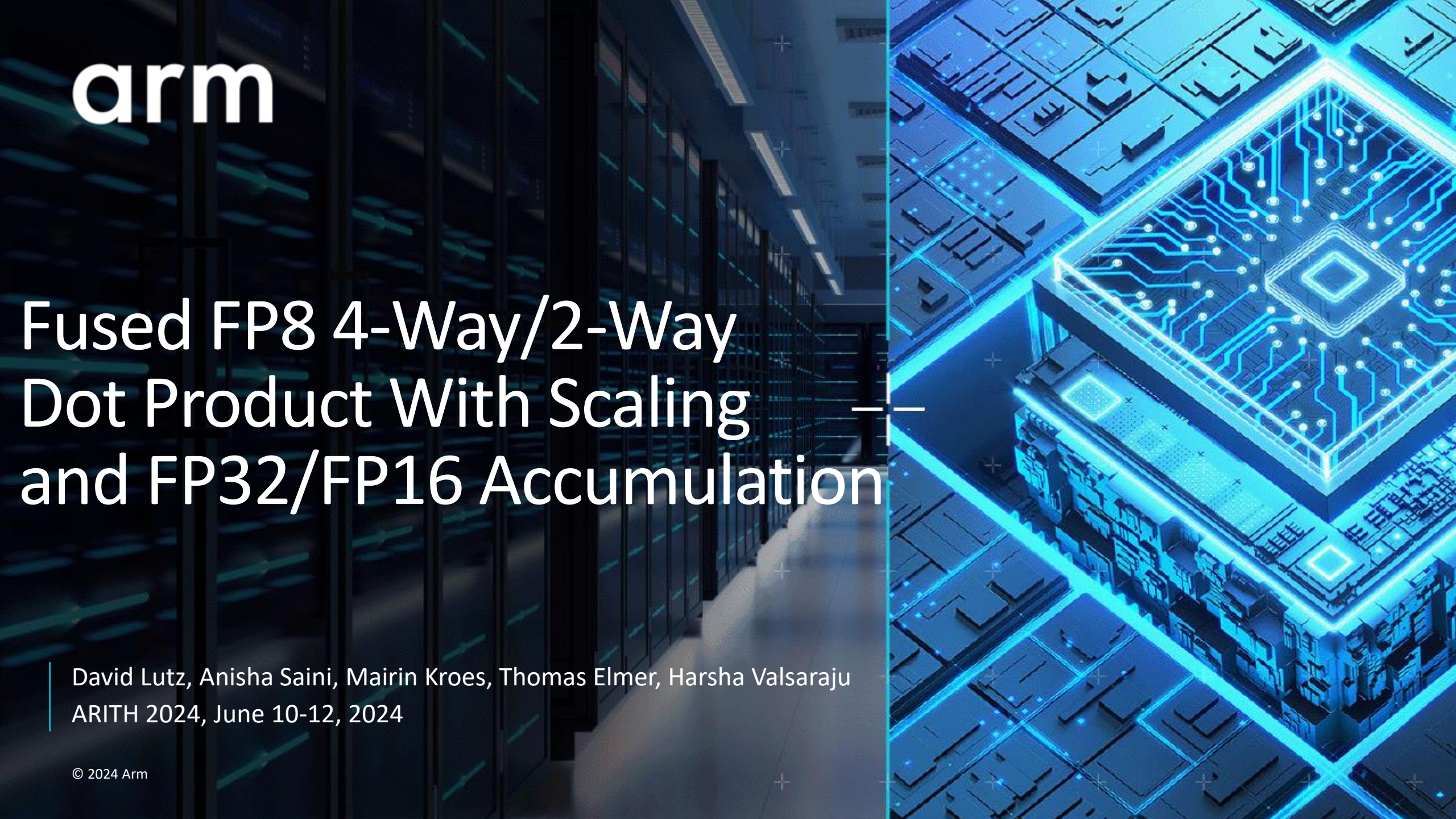


Fused FP8 4-Way/2-Way Dot Product With Scaling and FP32/FP16 Accumulation

David Lutz, Anisha Saini, Mairin Kroes, Thomas Elmer, Harsha Valsaraju

ARITH 2024, June 10-12, 2024

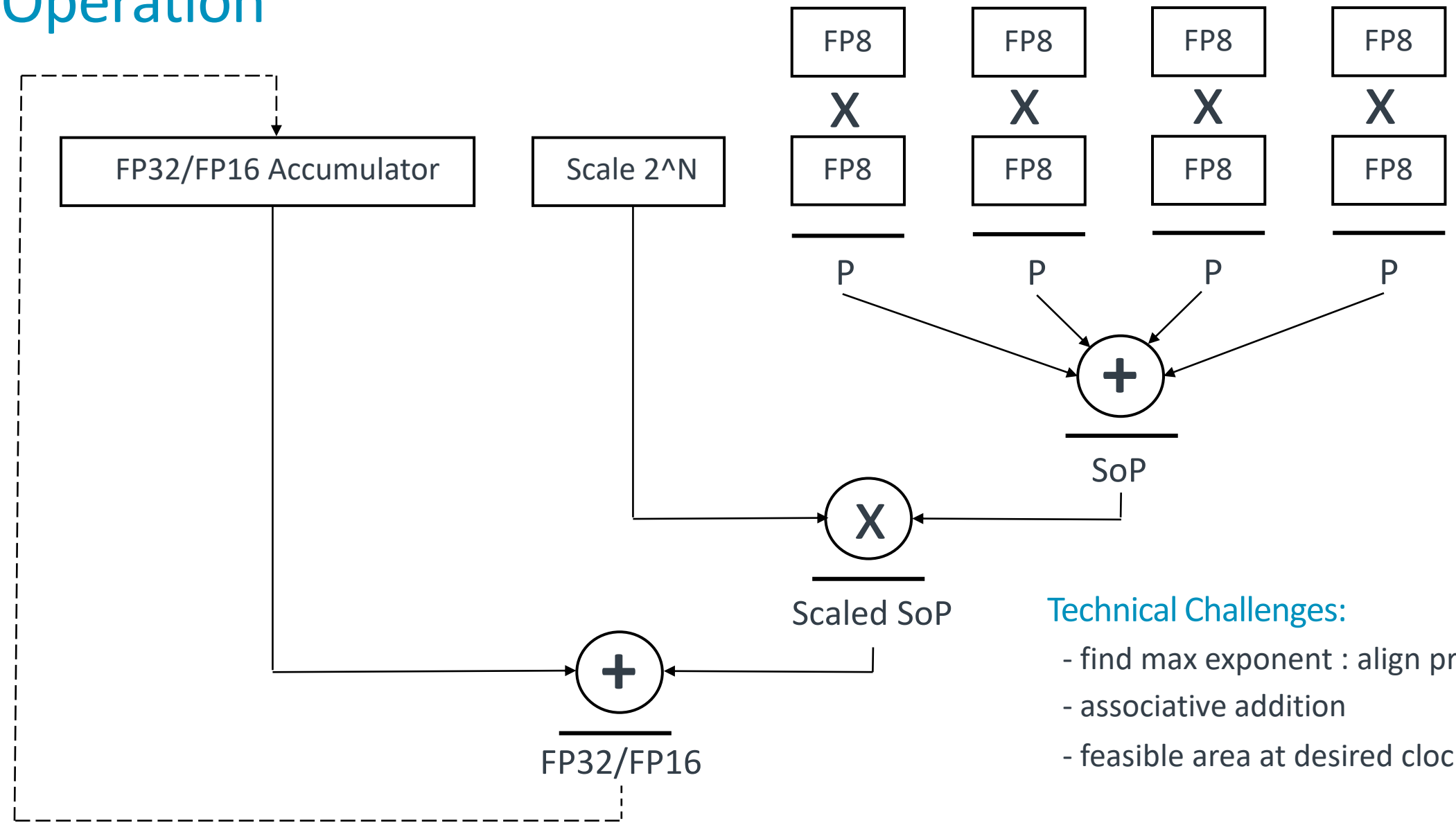
Agenda

- Motivation for Evaluating FP8-DOT4/2 Designs
- Arithmetic Operation Provided
- Recognized Technical Challenges
- Prior Art
- Overview: 2 Micro-architecture Solutions
- Micro-architecture Details
- Synthesis Results & Evaluation

Motivation

- Mixed-precision GEMM for ML now significant compute workload for CPU & dedicated accelerators
- Small input data types (FP8, INT8) now popular for ML
  - Reduce memory footprint as model sizes increase
  - Increase memory I/O efficiency (data element bandwidth, power-per-datum transfer)
  - Increase computational density (TOPs/mm^2)
  - Reduce computational power (Joules/TOP)
- FP8 : natively provides flexible range, precision : +/-e5m2, +/-e4m3, etc (IEEE P3109)
  - Simpler quantization/de-quantization of data in ML processing
- DOT Product Increases Computational Throughput
  - DOT4 consumes 4 operand pairs in parallel from 32-bit elements -vs- serial MAC
- Current literature demonstrates some apps need FP32/FP16 accumulation range/accuracy for acceptable performance
  - Mini-block (DOT4) summation preserves tiny values
  - Allow dynamic SOP scaling into FP32/FP16 accumulator ranges
  - Fully fused SOP + FP32 accumulator & Single rounding (RNE) step
- Evaluate design techniques, area cost & timing for datapath solutions

Operation



Technical Challenges:

- find max exponent : align prod, acc
- associative addition
- feasible area at desired clock speed

arm

Prior Art

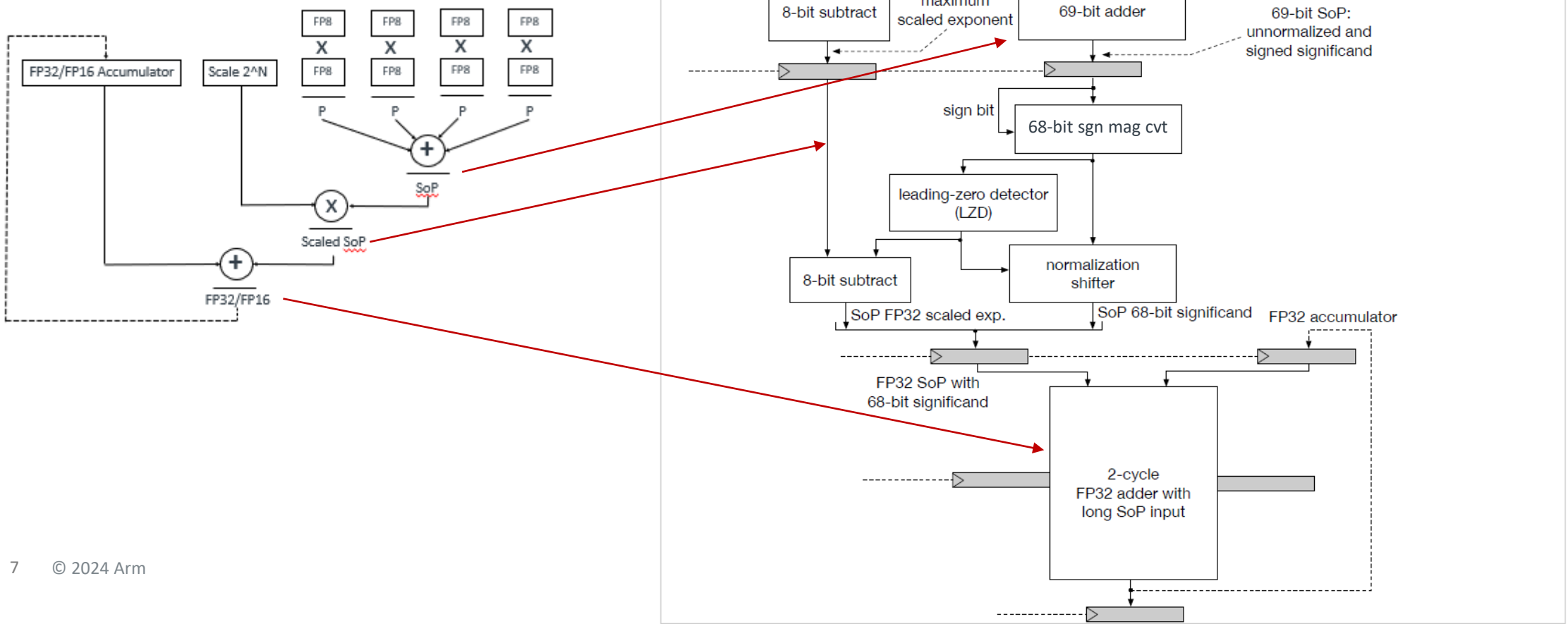
- Several papers report reduced precision DOT product work
- [9] Sohn, Swartzlander, 2016, " A Fused Floating-Point Four-Term Dot Product…"
  - 4-way DOT => FP32/FP64, no SoP scaling
  - max exp circuit for alignment
  - discard sticky bits beyond max shift
- [7] Kaul, et al., 2019, "Optimized Fused Floating-Point Many-Term Dot-Product…"
  - optimized timing complexity : max exp calc => local + global alignment stages
  - 8-to-32-way BF16 DOT with FP32 accumulation
  - SoP calc truncated to optimize PPA
- [6] Hickmann, et al., 2020, "Intel Nervana Neural Network Processor-T (NNP-T)…"
  - 32-way BF16 DOT – 2's comp SoP => FP32 accumulator, no SoP scaling
  - timing complexity - max exp calc for alignment
  - SoP truncated (37 bits) to optimize PPA – 9 stage pipeline
- [13] Desrentes, et al., 2023, "Exact Dot Product Accumulate Operators for 8-bit Floating-Point…"
  - 16-way DOT with 2's comp FX accumulation, no SoP scaling
  - large FX Acc for long FP8 product summations
  - separate convert ops to FP32, FP8

Overview

- Two Different Micro-architecture Solutions
  - FP8-DOT4/2-LA : Late Accumulation
  - FP8-DOT4/2-EA : Early Accumulation
  - {S,E,F} = +/-e5m2, +/-e4m3
  - 4 cycle latency from multiply operands : 5nm 3.6GHz
  - Scaled SoP
- FP8-DOT4/2-LA
  - Acc consumed 2 cycles after dispatch
  - 2 cycle Acc latency : Higher dispatch rate for ops with strict Acc dependency : 2-2-2-2…
  - Allows fully pipelined, interleaved accumulation (e.g. 2 accumulators)
  - Can share (modified) FP32 adder with other operations
  - Requires small incremental area in (eg) CPU vector unit
- FP8-DOT4/2-EA
  - Acc consumed at dispatch cycle
  - 4 cycle Acc latency : Dispatch rate for ops with strict Acc dependency: 4-4-4-4…
  - Allows fully pipelined, interleaved accumulation (e.g. 4 accumulators)
  - Dedicated/”Stand-alone” datapath – no logic sharing assumed
  - Requires less total area in (eg) dedicated accelerator application

FP8-DOT4-LA

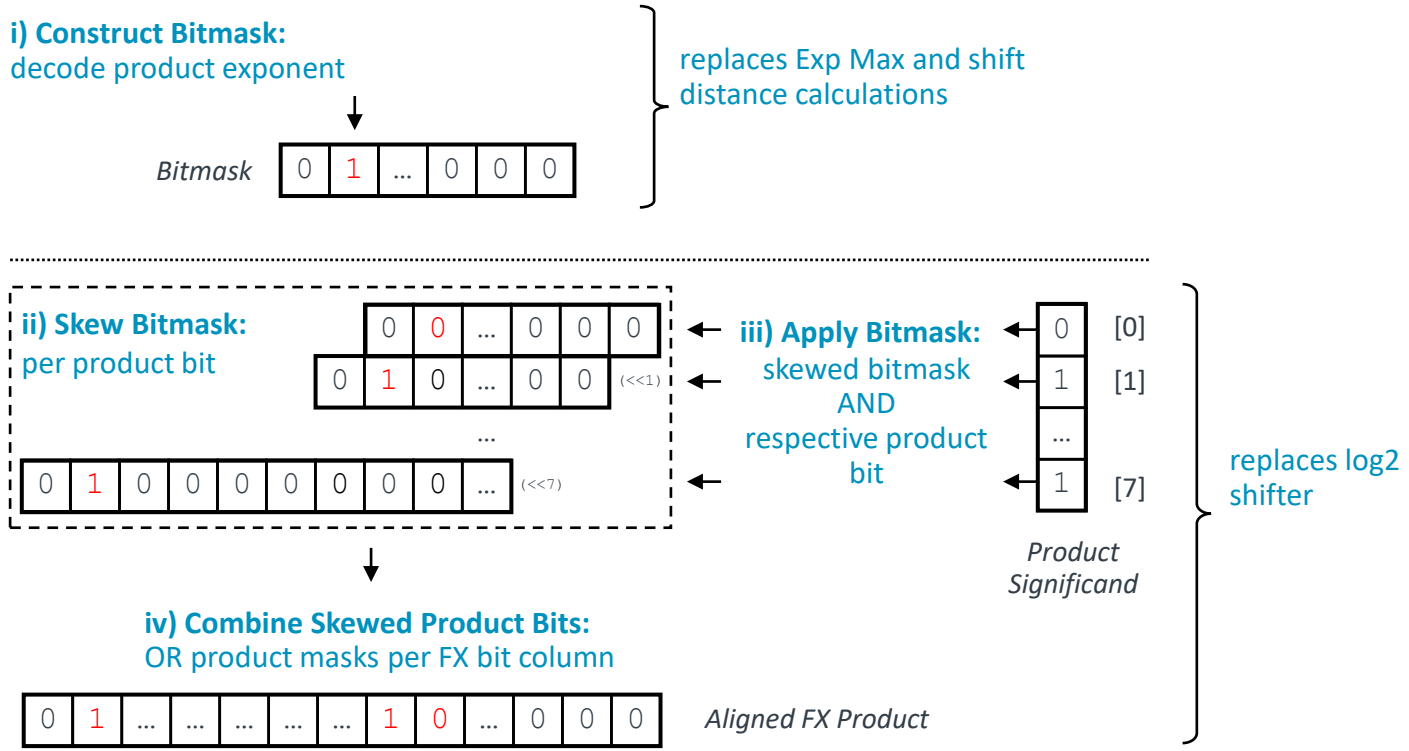
+/-e5m2, +/-e4m3

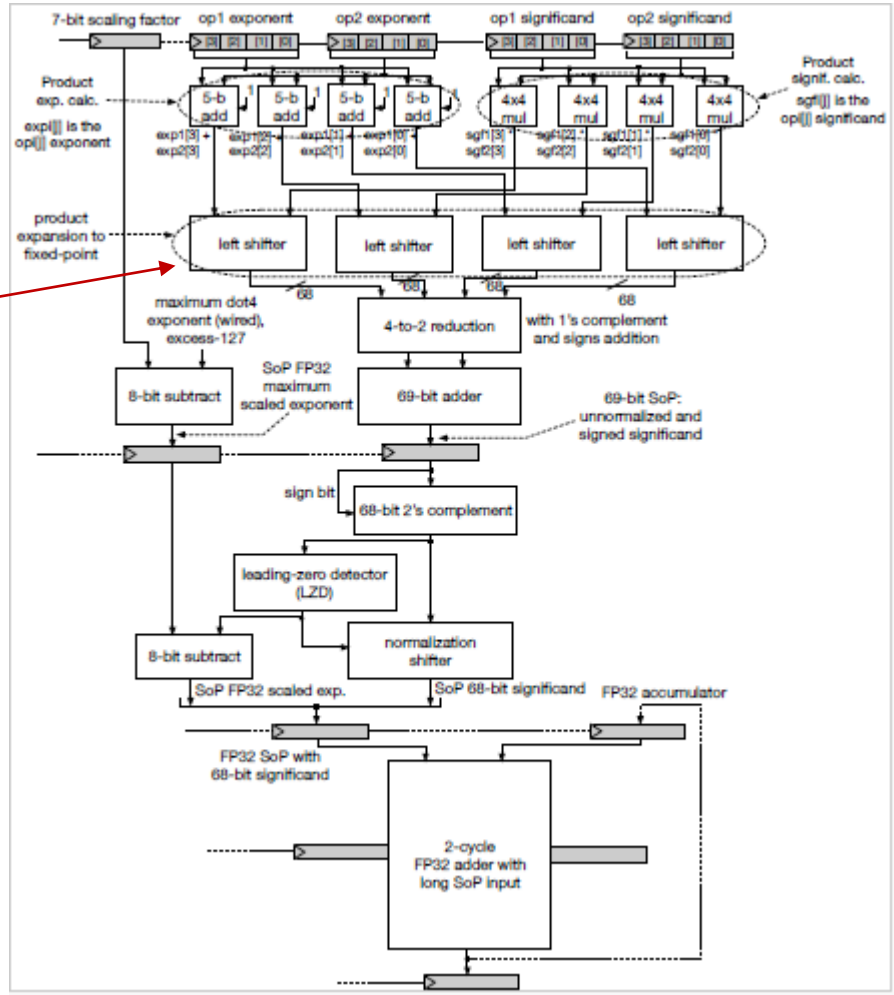


FP8-DOT4-LA

Mask Alignment - No Circuit Delay/Complexity for Exp Max Calc

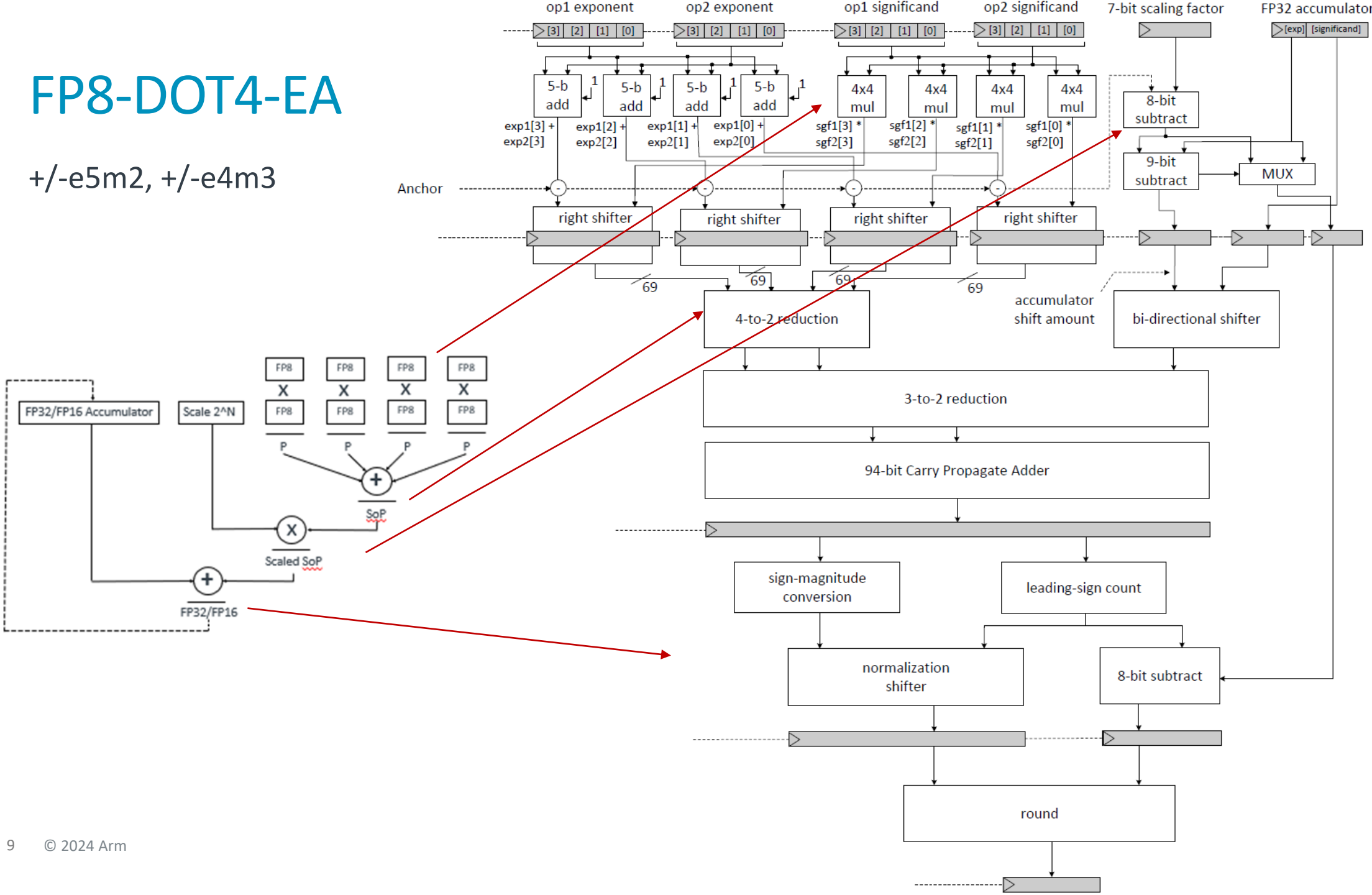
Convert FP8 Product to FX with Bitmask





FP8-DOT4-EA

+/-e5m2, +/-e4m3

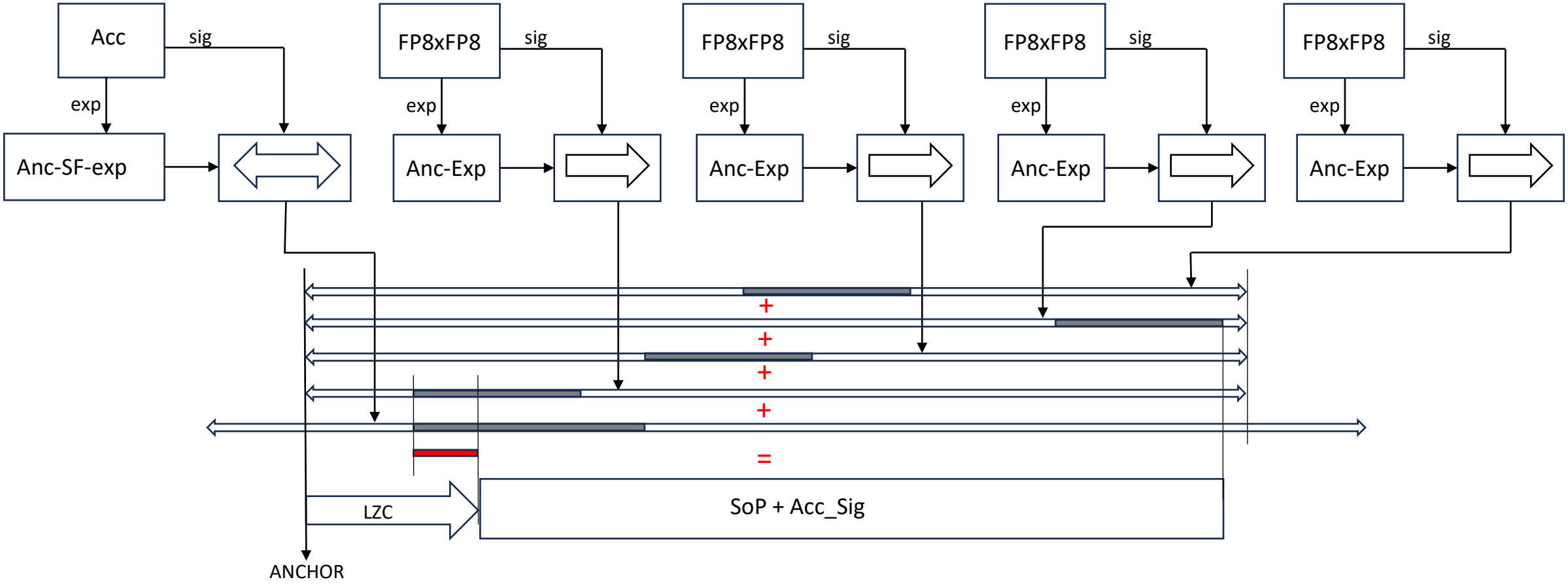


FP8-DOT4-EA : Anchor Shift Alignment

No Circuit Delay/Complexity for Exp Max Calc

Anchor = Max_Prod_Exp + ceil(Log2(N)) + 1 : (for N products)

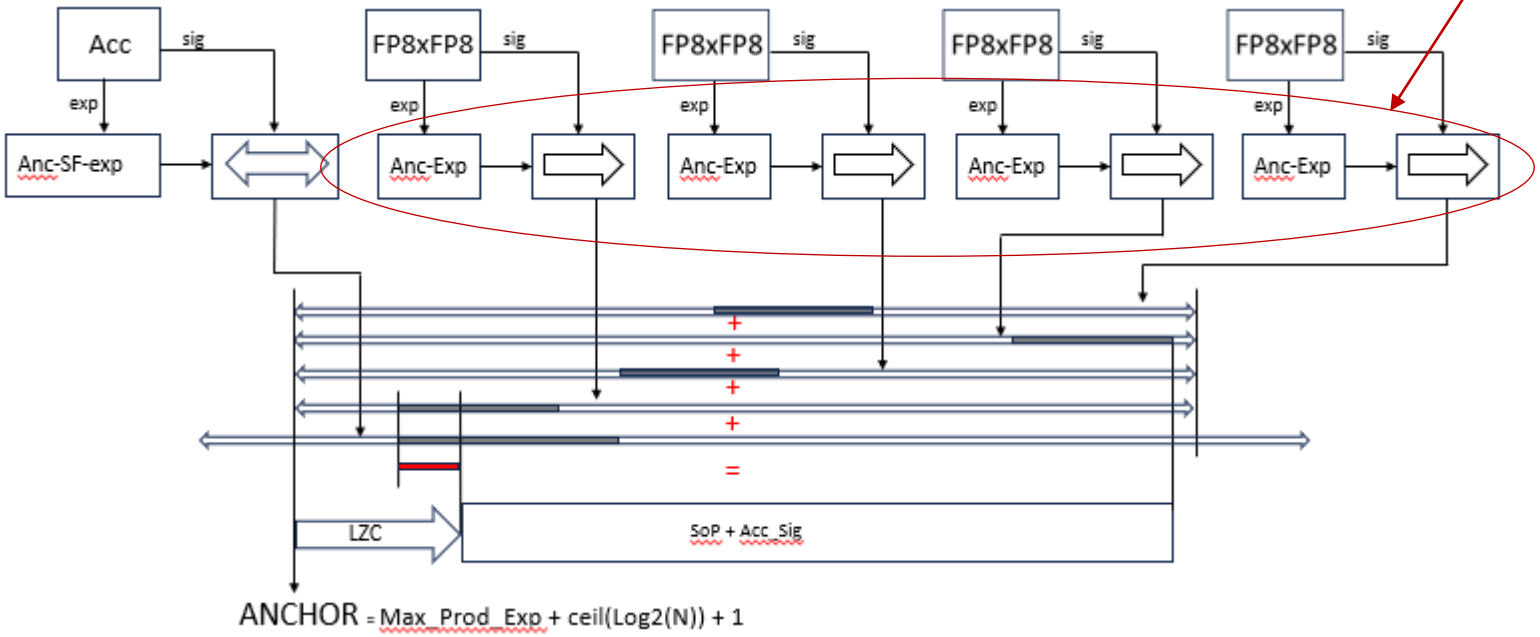
Scale Factor “Moves” Anchor To Lesser Magnitude

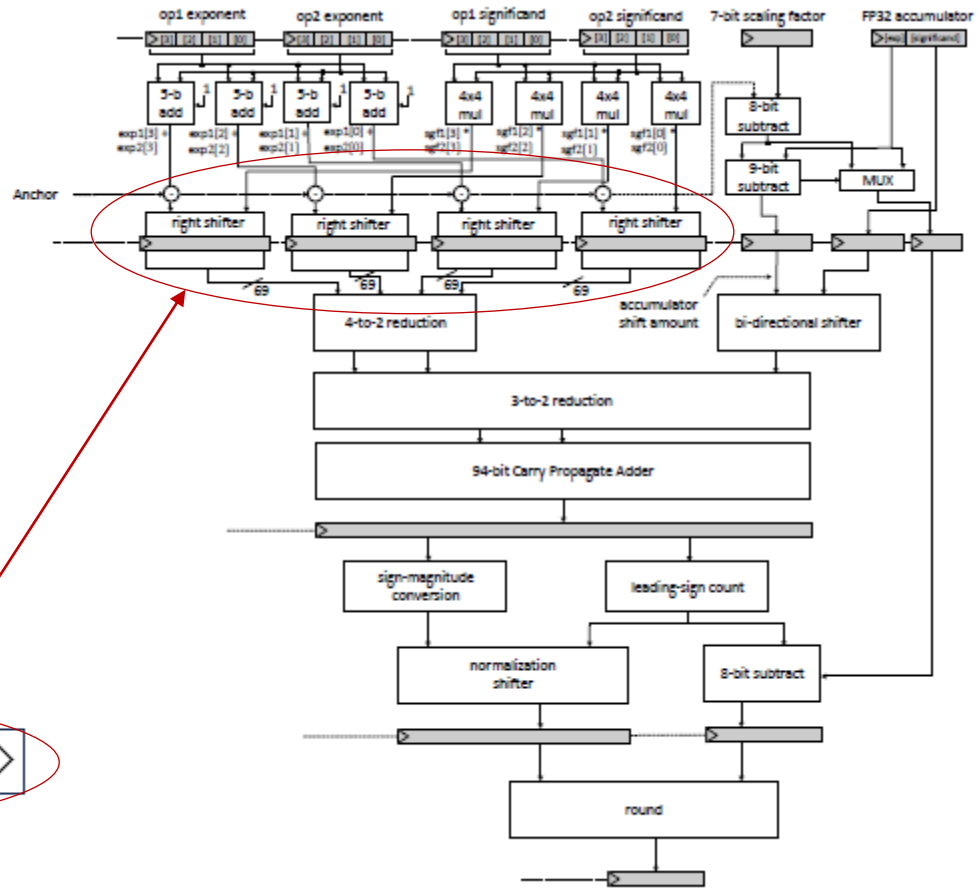


FP8-DOT4-EA

Product Shift :

- From Anchor
- Small First / Big Last : Reduce Reg Bits Req'd

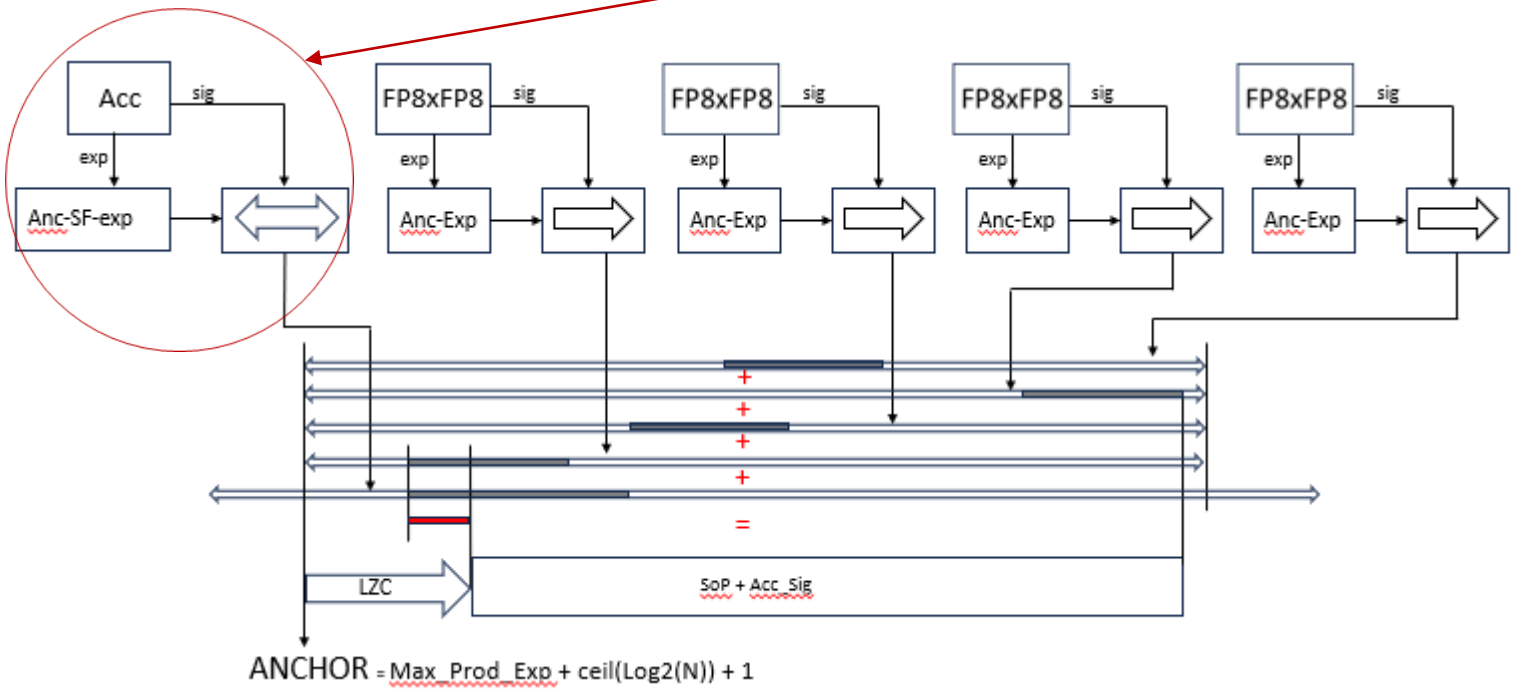


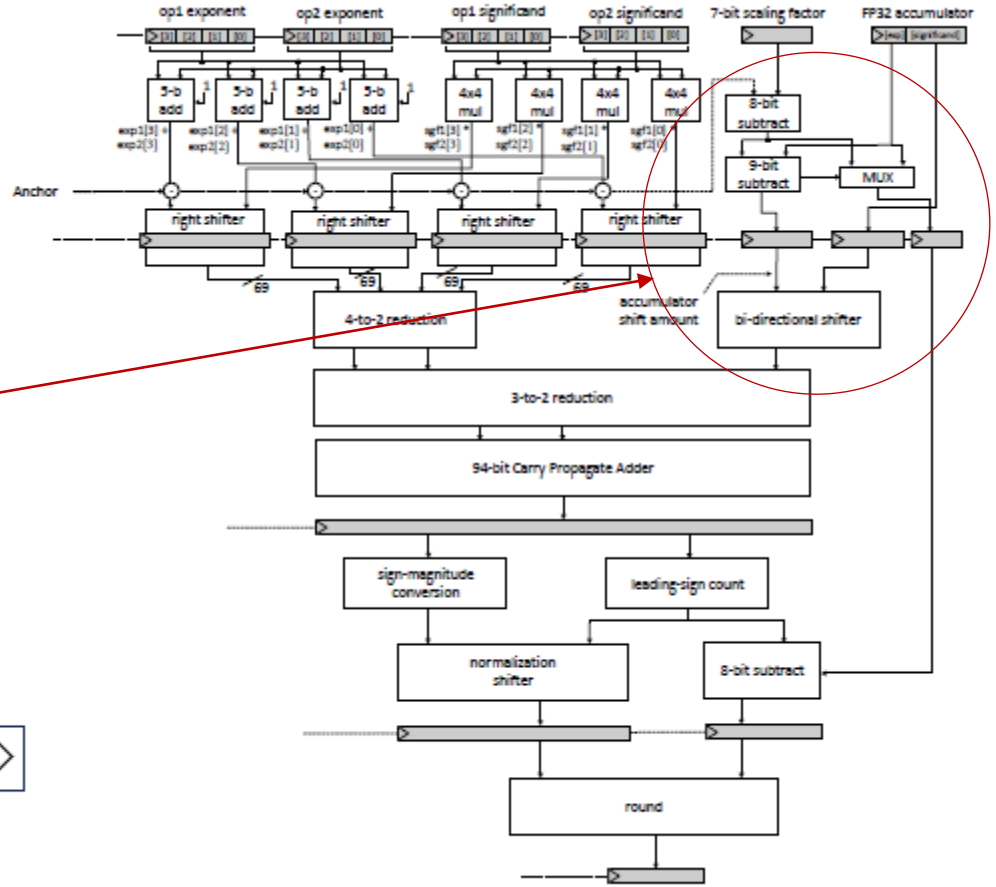


FP8-DOT4-EA

SoP Exp Scaled : Anchor–SF

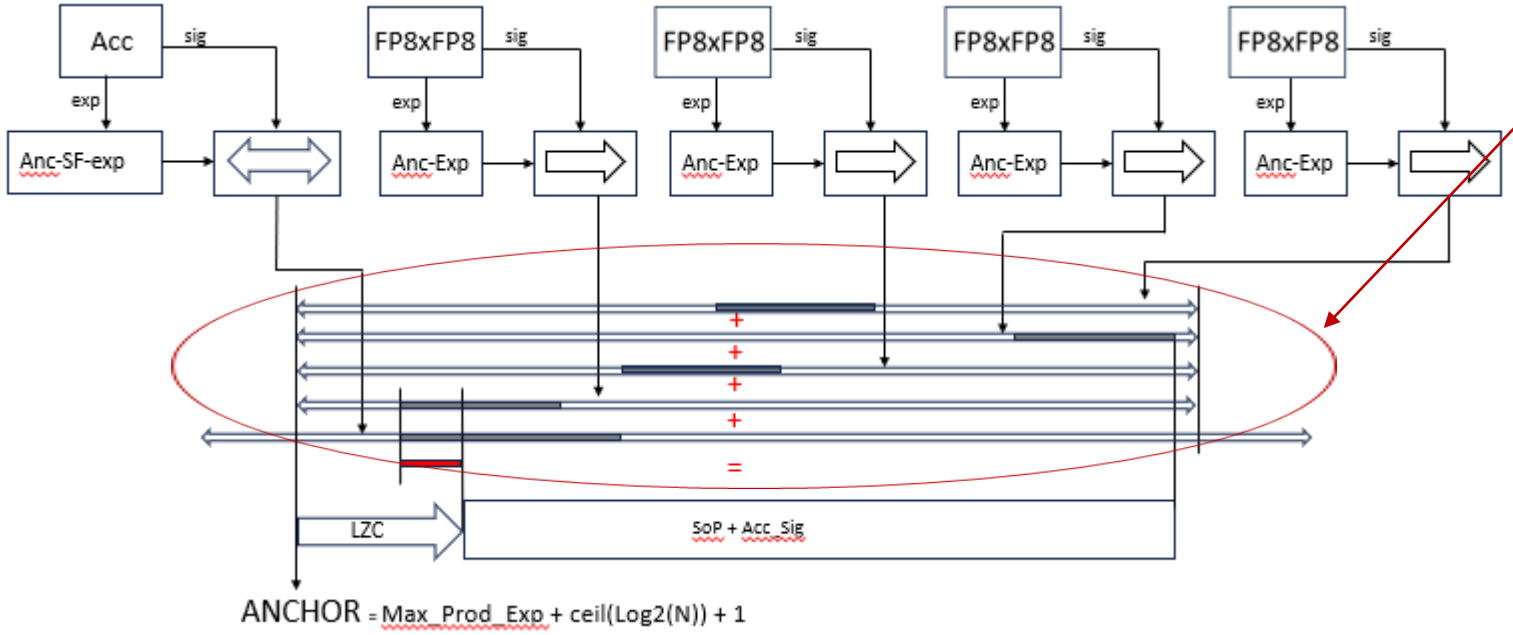
Major Exp Calc : Anchor-SF > AccExp

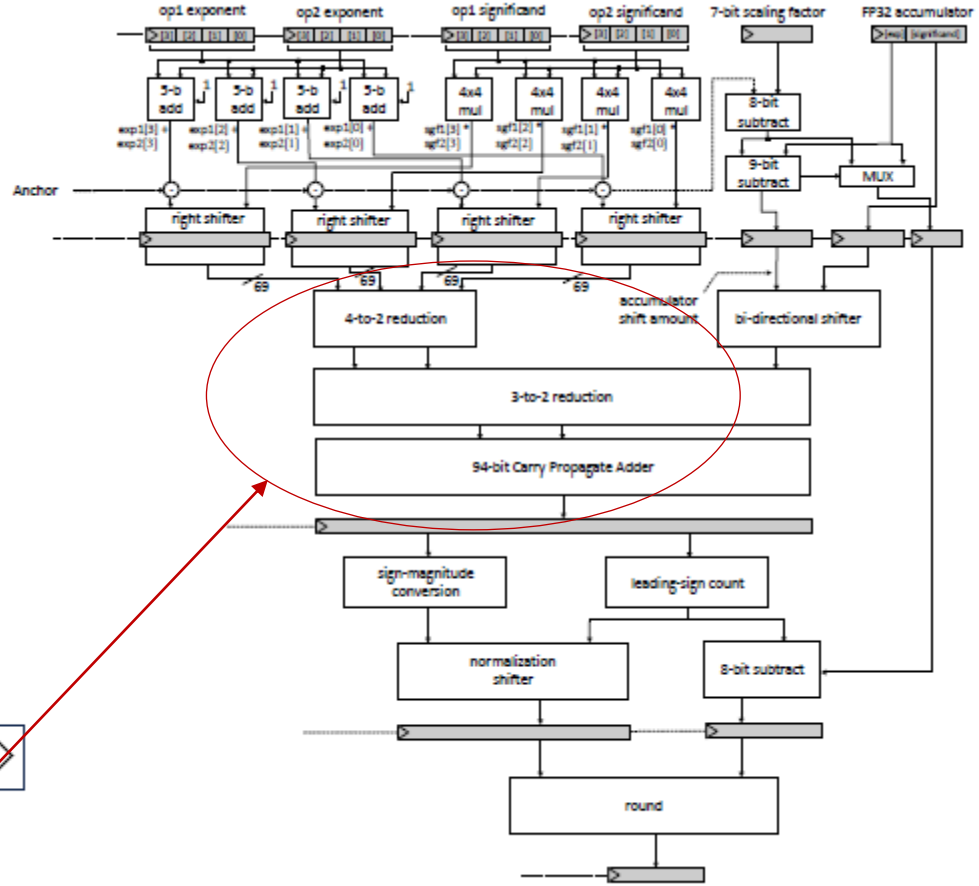




FP8-DOT4-EA

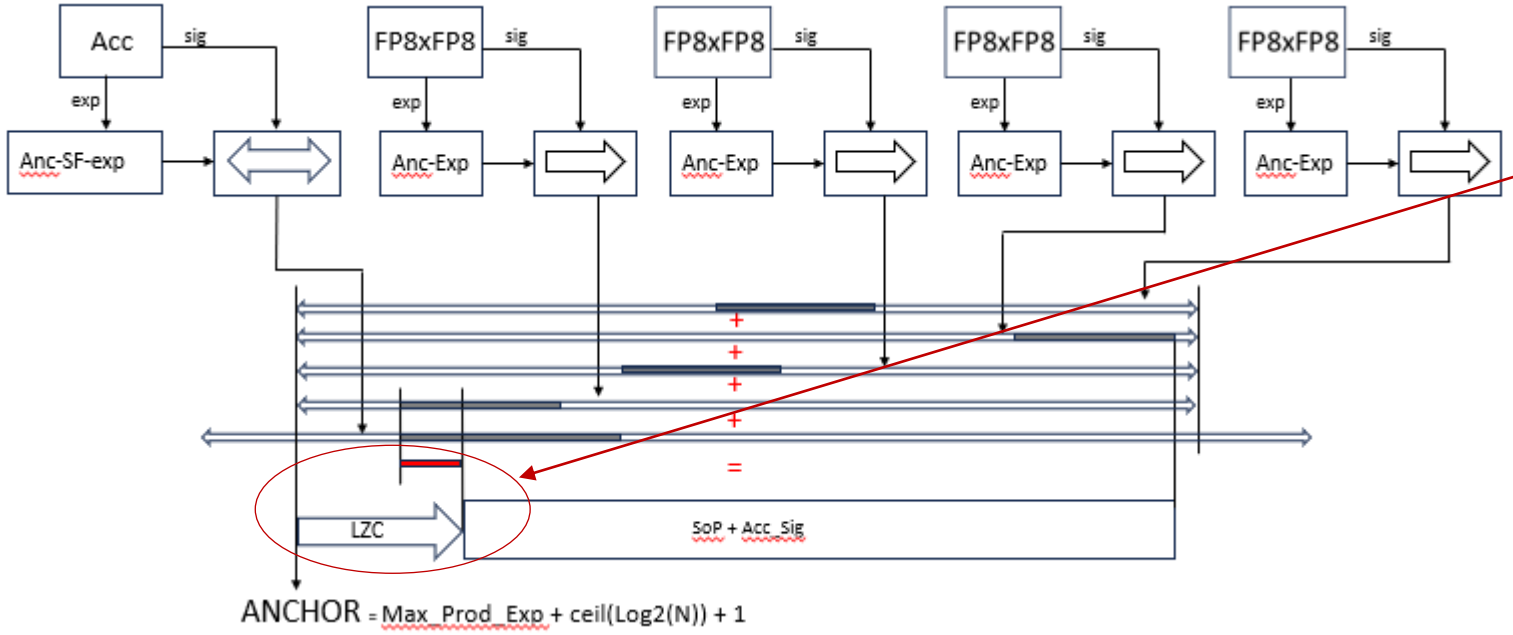
Exact SoP + Acc_Sig

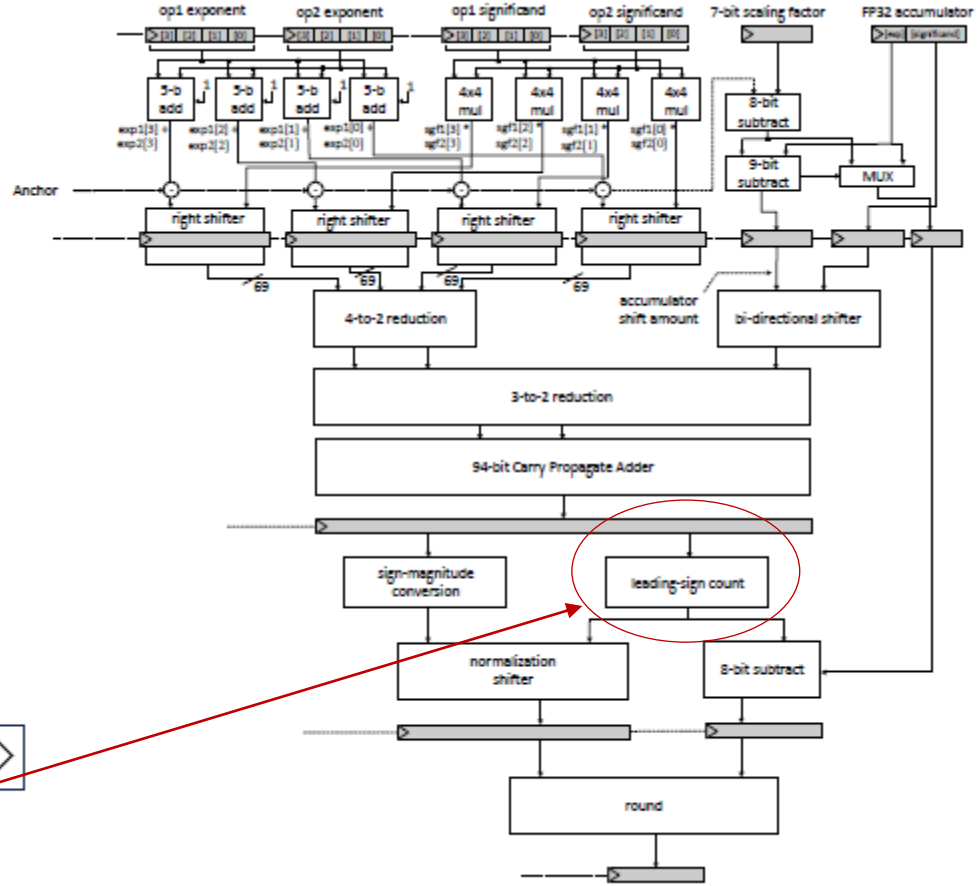




FP8-DOT4-EA

Exp Adjust For LZC





Synthesis Results – 5nm – 3.6 GHz

Unit	Register Count	Area (sq.um)
FP8-DOT4-LA	**345**	**1133**
DOT4 SoP Logic	255	572
FADD32 (Modified)	90	561
FP8-DOT4-EA	**434**	**674**
FP8-DOT4/2-LA	**506**	**1758**
DOT2/4 SoP Logic	407	926
FADD32/16 (Modified)	99	832
FP8-DOT4/2-EA	**624**	**975**
FADD32 for FMA	**N/A**	**512***
FADD32 (Split-Path)	**N/A**	**404***
FADD16 for FMA	**N/A**	**215***

* 3.4 GHz

arm

Thank You
Danke
Gracias
Grazie
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה
ధన్యవాదములు