

MXFP4 Microscaling Formats & Neural Quantization Analysis

Block-Scaled 4-Bit Floating Point Arithmetic, Underflow Cliffs, and Activation Outliers

Research Desk: Lesty Tech AI Compute Systems Desk | Date of Assessment: October 2026 | Distribution: Public Technical Release

1. The Precision vs. Memory Bandwidth Wall

Serving massive foundation models (400B+ parameters) is strictly memory-bandwidth bound. Moving from 16-bit floating point (FP16/BF16) to 4-bit representation reduces model weight memory footprints by 75% and quadruples mathematical tensor throughput on modern matrix execution units.

2. The Microscaling Format (MX) Architecture

Standard FP4 (E2M1) provides only 1 sign bit, 2 exponent bits, and 1 mantissa bit, mapping to only 15 representable values. This narrow dynamic range causes catastrophic gradient underflow.

The OCP Microscaling (MX) standard resolves this by grouping 32 contiguous floating-point numbers into a microscopic block sharing an 8-bit scale factor (E8M0). Individual weights are stored in ultra-dense FP4, while the shared scale factor preserves numerical magnitude.

3. The Outlier Underflow Cliff

CRITICAL: Generative transformers exhibit massive activation outliers—individual hidden dimensions carrying values 80x to 150x greater than the layer average. In an MXFP4 block, a single massive outlier forces the shared 8-bit scale factor to inflate, causing all remaining 31 normal values in the block to truncate straight to digital zero, degrading mathematical perplexity.

Empirical Laboratory & Architectural Benchmark Telemetry

Format	Bits per Element	Dynamic Range	LLM Perplexity Degradation
Baseline (FP16 Reference)	16 bits	~1e-38 to ~3e38	Baseline Reference (0.00)
FP8 (E4M3)	8 bits	~2e-7 to 448	< 0.08 Perplexity Increase
FP4 (E2M1)	4 bits	0.0625 to 6.0	+3.85 Perplexity (Model Failure)
OCP Block 32)	4.25 bits (with scale)	Matches FP8 Range	< 0.18 Perplexity (Production Viable)
Sensitive Layers	Attention QK Projections	Requires Hybrid FP8/MXFP4 Routing	

