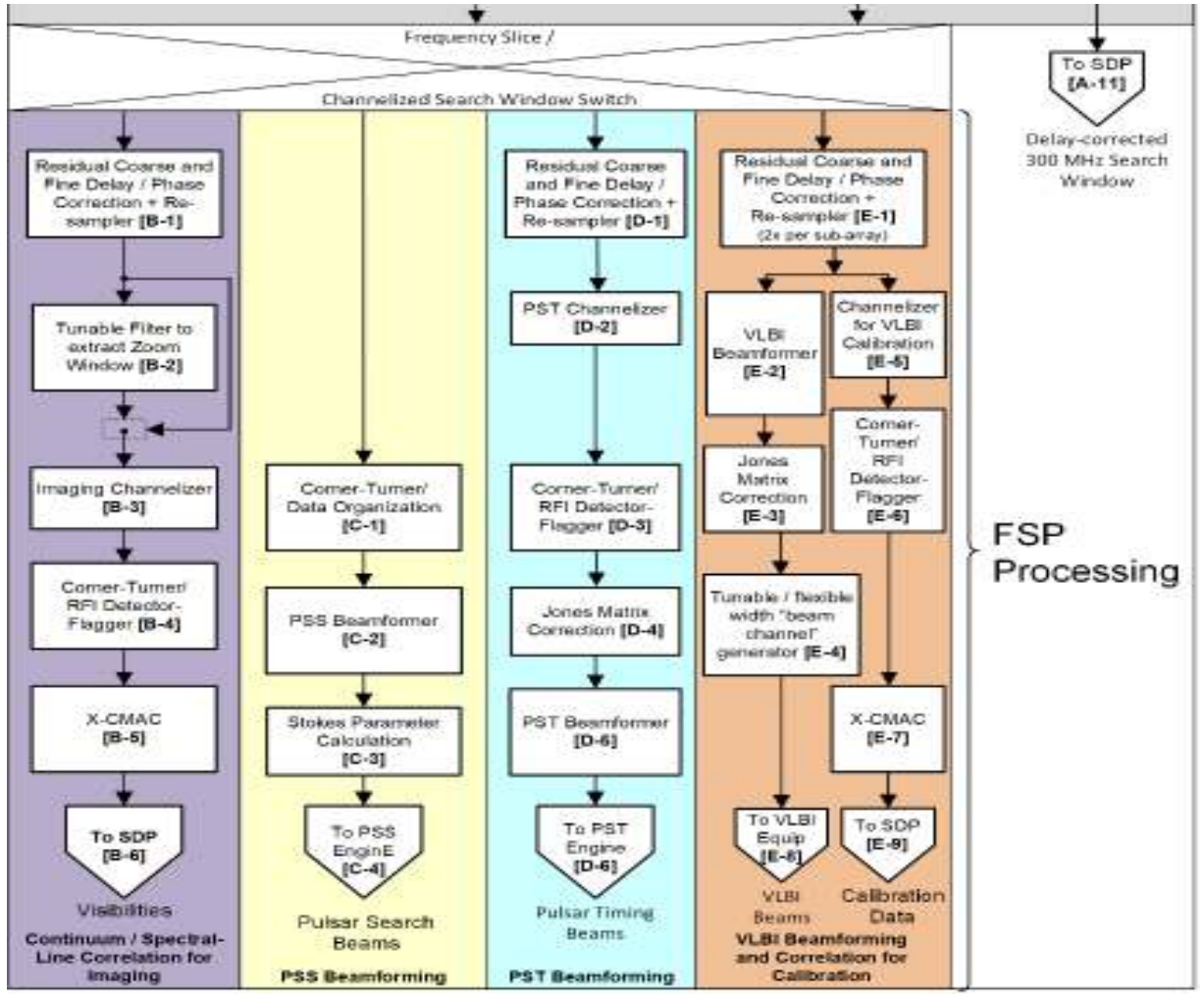
- Radio astronomers' desire to observe fainter sources with better accuracy
- Increasing strength, bandwidth and occurrence of Radio Frequency Interference (RFI)
  - Quantized strong RFI signals can lead to spectral-confusion
- This requires increase in the sample widths resulting in higher power, area and longer design times
- Input signals are 'Gaussian distributed'; 25 – 50 % toggle rate
- Straight forward way to handle the increased signal dynamic range → Floats
- Key FPGA/GPU vendors began to support 'native' low-precision floating point formats (e.g. Float16, BFloat16 and BFloat16+ (aka NVIDIA TensorFloat32))

Case Study:
Correlator Beamformer for the Square Kilometre Array

- The FX-type TALON frequency-slice architecture has been down selected for the correlator and beamformer (CBF) of the Square Kilometre Array (SKA), mid frequency radio telescope.
- Construction of the phase-1 started in December 2022 in the ‘Karoo’ region of South Africa.
- It will consisted of 197 dishes with diameters up to 15 m spanning 150 km
- Observes the spectrum 0.35 – 15.3 GHz with 6 overlapping bands. The maximum instantaneous bandwidth is 5 GHz.
- Facilitates multiple simultaneous observation modes.

Functional Architecture: SKA1 Mid CBF





Key Signal Processing Modules in SKA1 Mid CBF

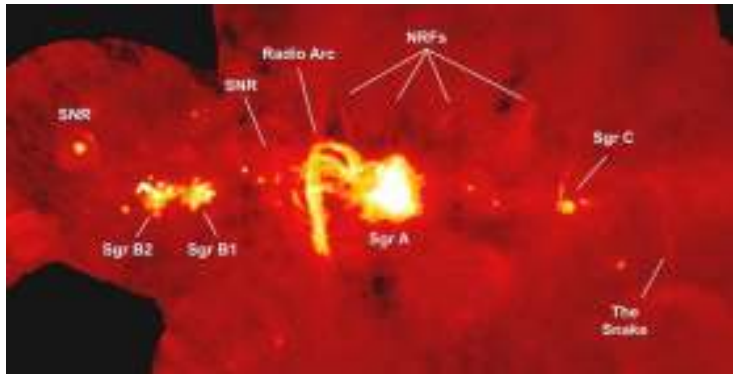
- Channelizers – To segment the input bandwidth into uniform channels
- Tunable Filters – To extract a part of the input bandwidth
- ReSamplers – To change the sample rate, apply delay corrections
- Complex-mixers – To shift the spectrum signals
- Cross - Complex Multiply and Accumulator – To evaluate auto- and cross-correlations
- Beamformers – To selectively enhance signals depending on their direction of arrival
- 2x2 matrix-multipliers – To apply corrections to polarization mismatches

The Considered Low Precision Floating Point Formats

Format	Specs	Minimum	Maximum	Comments
float16	1S : 5E : 10M	$2^{-14} \approx$ 6.1035e-5	65504	Supported in Intel Agilex Variable Precision DSP Block
bfloat16	1S : 8E : 7M	$2^{-126} \approx$ 1.1755e-38	3.390e+38	
bfloat16+ (NVIDIA TensorFloat32)	1S : 8E : 10M	$2^{-126} \approx$ 1.1755e-38	3.400e+38	
posit16	1S : 1ES	$4^{-14} \approx$ 3.7253e-9	$4^{14} \approx$ 268,435,456	
fixed19	2's complement	$2^{-18} \approx$ 3.8147e-6	$1-2^{-18} \approx$ 0.9999962	Normalized range [-1, 1]

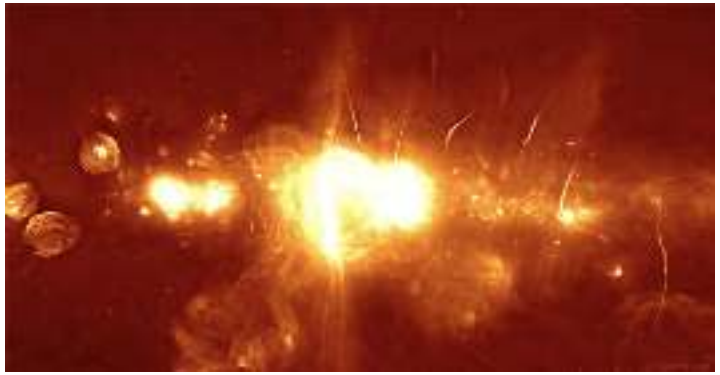
Qualitative Study:
The Frequency Responses of Key Signal Processing Modules

- In terms of the imaging dynamic range, the performance of a radio telescope depends on the ability to suppress the interference from celestial and other sources.
- For the SKA1 Mid telescope, the targeted imaging dynamic range is 1 : 1 000 000
- This translated to > 60 dB suppression of accumulated interferences power
- The integrated stopband attenuation should exceed 60 dB!



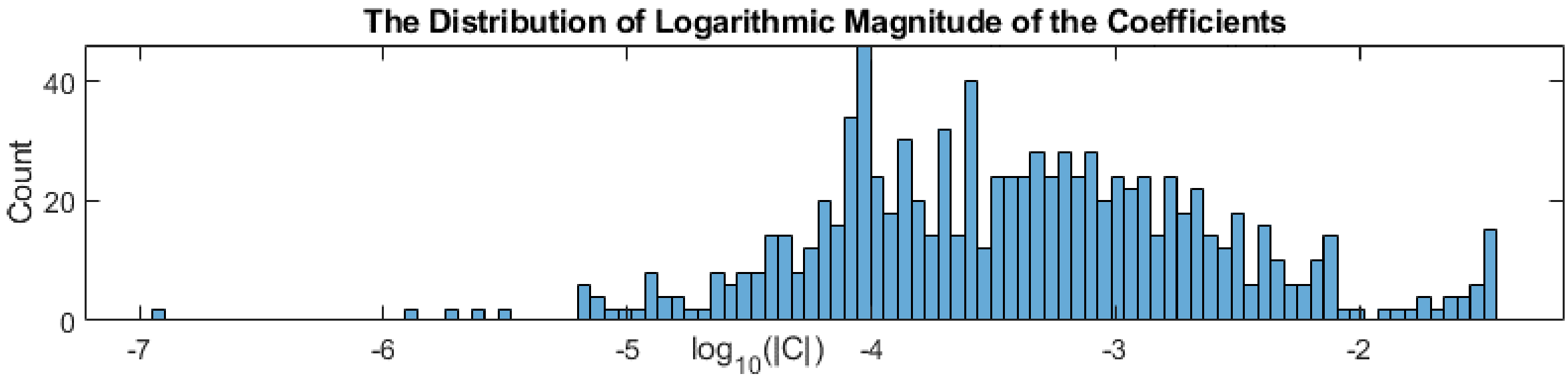
Very Large Array (VLA) 2004

Galactic Center of the Milky Way galaxy

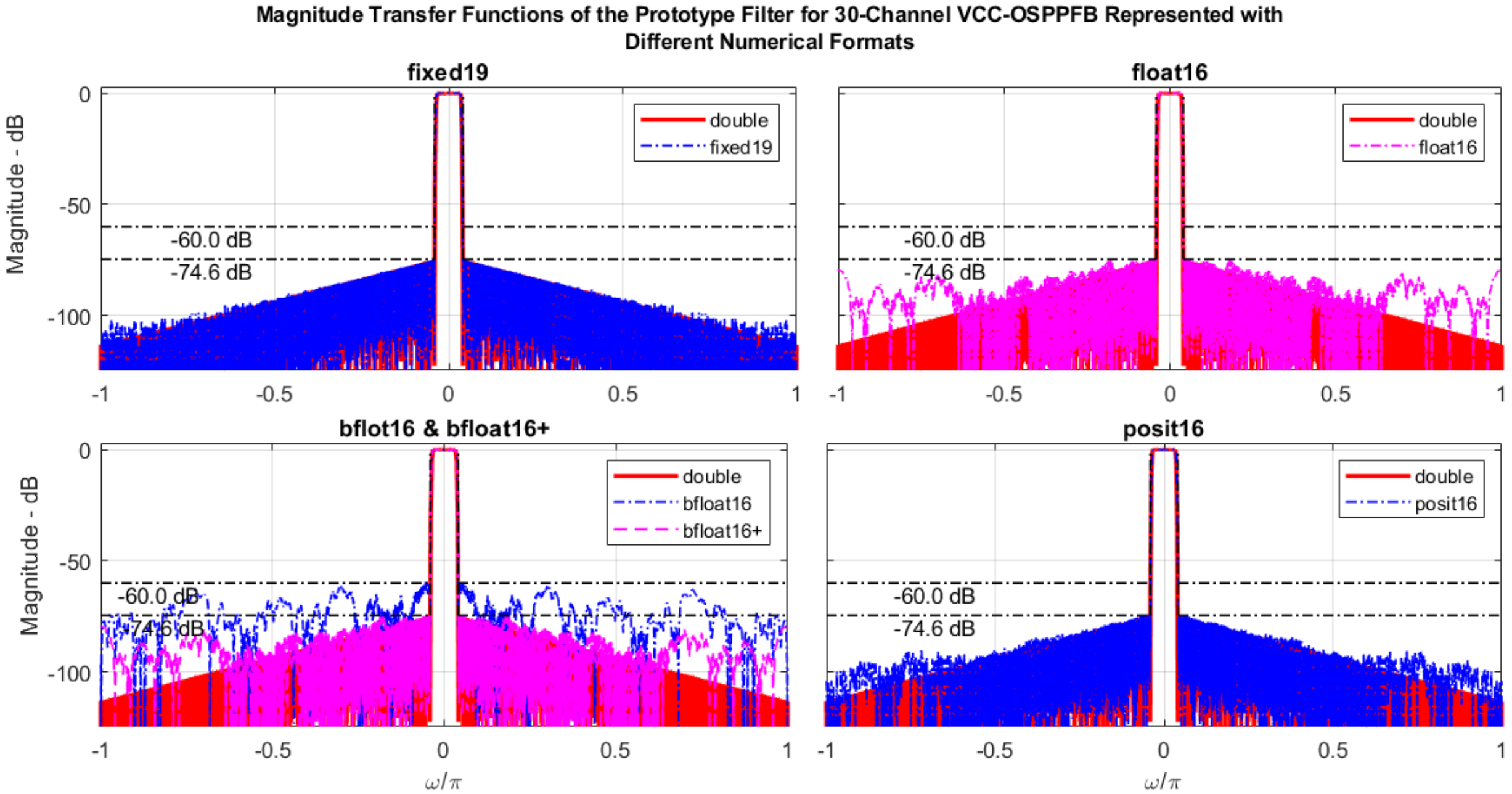


MeerKAT 2019

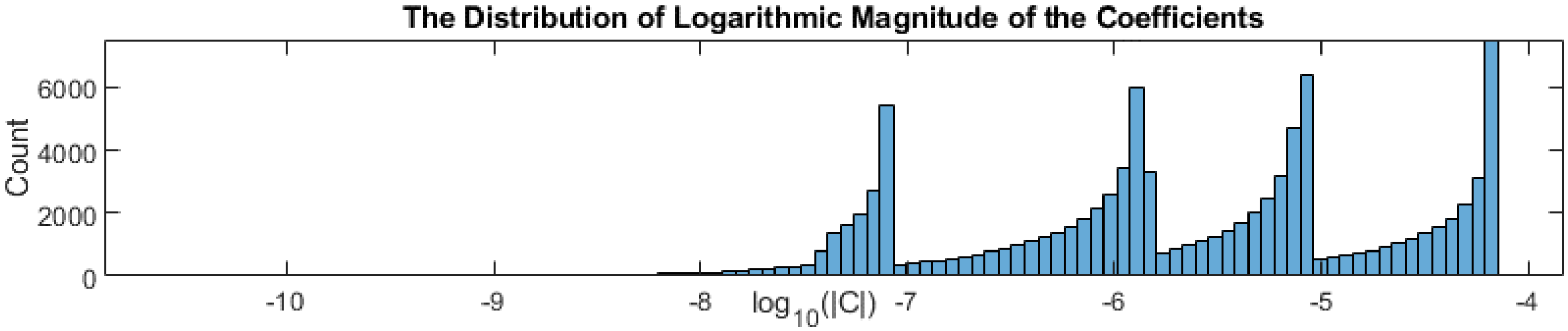
Frequency Responses of the 30 Channel Coarse-Channelizer Represented with float16, bfloat16, bfloat16+, posit16 and fixed19 Format



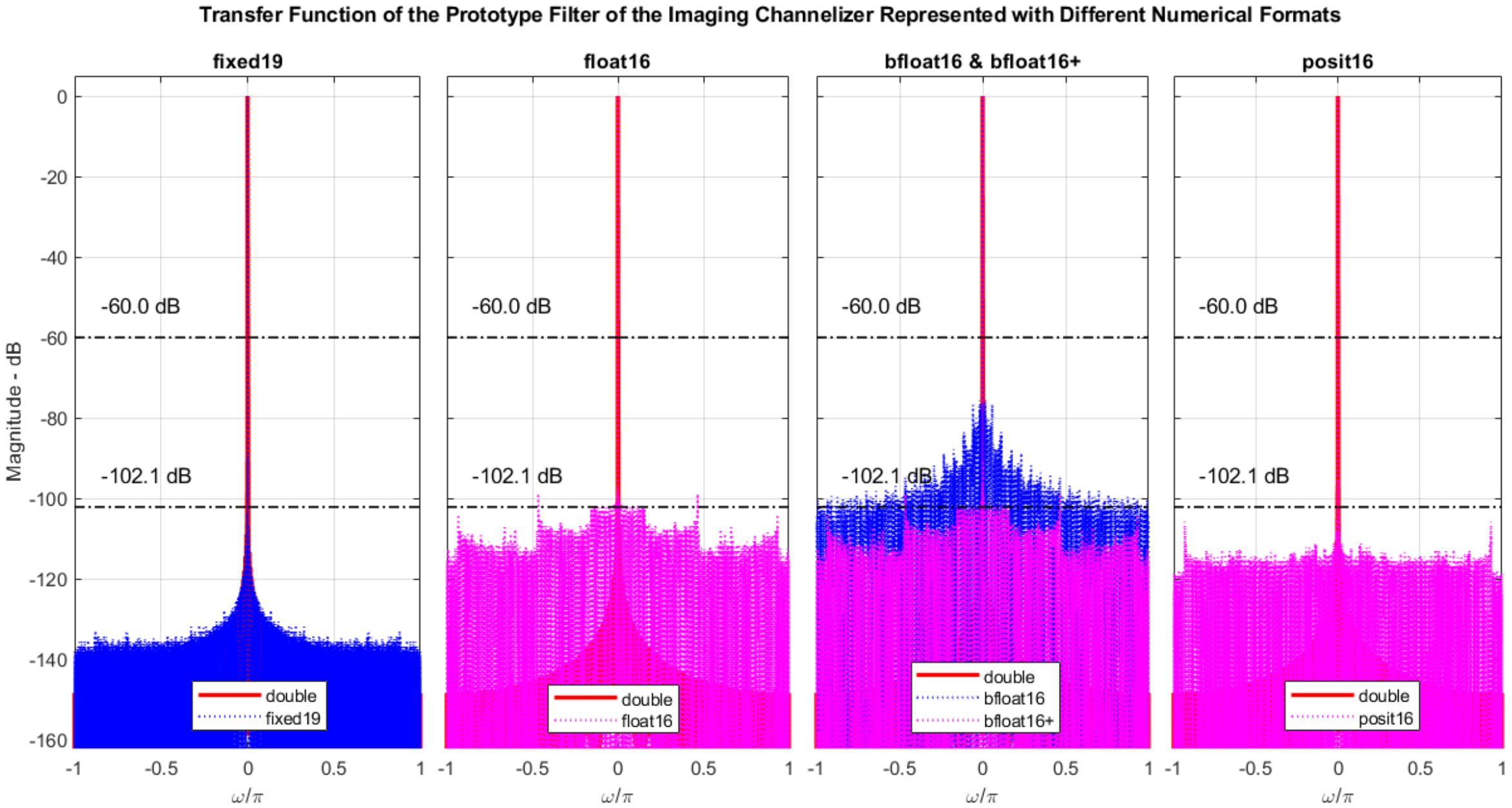
Applied Scaling Factors				
float16	bfloat16	bfloat16+	posit16	fixed-19
2^{10}	2^{10}	2^{10}	2^{6}	2^{4}


Magnitude Transfer Functions of the Prototype Filter for 30-Channel VCC-OSPPFB Represented with Different Numerical Formats
fixed19
float16
bflot16 & bfloat16+
posit16
Magnitude - dB
0
-50
-100
-60.0 dB
-74.6 dB
double
fixed19
float16
bfloat16
bfloat16+
posit16
-1
-0.5
0
0.5
1
ω/π

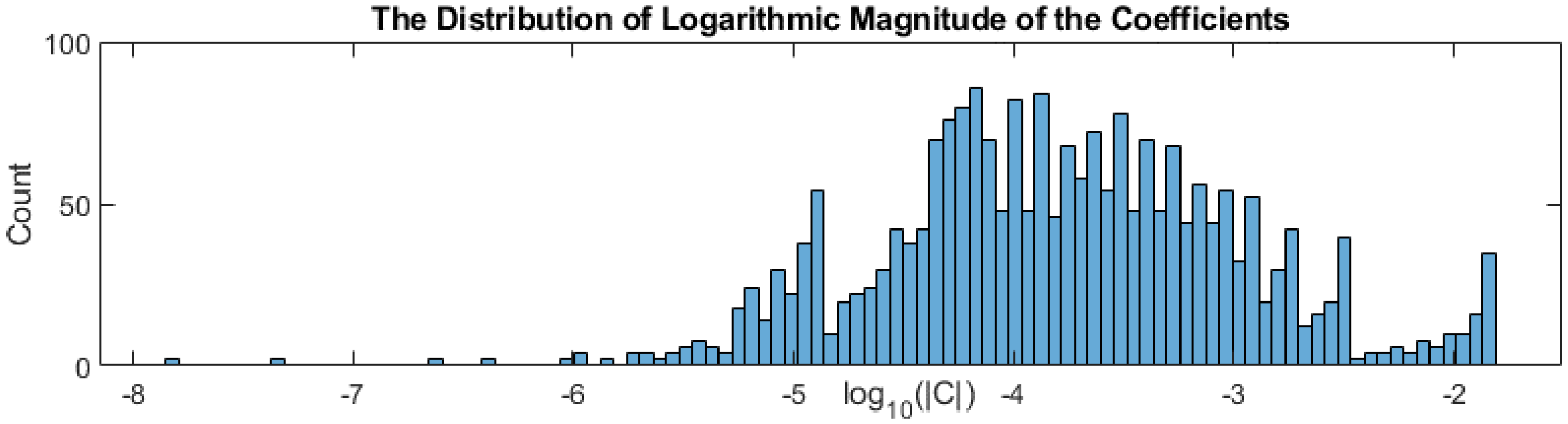
Frequency Responses of the 16,384 Channel Imaging-Channelizer Represented with float16, bfloat16, bfloat16+, posit16 and fixed19 Formats



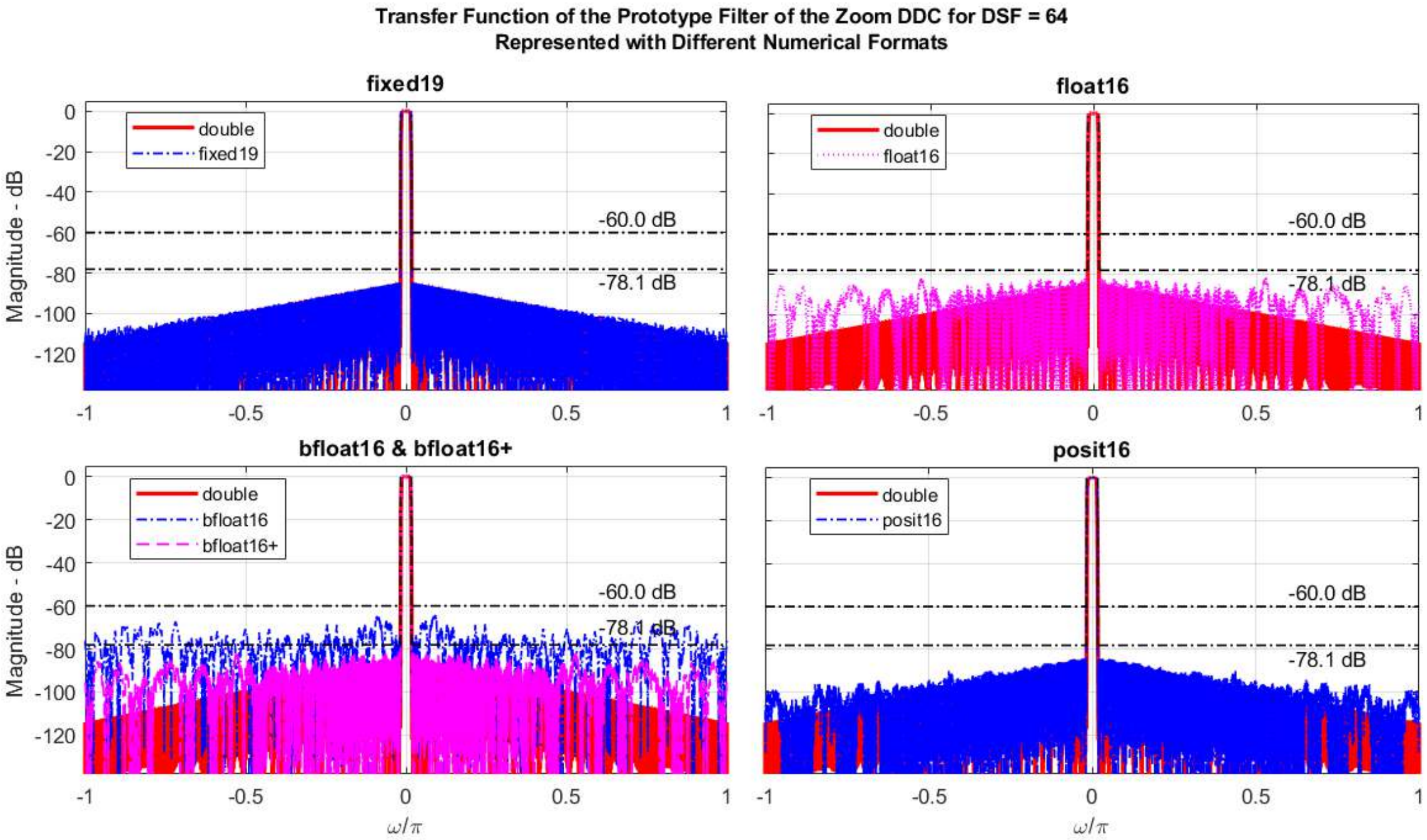
Applied Scaling Factors				
float16	bfloat16	bfloat16+	posit16	fixed-19
2^{20}	2^{20}	2^{20}	2^{12}	2^{13}


Transfer Function of the Prototype Filter of the Imaging Channelizer Represented with Different Numerical Formats
fixed19
float16
bfloat16 & bfloat16+
posit16
Magnitude - dB
0
-20
-40
-60
-80
-100
-120
-140
-160
-60.0 dB
-102.1 dB
double
fixed19
float16
bfloat16
bfloat16+
posit16
-1
-0.5
0
0.5
1
ω/π

Frequency Responses of the Tunable Filter for 64-fold Down-Sampling Represented with float16, bfloat16, bfloat16+, posit16 and fixed19 Formats



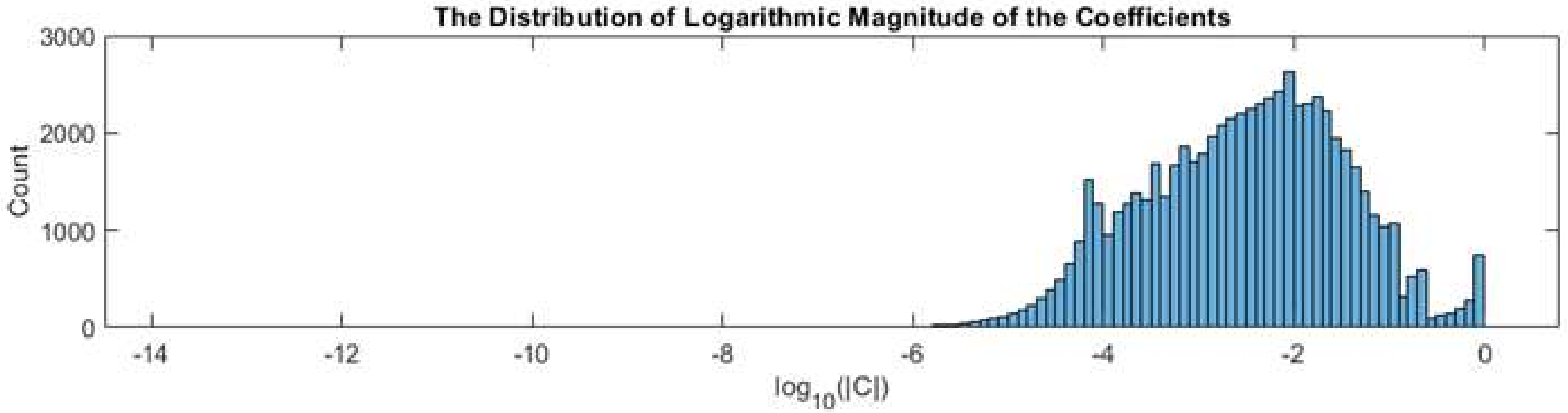
Applied Scaling Factors				
float16	bfloat16	bfloat16+	posit16	fixed-19
2^{16}	2^{16}	2^{16}	2^{8}	2^{6}


Transfer Function of the Prototype Filter of the Zoom DDC for DSF = 64
Represented with Different Numerical Formats
fixed19
double
fixed19
-60.0 dB
-78.1 dB
float16
double
float16
-60.0 dB
-78.1 dB
bfloat16 & bfloat16+
double
bfloat16
bfloat16+
-60.0 dB
-78.1 dB
posit16
double
posit16
-60.0 dB
-78.1 dB
Magnitude - dB
0
-20
-40
-60
-80
-100
-120
-1
-0.5
0
0.5
1
ω/π

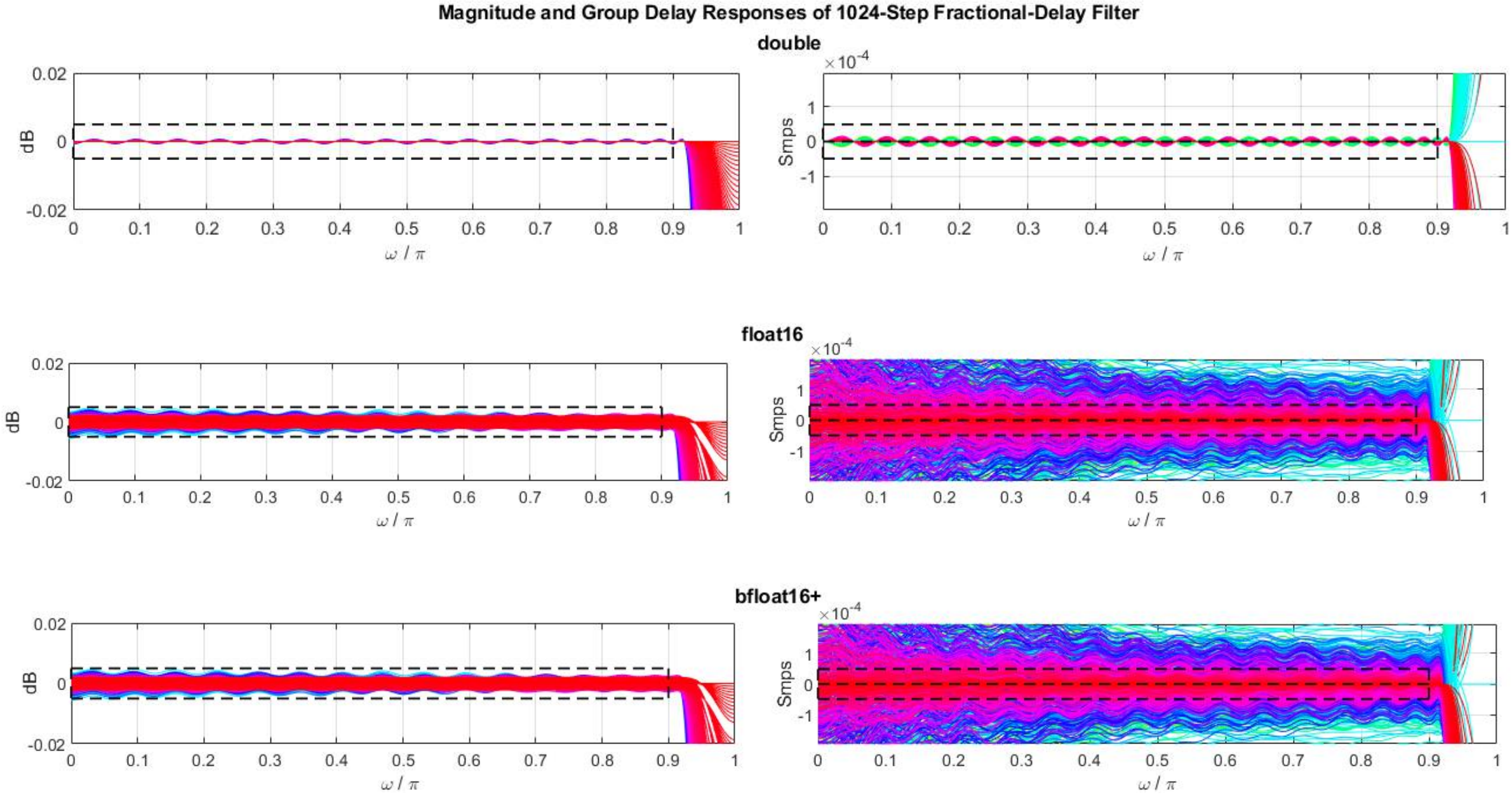
Frequency Responses of the Fractional Delay Filter of the ReSampler Represented with float16, bfloat16, bfloat16+, posit16 and fixed19 formats

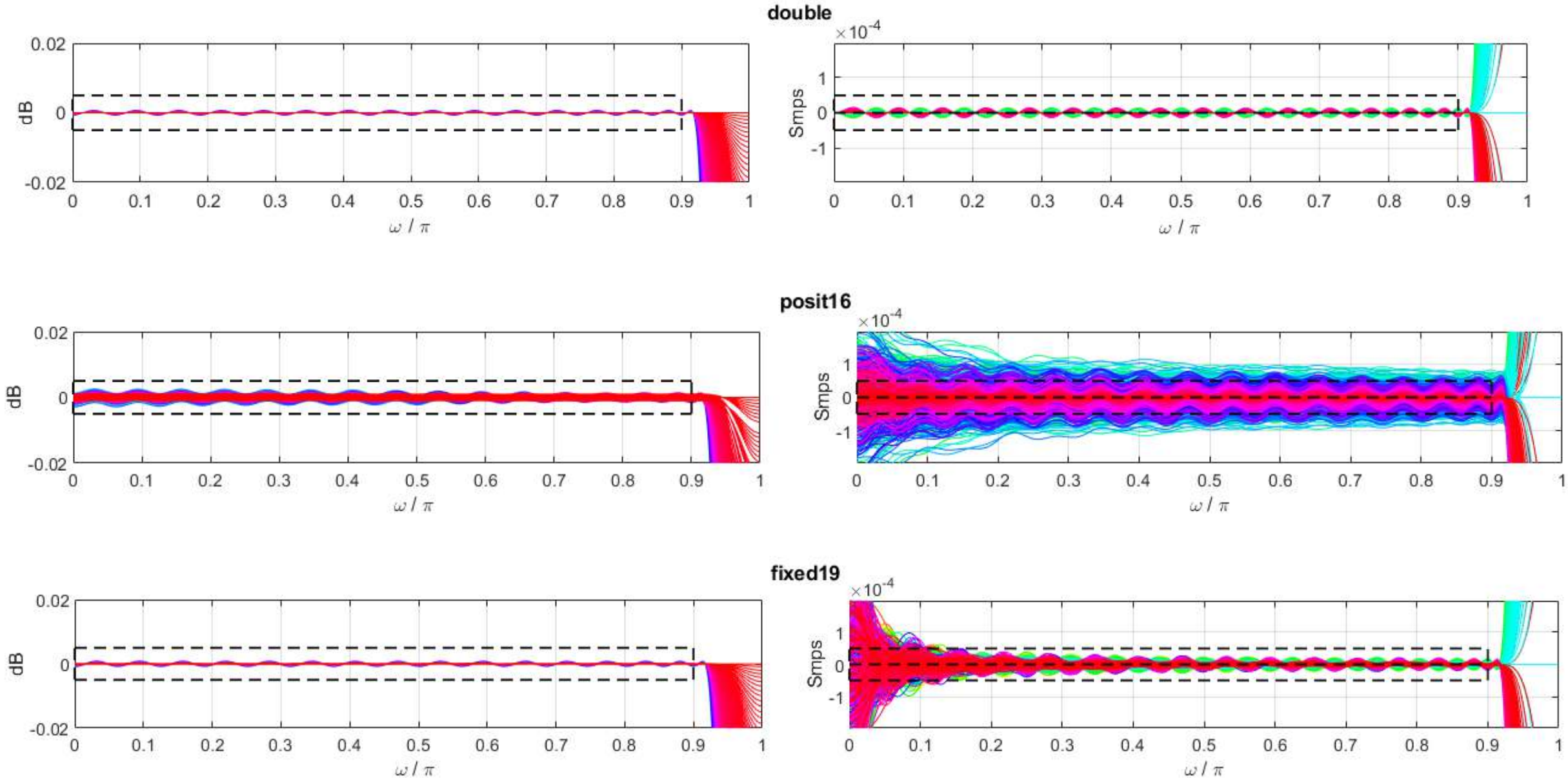
- In ReSamplers, the passband ripple and the group-delay response are of concern.
- Unlike other signal processing block in the signal chain of Mid.CBF passband ripple of the ReSampler can not be calibrated out
- The variations in the group-delay response lead decorrelation and artefacts in the synthesized images.
- In Mid.CBF, a 1024 step fractional-delay filter-bank has been used in the ReSampler
- The maximum passband ripple should be ≤0.005 dB and the maximum error in group delay should *ideally* be ≤0.0005 sample-period (i.e. ≤ 0.5 delay-step)

Frequency Responses of the Fractional Delay Filter of the ReSampler Represented with float16, bfloat16, bfloat16+, posit16 and fixed19 formats



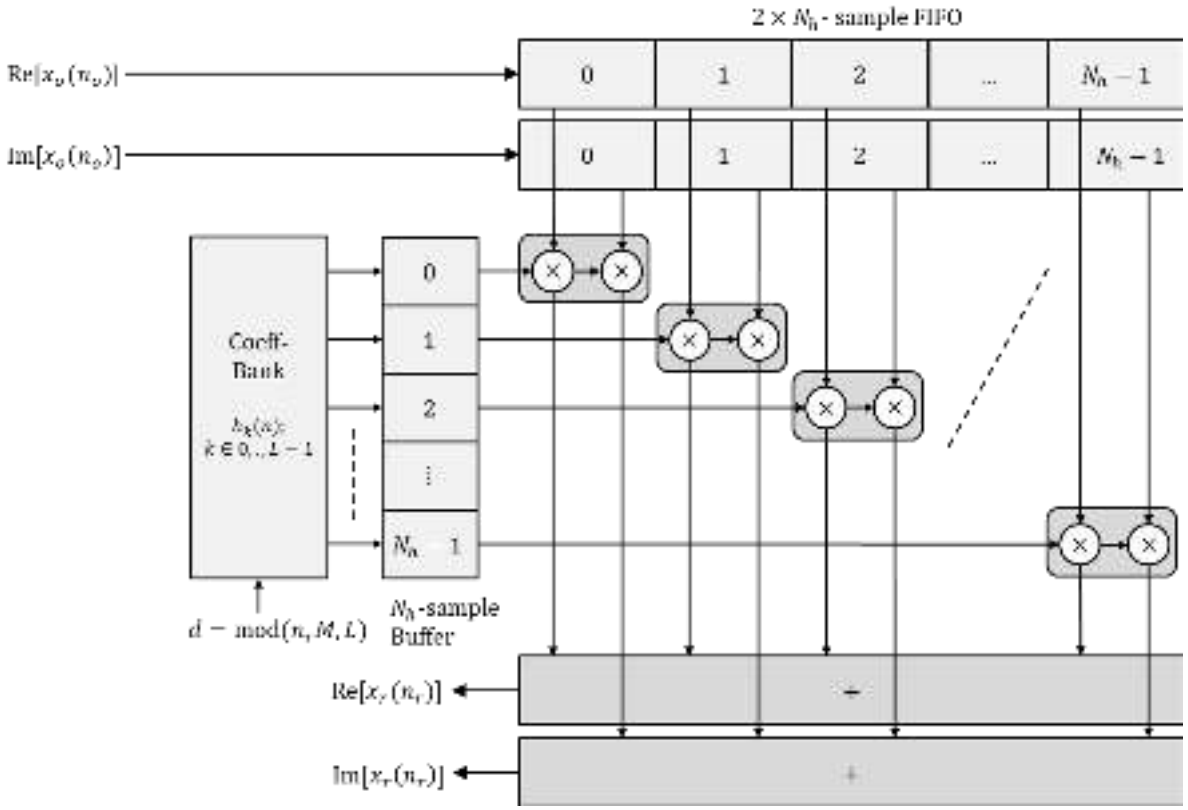
Applied Scaling Factors			
float16	bfloat16+	posit16	fixed-19
2^4	2^4	2^4	1


Magnitude and Group Delay Responses of 1024-Step Fractional-Delay Filter
double
float16
bfloat16+
dB
Smps
×10^-4
ω / π

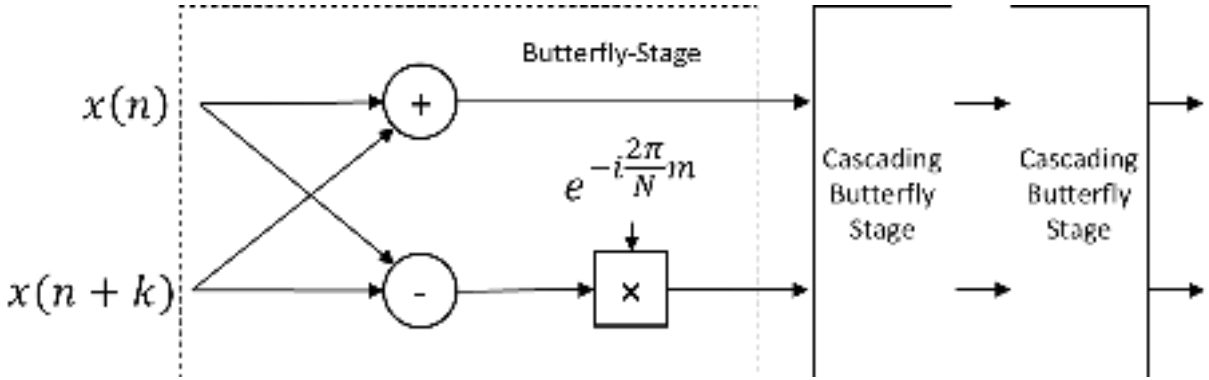

Magnitude and Group Delay Responses of 1024-Step Fractional-Delay Filter
double
posit16
fixed19
dB
Smps
ω / π
×10^-4

Quantitative Study: Key Arithmetic Operations

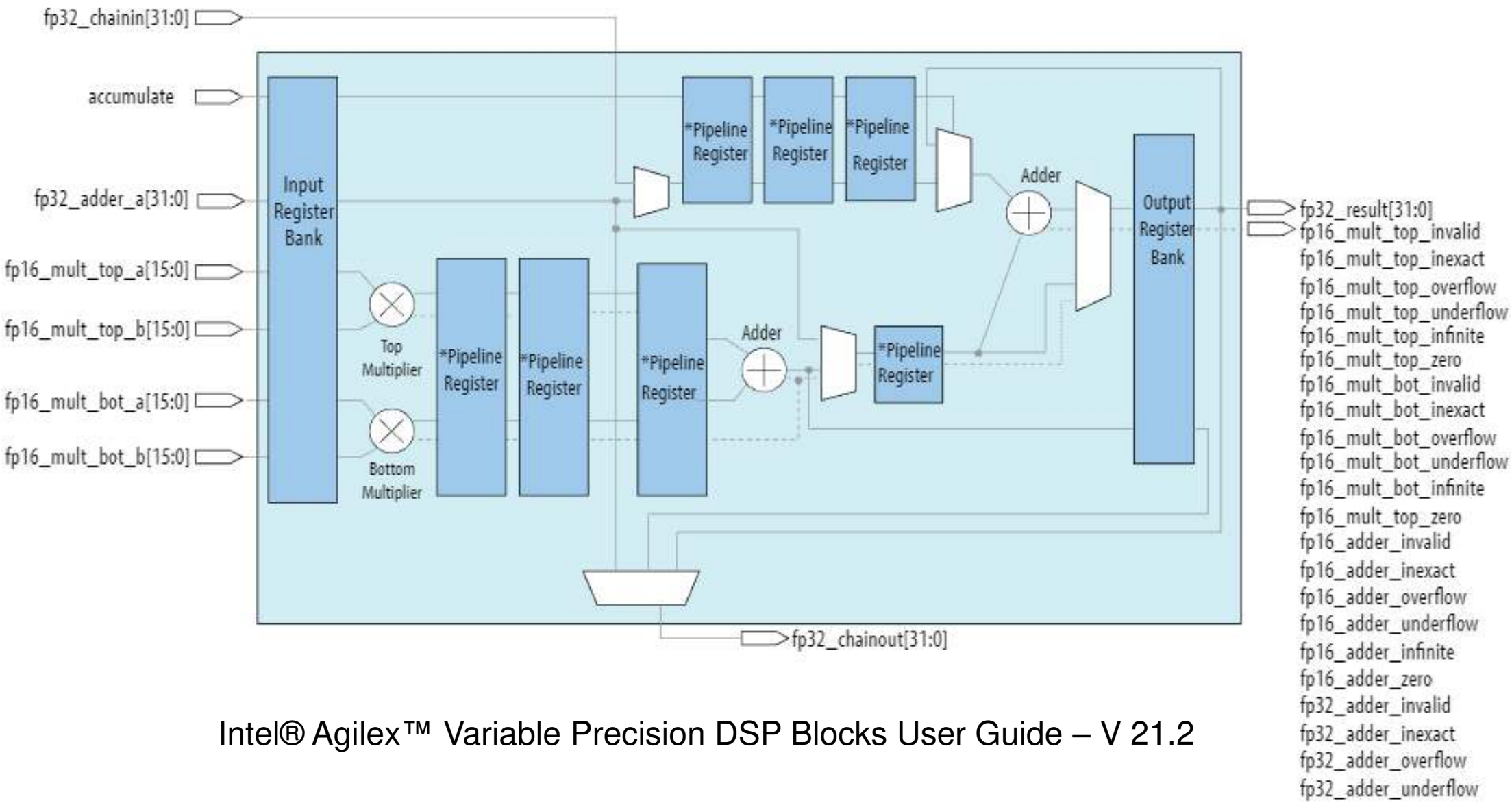
- Multiply and Add (e.g. in FIR filters)



- Chain Multiplications (e.g. FFT)



Floating-point Arithmetic 16-bit Half-Precision Mode



Intel® Agilex™ Variable Precision DSP Blocks User Guide – V 21.2

*This block diagram shows the functional representation of the DSP block. The pipeline registers are embedded within the various circuits of the DSP block.

Summary

- Vendors are now supporting low-precision floating-point arithmetic in FPGAs and GPUs.
- A preliminary qualitative and quantitative study has been conducted to see whether float16, bfloat16, bfloat16+ and posit16 formats can be used for the implementation of signal processing modules for SKA1 Mid CBF
- For coarse and imaging channelizers and tunable filters, float16, bfloat16+ and posit16 formats are capable of achieving the desired attenuation
- For the ReSampler, the qualitative study is inconclusive on whether it meets the accuracy in delay/phase
- “The quantitative study on the FIR filters implies the float16, bfloat16+ and posit16 are capable of achieving the desired SNR”
- “For FFTs, the data has to be converted for float32 to achieve the desired SNR”

Thank You!

Questions?