Dual Fixed-Point

C. Ewe, P. Cheung, G. Constantinides
Imperial College, London, UK

ACME Seminar

Mark L. Chang

May 4, 2004

Floating-Point vs. Fixed-Point

- Fixed-point
  - Partition a binary word into integer and fractional
  - Radix point is in a fixed position
  - Two's complement:
    q (integer) + p (fractional) = n

$$X = -x_{q-1} \cdot 2^{q-1} + \sum_{i=-p}^{q-2} x_i 2^i$$

Floating-Point vs. Fixed-Point

- Floating-point
  - Large dynamic range
  - Composed of a mantissa and exponent

$$F = M \cdot \beta^{E}$$

- Fixed point good for most hardware designs
- Floating-point necessary for problems with large dynamic range

Dual FiXed-point (DFX)

- Single Exponent bit (E) selects between two scalings for significand
- Two possible ranges for the number
- Lower: *Num0*
- Higher: *Num1*

Scaling DFX Numbers

- Define two radix points
  - p_0 and p_1, $p_0 > p_1$
  - Represent number of bits from the LSB

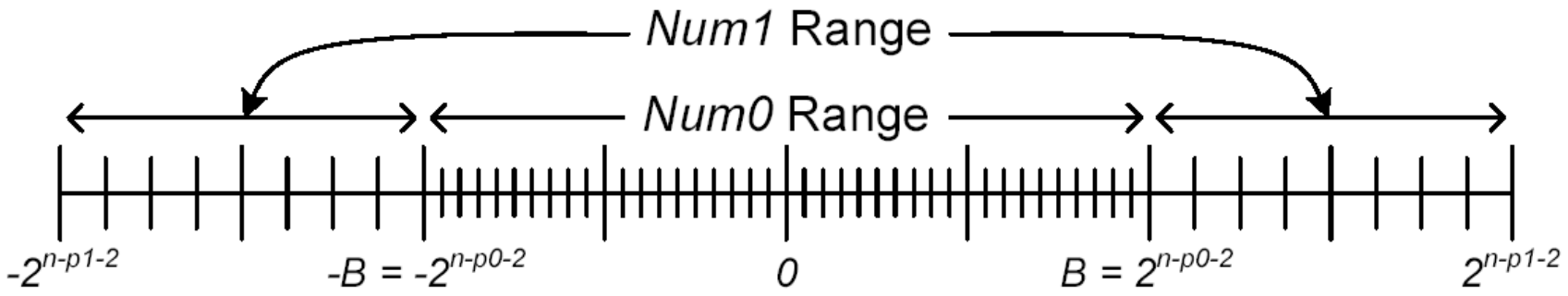
$$D = \begin{cases} X \cdot 2^{-p_0} & \text{if } E = 0 \\ X \cdot 2^{-p_1} & \text{if } E = 1 \end{cases}$$

- *Boundary value*, B used to decide scaling

$$E = \begin{cases} 0 & \text{if } -B \leq D < B \\ 1 & \text{if } D < -B \text{ or } D \geq B \end{cases}$$

Range of DFX Numbers

- DFX Number: n_p_0_p_1
  - Number of bits n
  - High radix position p_0 yields Num0
  - Low radix position p_1 yields Num1



Dynamic Range

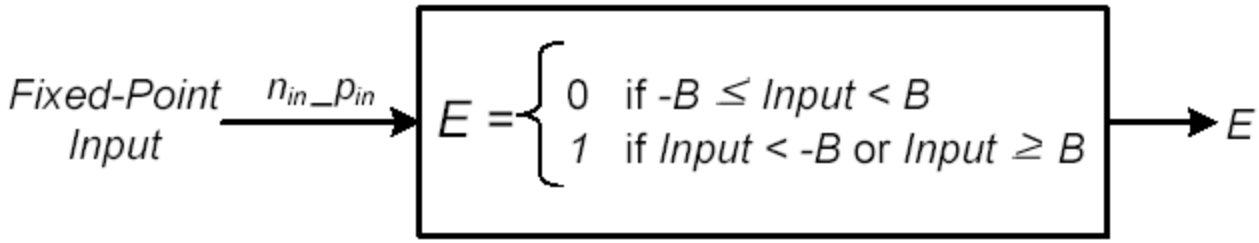
- Dynamic range = |largest| / |smallest|
- Smallest DFX = 2^{-p0}
- Largest DFX = 2^{n-p1-2}

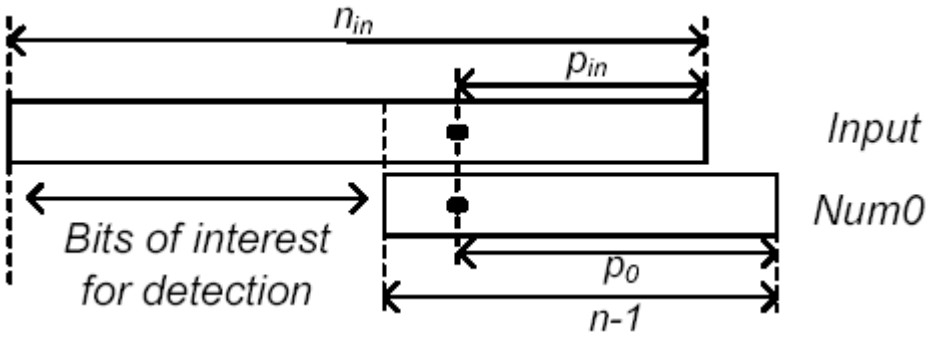
$$\text{Dynamic range} = 20\log_{10}(2^{n+p_0-p_1-2})\ \text{dB}$$

Number System	*Dual FiXed-point*	*Dual FiXed-point*	*Fixed Point*	*Floating Point*
Format	32_30_0	32_16_4	32-bit	32-bit IEEE
Dynamic Ranges	$2^{60} \approx$ 361dB	$2^{46} \approx$ 276dB	$2^{31} \approx$ 187dB	$2^{254} \approx$ 1529dB

Range Detector

- Generates exponent E based on fixed-point input

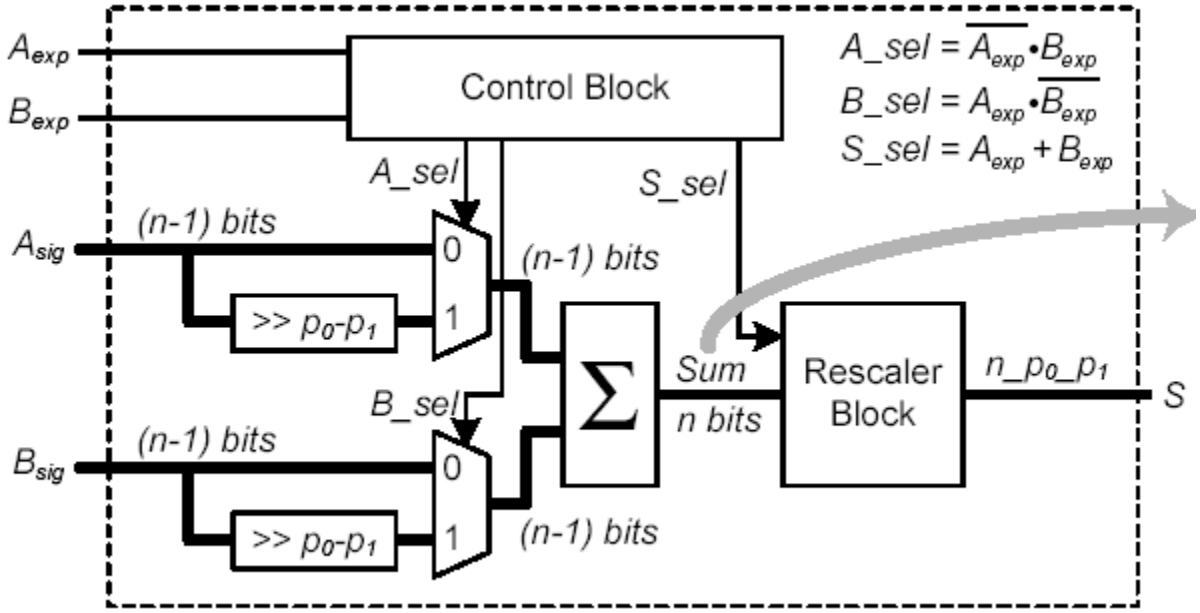




Range Detector Function

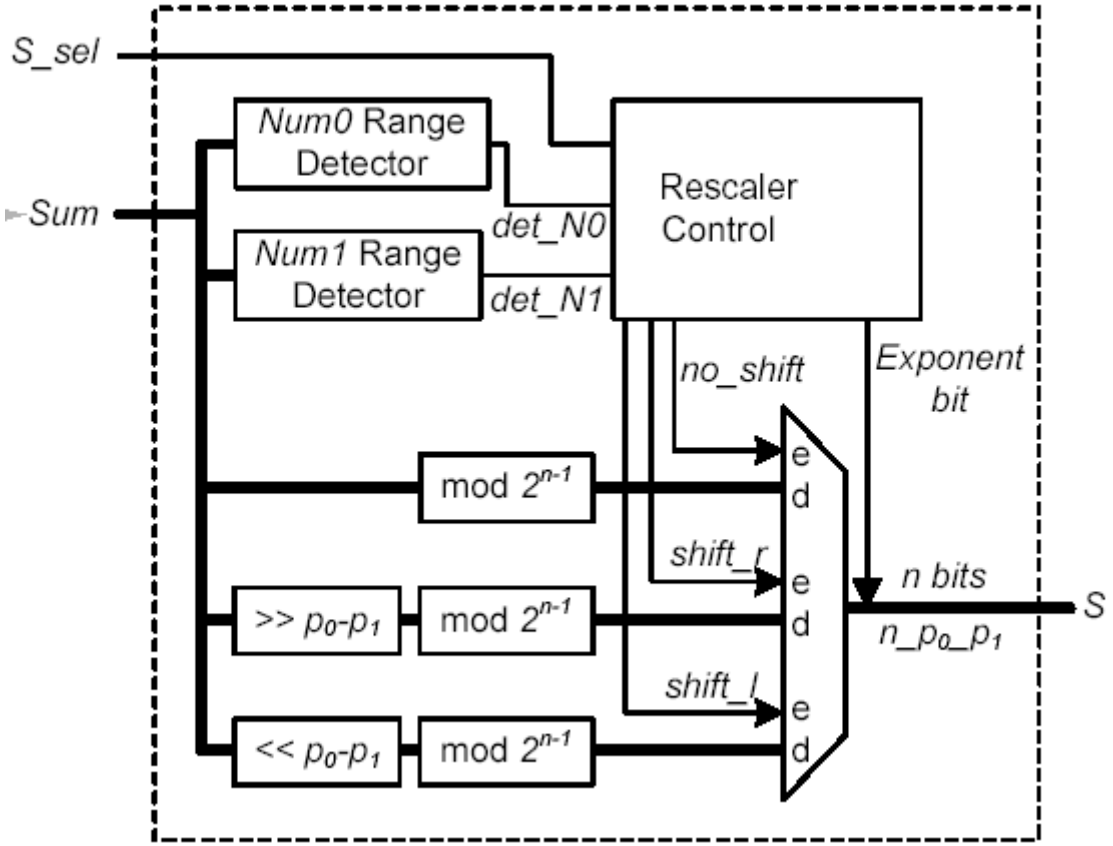
$$E = \overline{d_{n_{in}-1} \cdot d_{n_{in}-2} \cdot \ldots \cdot d_{n_{in}-(n-p_0-2)-p_{in}}} + \overline{\overline{d_{n_{in}-1}} \cdot \overline{d_{n_{in}-2}} \cdot \ldots \cdot \overline{d_{n_{in}-(n-p_0-2)-p_{in}}}}$$

DFX Adder



- Fixed-length shift for radix alignment

DFX Adder Rescaler



DFX Adder Results

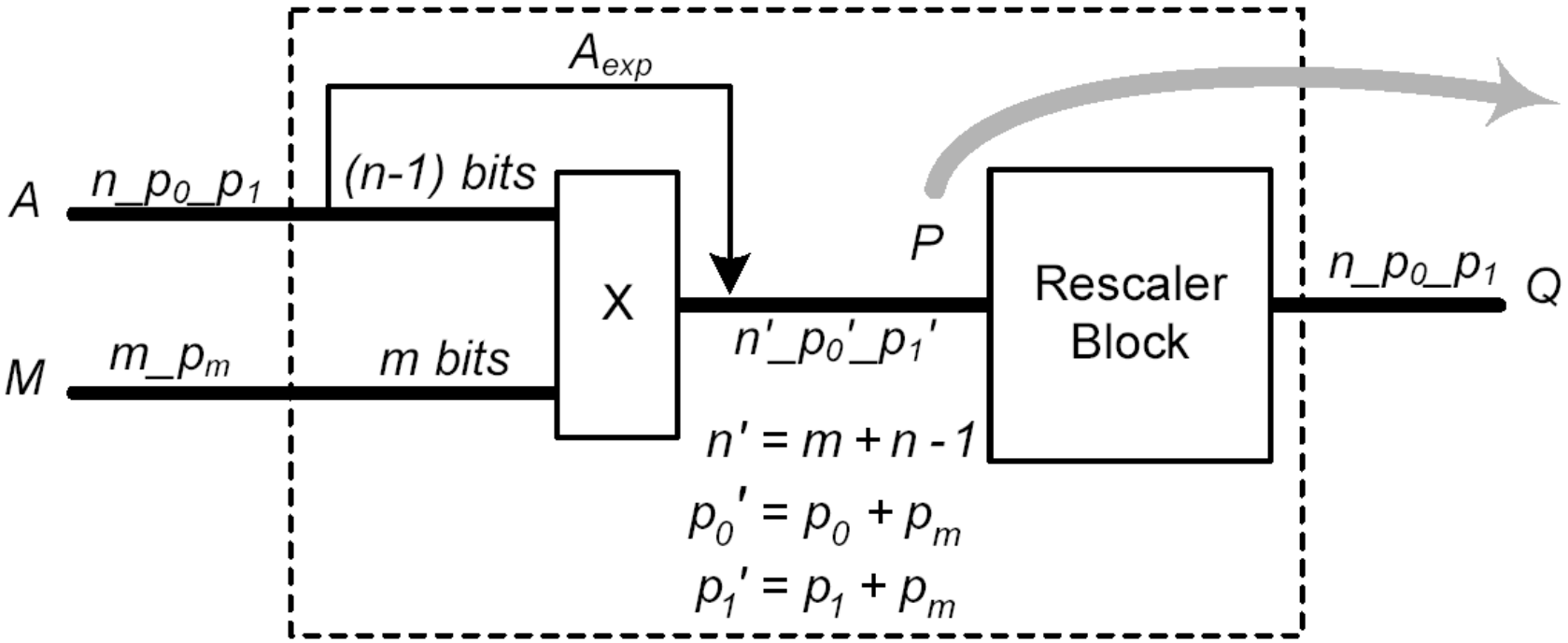
Adder Type	Size (Slices)	Latency (ns)
Fixed-Point	17	2.5
DFX	64	10.28
Floating-Point (IEEE)	255	34.48

- Xilinx Virtex XC2V80
- 32-bit adders
- 4x larger and slower than fixed-point
- 4x smaller and faster than floating-point

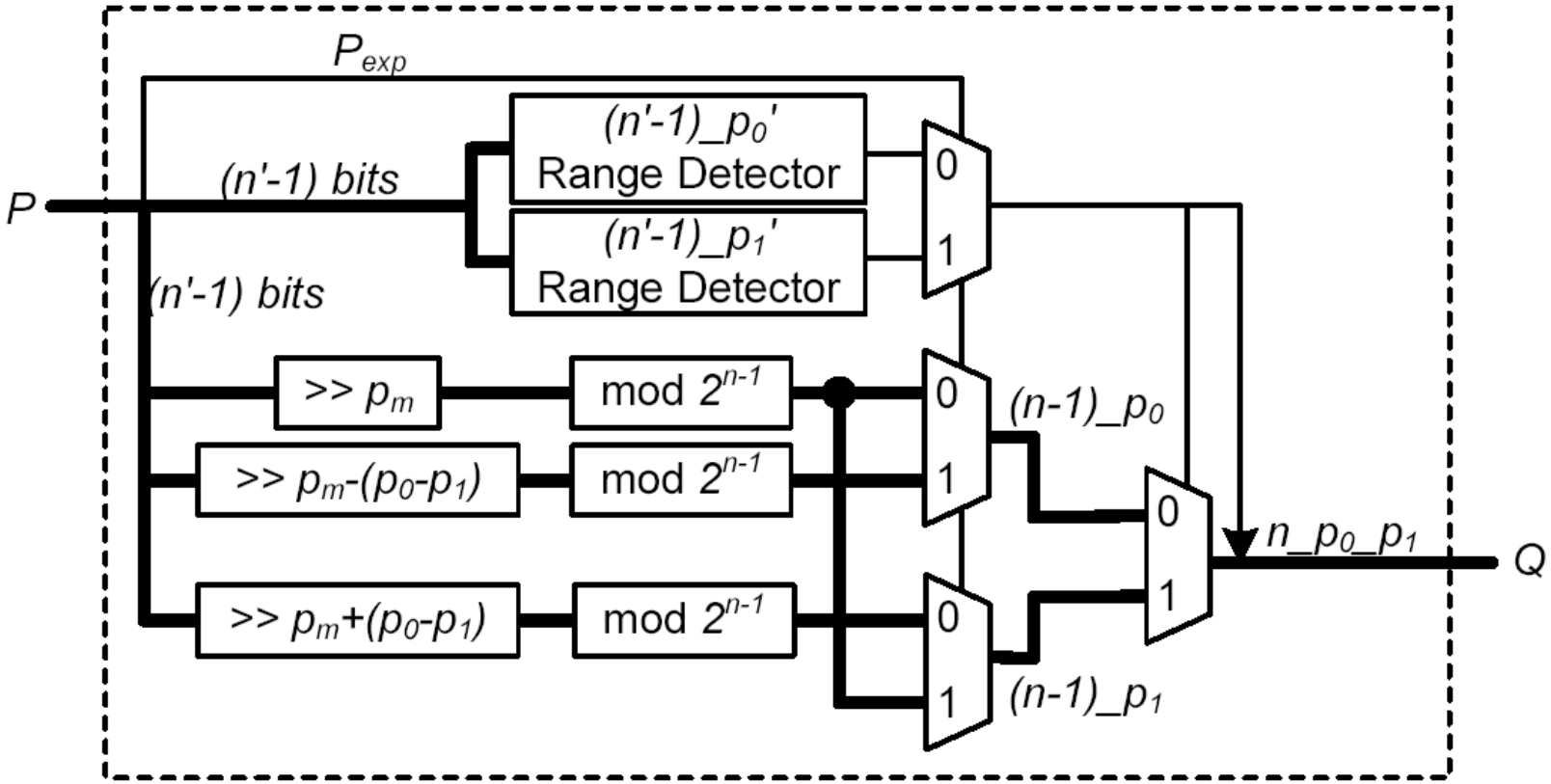
Module	Size (Slices)	Latency (ns)
Encoder	17.5	7.8
Decoder	10	5.8

DFX-H Multiplier

- DFX-H and DFX-F designed



DFX Multiplier Rescaler



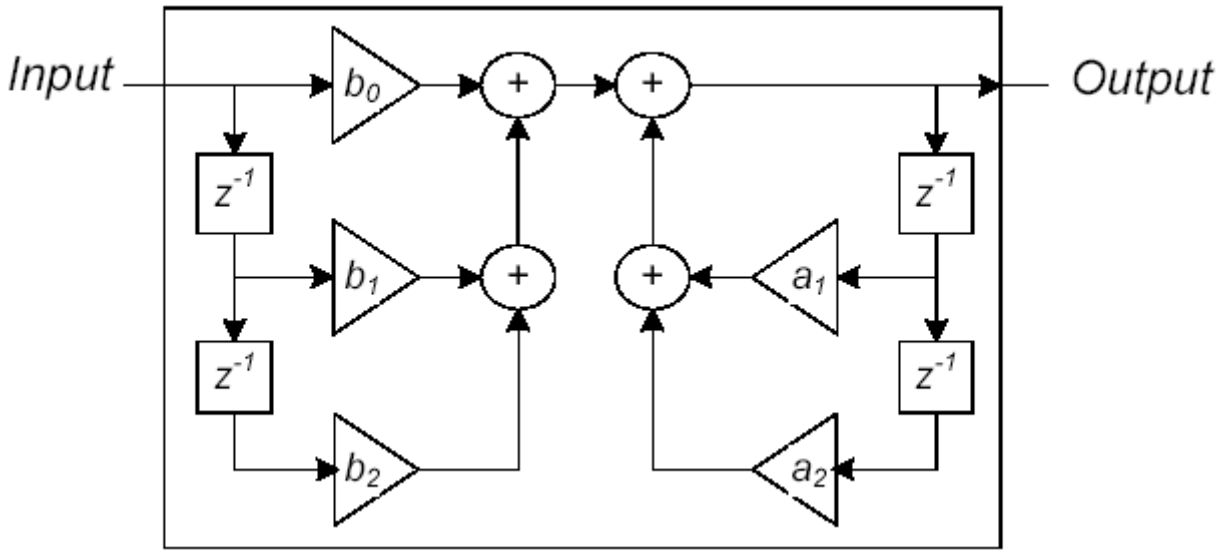
DFX Multiplier Results

Multiplier Type	Size (Slices)	Latency (ns)
Fixed-point	43	13.946
DFX-H Mult	58	17.308
DFX-F Mult	76	19.149
Floating-point	73	20.683

- DFX-H 1.2x larger and slower than fixed-point
- DFX-H 1.2x smaller and faster than FP
- DFX-F 1.5 larger and slower than fixed-point
- DFX-F comparable to floating-point

Benchmark: IIR Filter

- 2nd order notch IIR filter

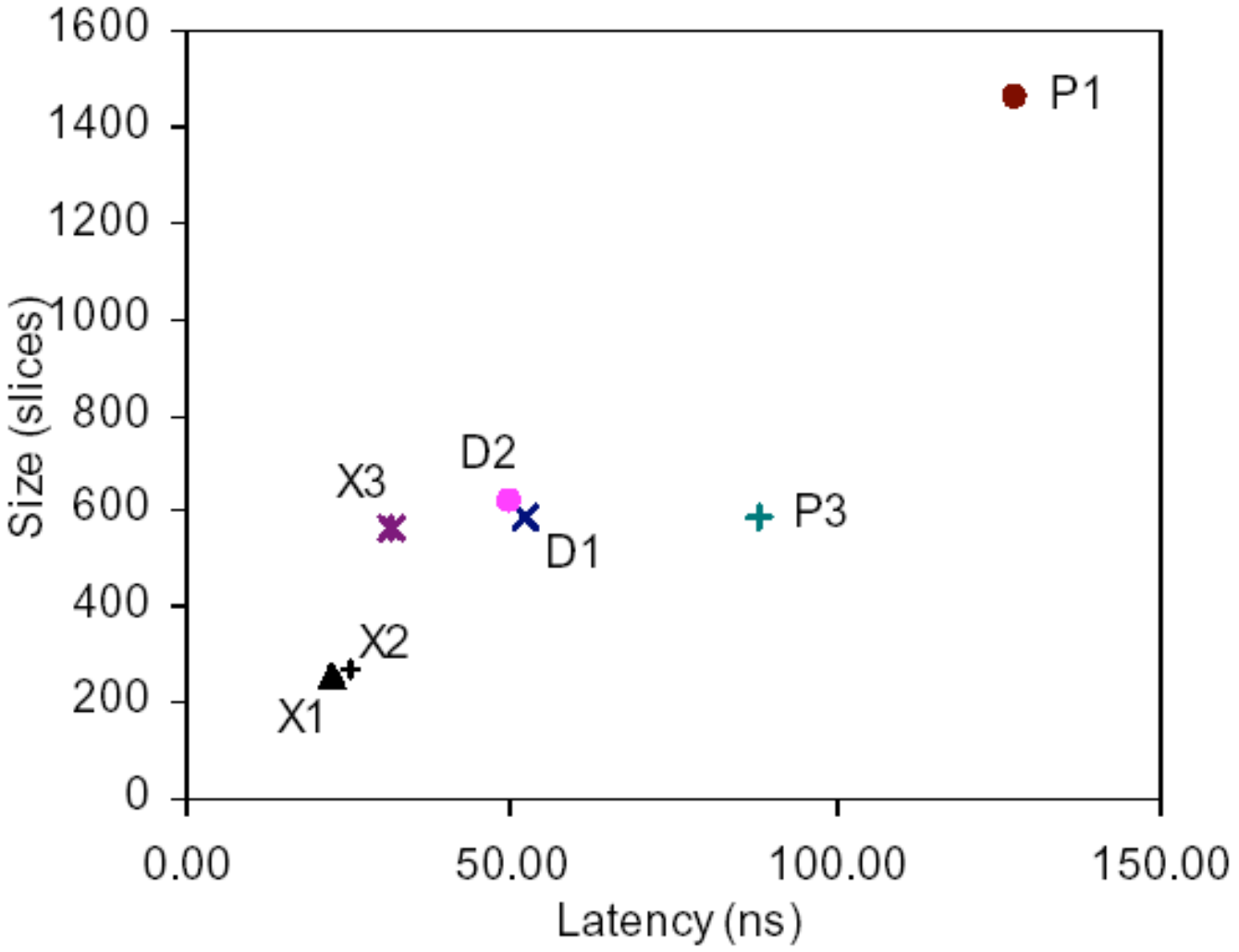


Benchmark Results

- Xilinx Virtex II XC2V500

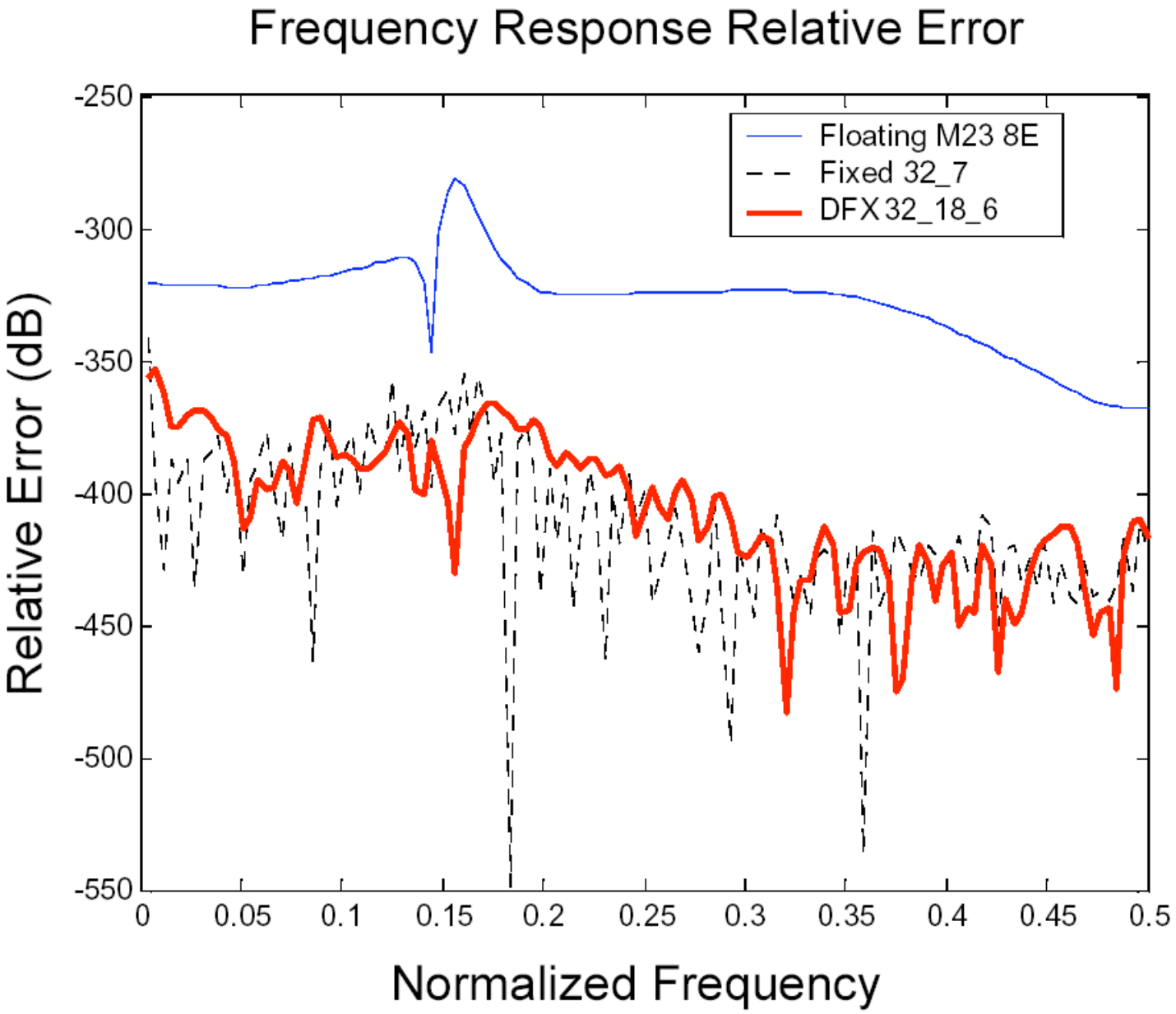
Filter Type	Design	Format	Size (Slices)	Latency (ns)
DFX	D1	32_18_6	584	52.29
	D2	32_9_6	580	51.28
Fixed Point	X1	32_7	255	24.26
	X2	33_8	272	24.18
	X3	41_16	565	31.25
Floating Point	P1	32bit M23 E8	1459	127.39
	P3	17bit M10 E6	586	88.183

Benchmark Results



Error Analysis

Filter Type	Design	Format	Error Variance	Av % Relative Error	Max & Relative Error
DFX	D1	32_18_6	0.0005	0.0072%	0.89%
	D2	32_9_6	0.0001	3.2980%	450.43%
Fixed Point	X1	32_7	0.0002	12.2981%	1611.42%
	X2	33_8	0.0000	6.3465%	773.97%
	X3	41_16	0.0000	0.0268%	3.47%
Floating Point	P1	32bit M23 E8	0.0219	0.0051%	3.21%
	P2	17bit M10 E6	2.09E+06	43.4488%	25853.78%


Frequency Response Relative Error
Floating M23 8E
Fixed 32_7
DFX 32_18_6
Relative Error (dB)
-250
-300
-350
-400
-450
-500
-550
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Normalized Frequency