THE LANDSCAPE OF RISC-V FLOATING POINT SUPPORT WITH BF16 AT THE CENTRE

Kenneth C. Rovers
CPU Architecture, Imagination Technologies

Imagination

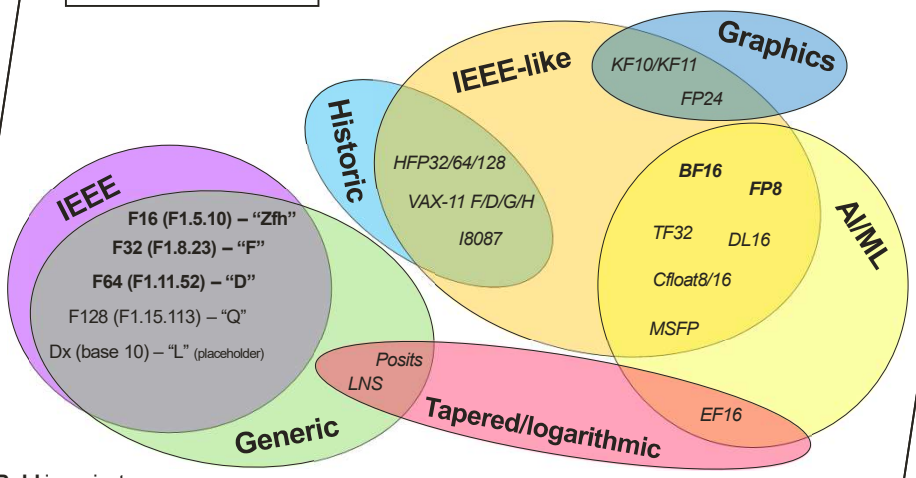
APPROACH

- Explore floating point landscape
- Consider all formats, support only mainstream
- Identify BF16 as an unsupported mainstream format
- Expose BF16 as non-standard
- Analyse BF16 use-cases and ISA support
- Provide recommendations

MAINSTREAM

- IEEE-754 F16/F32/F64
- BF16
- FP8 up and coming

LANDSCAPE

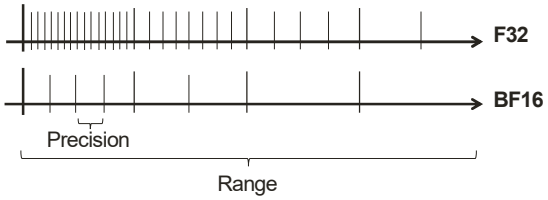


Bold is mainstream
Italic is not supported by RISC-V

BF16 FORMAT

- BF16 is a truncated F32

- Same range, less precision



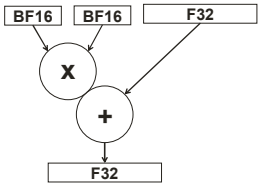
- Not an official standard
  - Introduced by the Google Tensor Processing Unit (TPU)
  - Not IEEE-754 compliant, subnormal support, rounding mode, and exception handling all undefined

BF16 USE CASES

- As a storage format (conversions only)

Half the storage requirements compared to F32

- For efficient matrix multiplications (+ FMA)



Quarter of the area of a F32 fused multiply-add

- As an arithmetic format (all operations)
  - All operations are in BF16 for further area savings
  - Mirrors existing Zfh/F/D extensions
  - Supported by e.g. CUDA and StableHLO/XLA
  - Internal analysis has shown a 20 times smaller error rate than half precision numbers on some benchmarks giving results with less than 0.2% error rate differences against F32

BF16 ISA SUPPORT

- Google TPU, Intel AVX-512_BF16, Armv8.2-A all flush subnormals
- Armv8.2-A extended BF16 makes it optional, and NVIDIA supports subnormals
- Rounding and exceptions handling differ

ISA	Instr.	Subnormal	Rounding
Google	CNV, FMA	flush	?
Intel	CNV, FMA	flush	RNE
ARM	CNV, FMA, MAT	flush	sel cnv, RTO
ARM_E	CNV, FMA, MAT	sel	sel
Nvidia	CNV, FMA, MAT	support	RTZ

RECOMMENDATIONS

- RISC-V should provide a ratified **BF16 extension** (and FP8)
- All **three use cases** should be supported as (sub-) extensions (as planned)
- **Flushing** subnormals should be the default (change)
- BF16 trades precision for implementation efficiency, so **RTZ** should be the default rounding mode
- Enable non-flushing and other rounding modes as **options**