International Institute of Information Technology
Bangalore

VLSI PROJECT ELECTIVE

VARIABLE-PRECISION APPROXIMATE FLOATING-POINT MULTIPLIER

B Sathiya Naraayanan(IMT2020534)

Iswarya I(MT2023530)

iiit-b

ज्ञानमुत्तमम्

INTRODUCTION

The Variable-precision approximate floating-point multiplier is proposed for energy efficient deep learning computation. The proposed architecture supports approximate multiplication with BFloat16 format.The approximation is done in form of truncation.

ARCHITECTURE

The inputs are divided into sign bits,exponent and mantissa bits seperately. There are three block rams used,one for each inputs and one for the output. Sign and exponent module is used for getting the sign and exponent of the multiplied result.The calculated exponent is used for creating mask which is used to decide the number of bits to be truncated.This is done using precision control module.The mantissas' are given as input to booth multiplier after appending '01' bits.The '1' bit is the implied bit that comes before the decimal point.The booth muliplication algorithm uses input in signed format, so '0' bit is added.All the outputs are processed and normalized for final output in the normalization module.

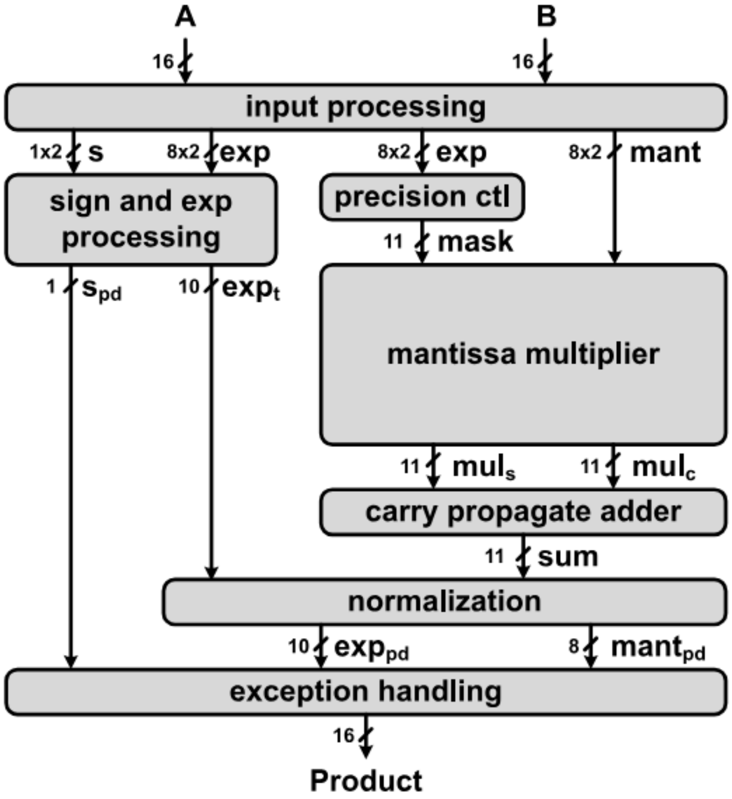


Figure 1: Multiplier Architecture

IMPLEMENTATION

The proposed architecture was implemented in Verilog for different data formats, namely bfloat16, TensorFloat32, posit, IEEE half precision, and IEEE single precision, and deployed on a Xilinx Basys-3 FPGA board using Vivado. The multiplier was implemented without masking and a precision control unit for comparing the results.
The posit datatype implemented was n=16,es=4(regime)

Data Formats	Sign	Exponent	Mantissa
bfloat16	1	8	7
Tensorfloat32	1	8	10
IEEE Half	1	5	10
IEEE Single	1	8	23
Posit	1	Not Fixed	Not Fixed

FPGA and ASIC RESULTS

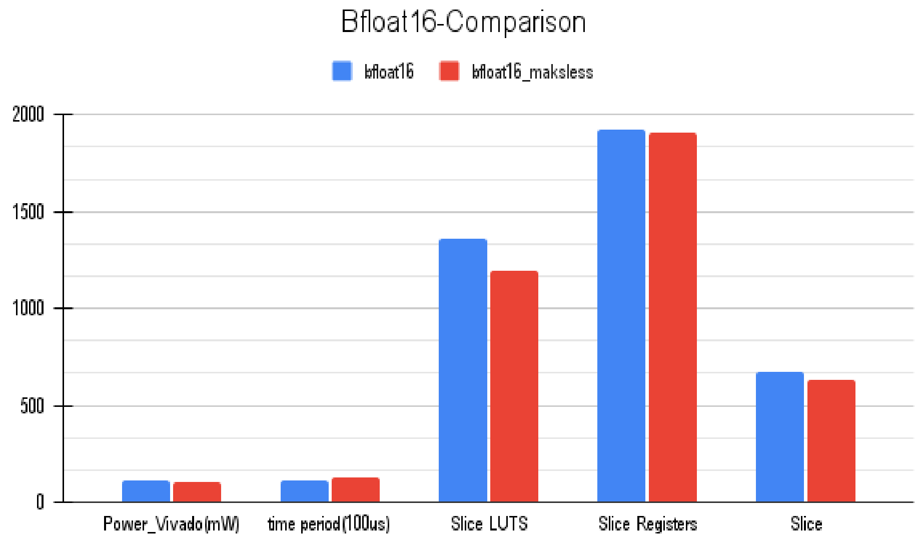


Figure 2: Bfloat16-Comparison

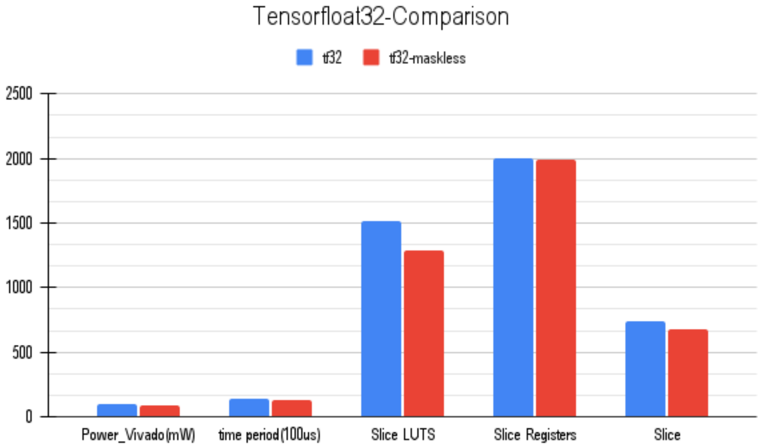


Figure 3: Tensorfloat32-Comparison

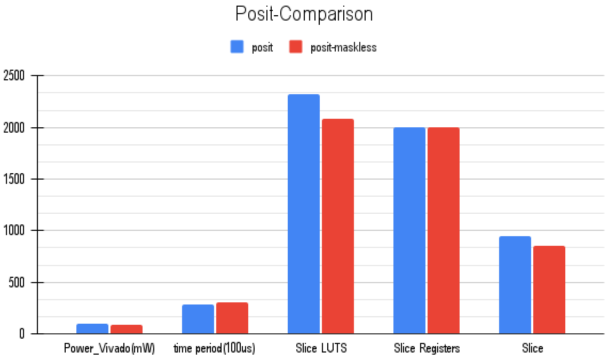


Figure 4: Posit-Comparison

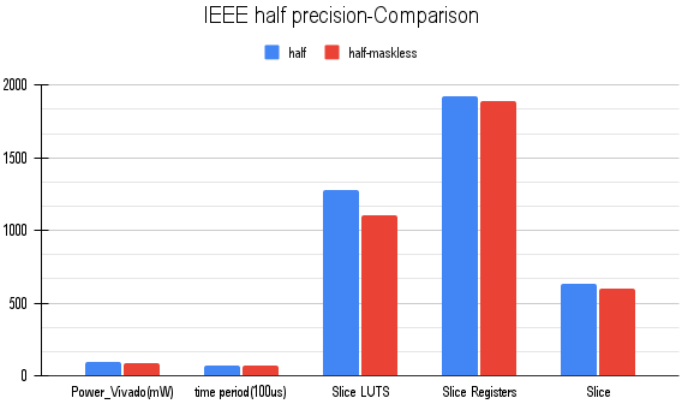


Figure 5: IEEE Half Precision-Comparison

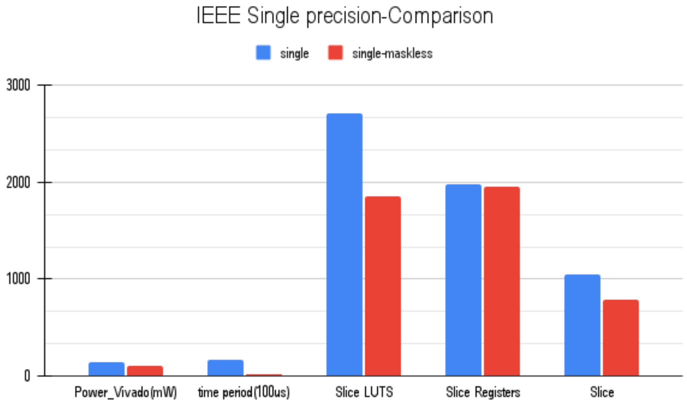


Figure 6: IEEE Single Precision-Comparison

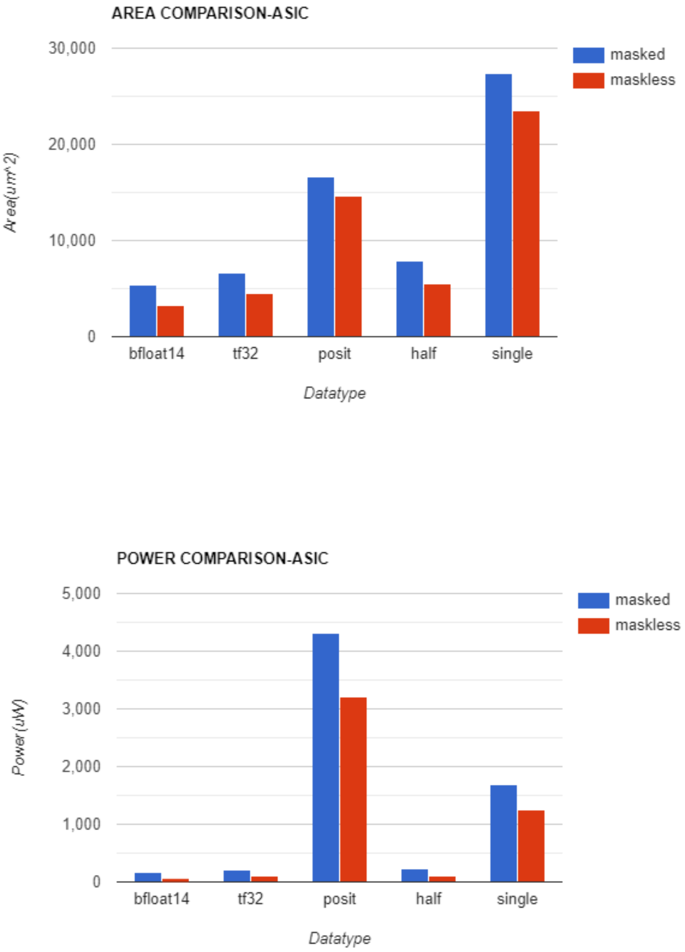
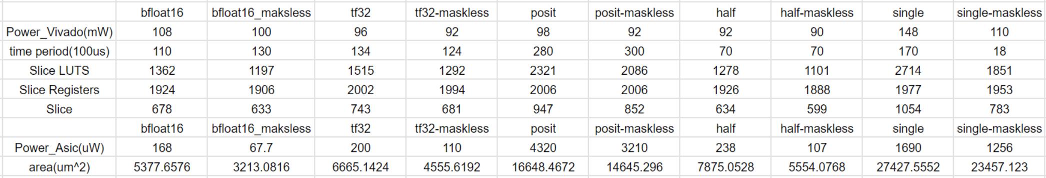


Figure 7: Power-Comparison



	bfloat16	bfloat16_maksless	tf32	tf32-maskless	posit	posit-maskless	half	half-maskless	single	single-maskless
Power_Vivado(mW)	108	100	96	92	98	92	92	90	148	110
time period(100us)	110	130	134	124	280	300	70	70	170	18
Slice LUTS	1362	1197	1515	1292	2321	2086	1278	1101	2714	1851
Slice Registers	1924	1906	2002	1994	2006	2006	1926	1888	1977	1953
Slice	678	633	743	681	947	852	634	599	1054	783
	bfloat16	bfloat16_maksless	tf32	tf32-maskless	posit	posit-maskless	half	half-maskless	single	single-maskless
Power_Asic(uW)	168	67.7	200	110	4320	3210	238	107	1690	1256
area(um^2)	5377.6576	3213.0816	6665.1424	4555.6192	16648.4672	14645.296	7875.0528	5554.0768	27427.5552	23457.123

Figure 8: Results

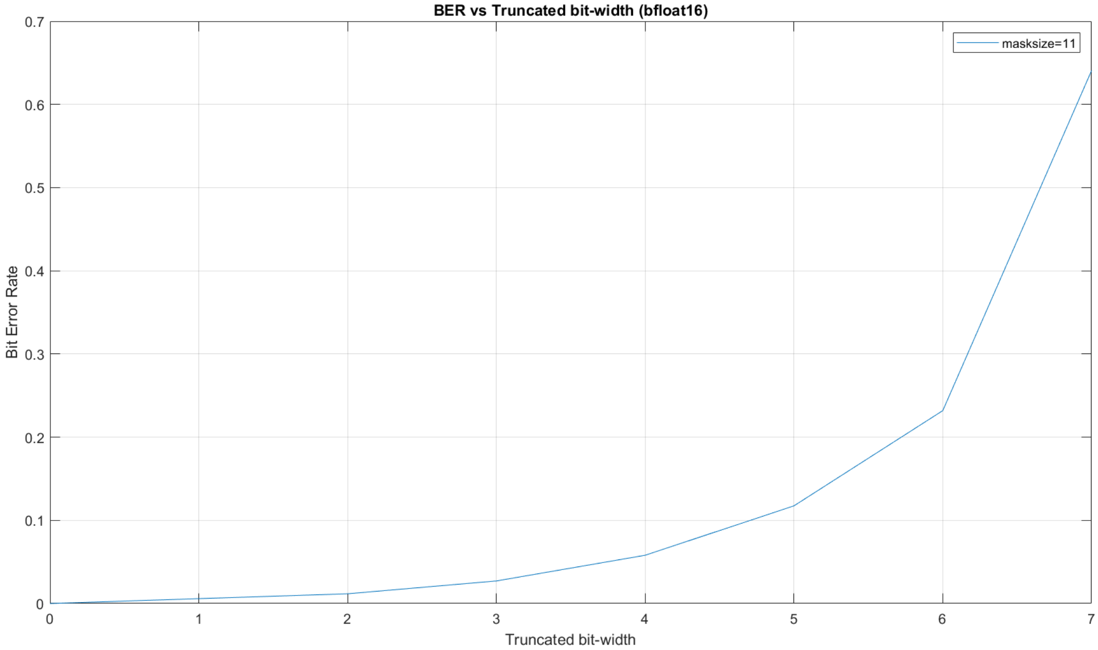


Figure 9: Bfloat16

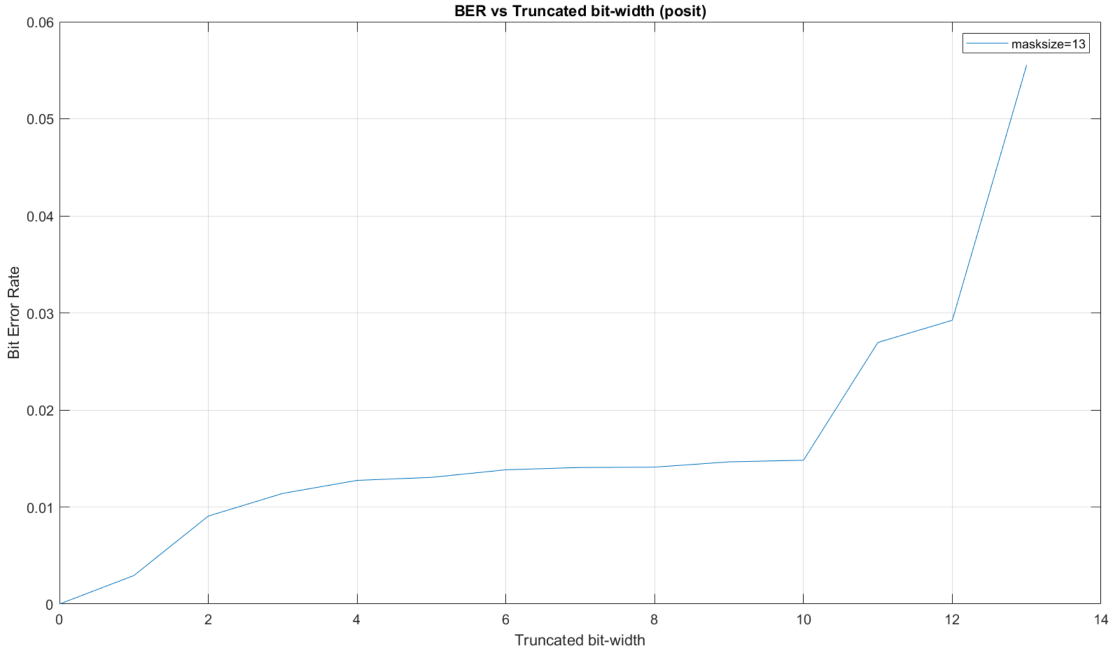


Figure 10: Posit

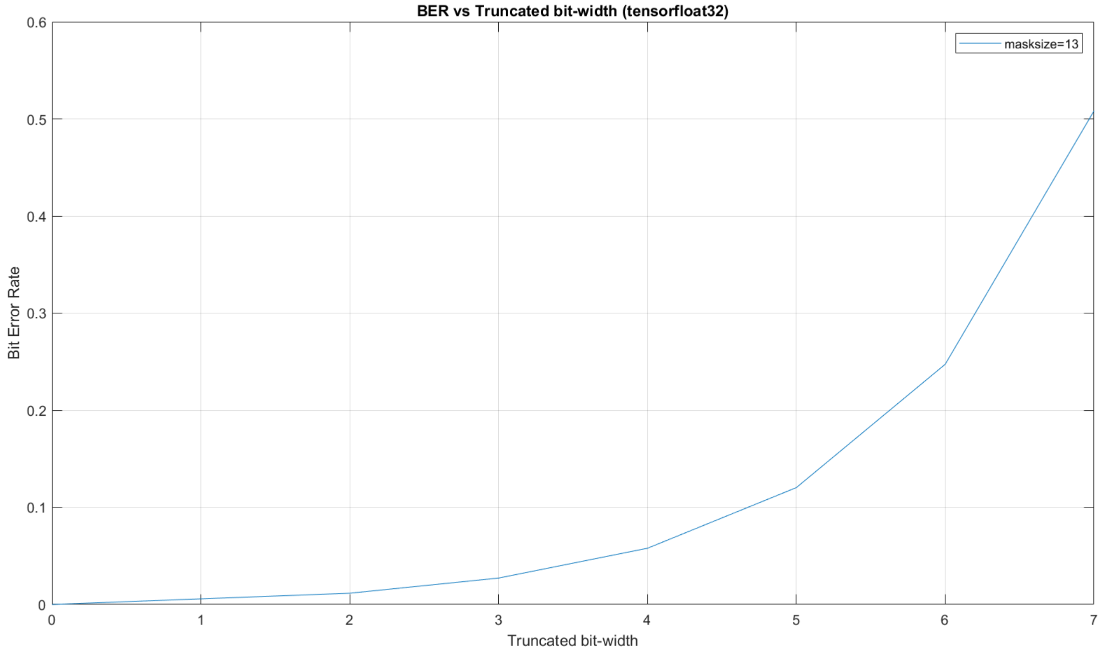


Figure 11: Tensorfloat32

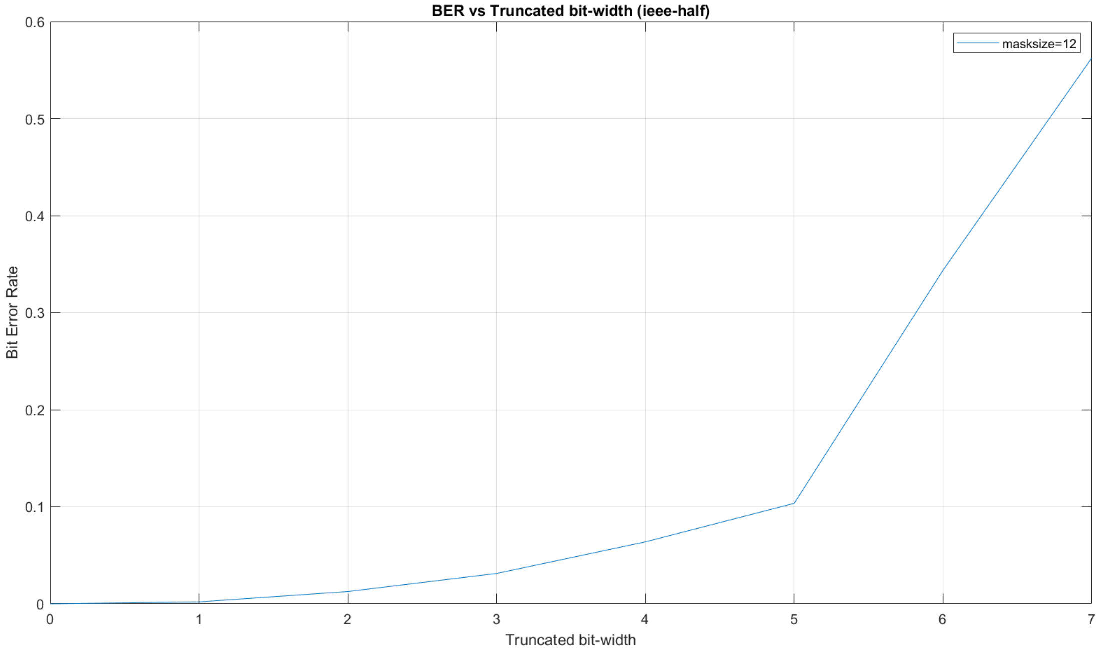


Figure 12: IEEE Half Precision

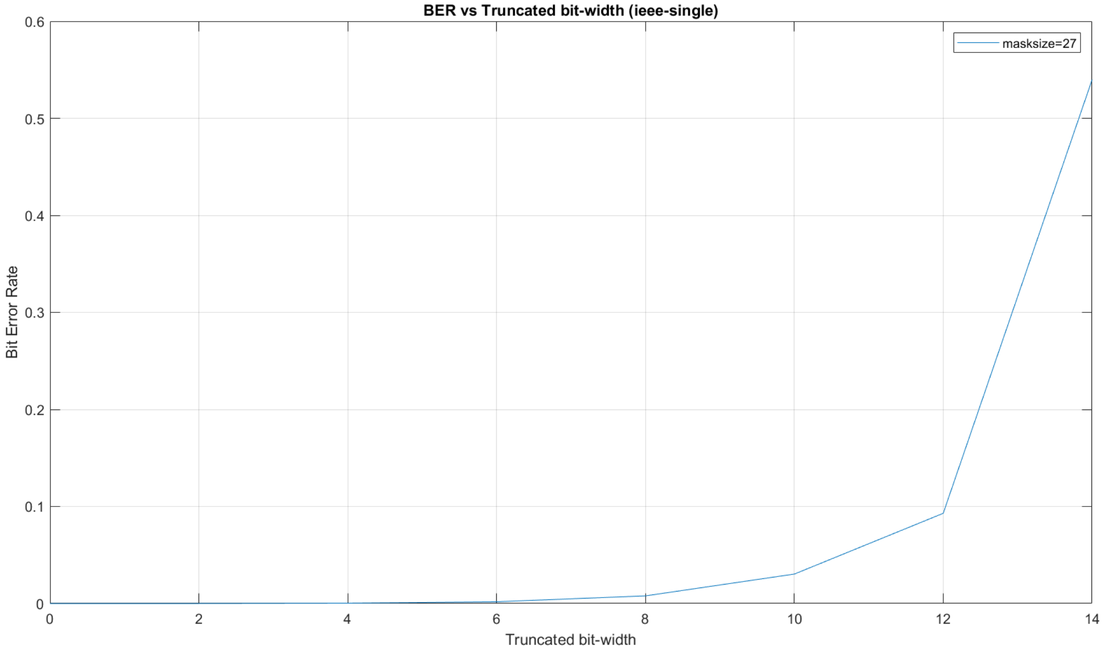


Figure 13: IEEE Single Precision

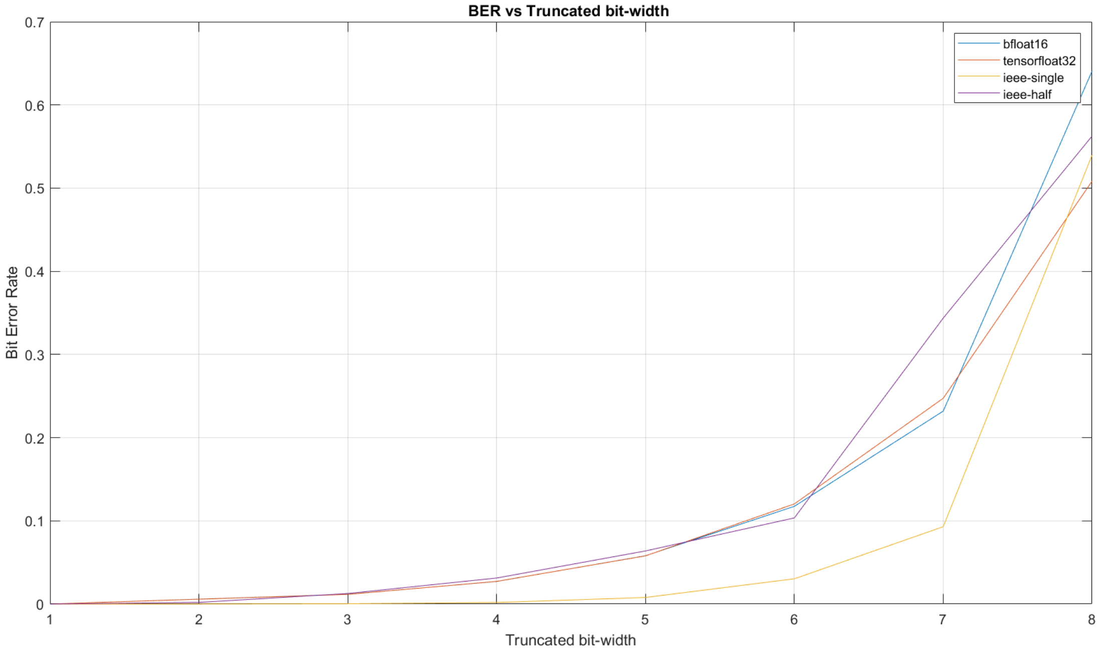


Figure 14: Bit Error Rate

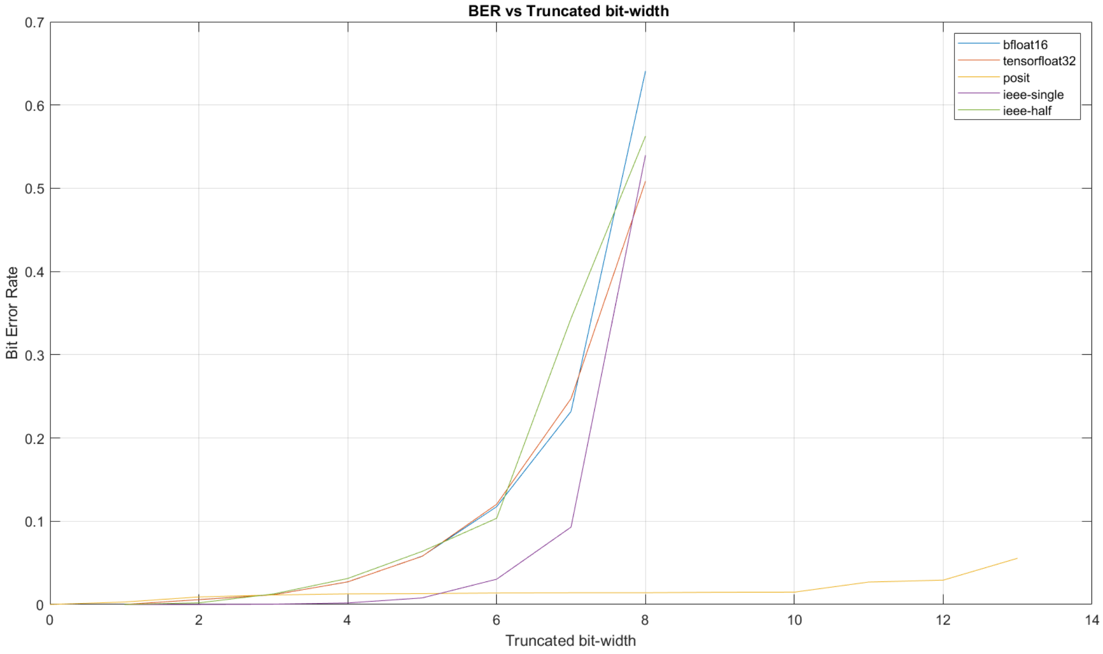


Figure 15: Bit Error Rate-Posit