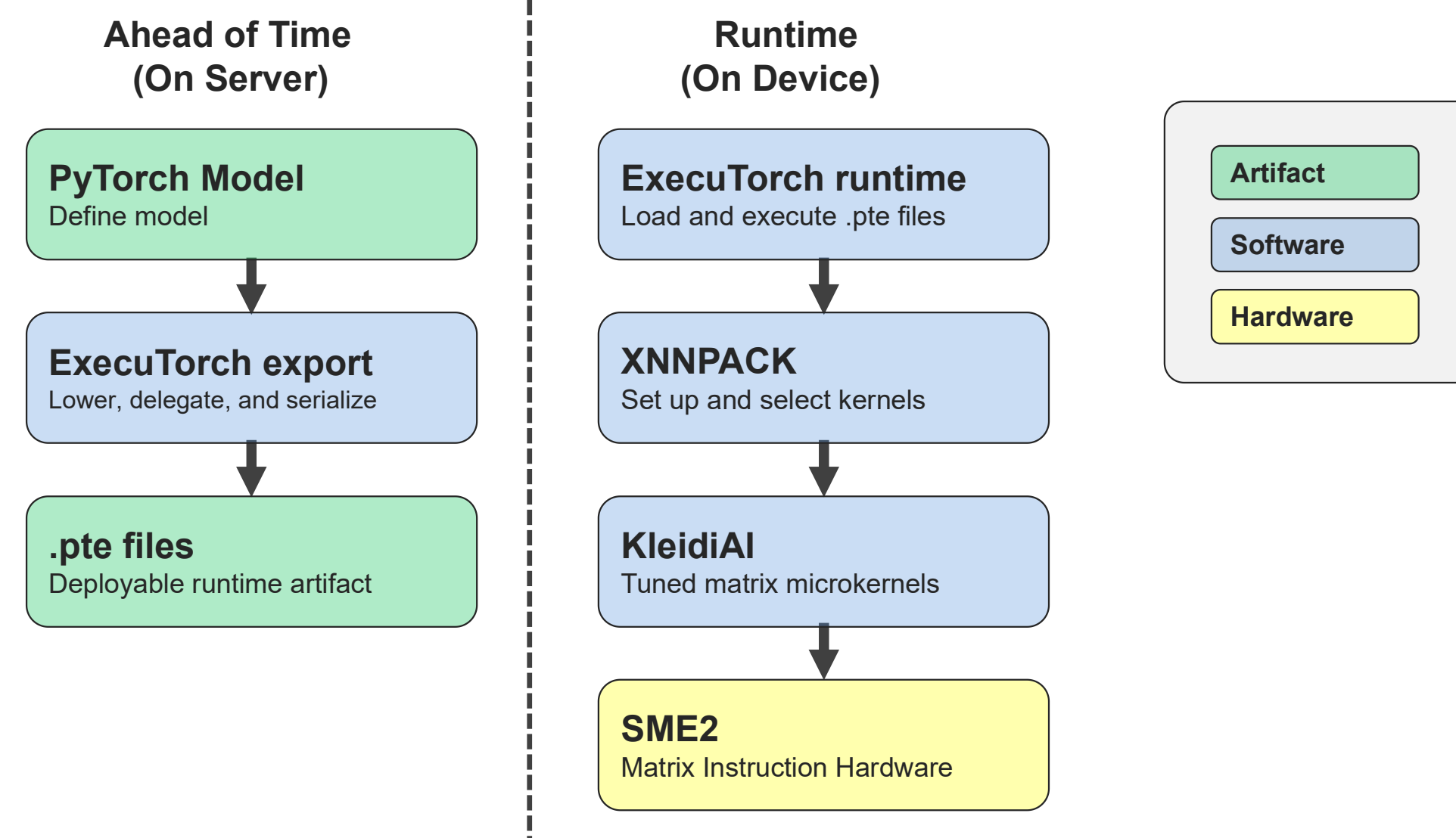


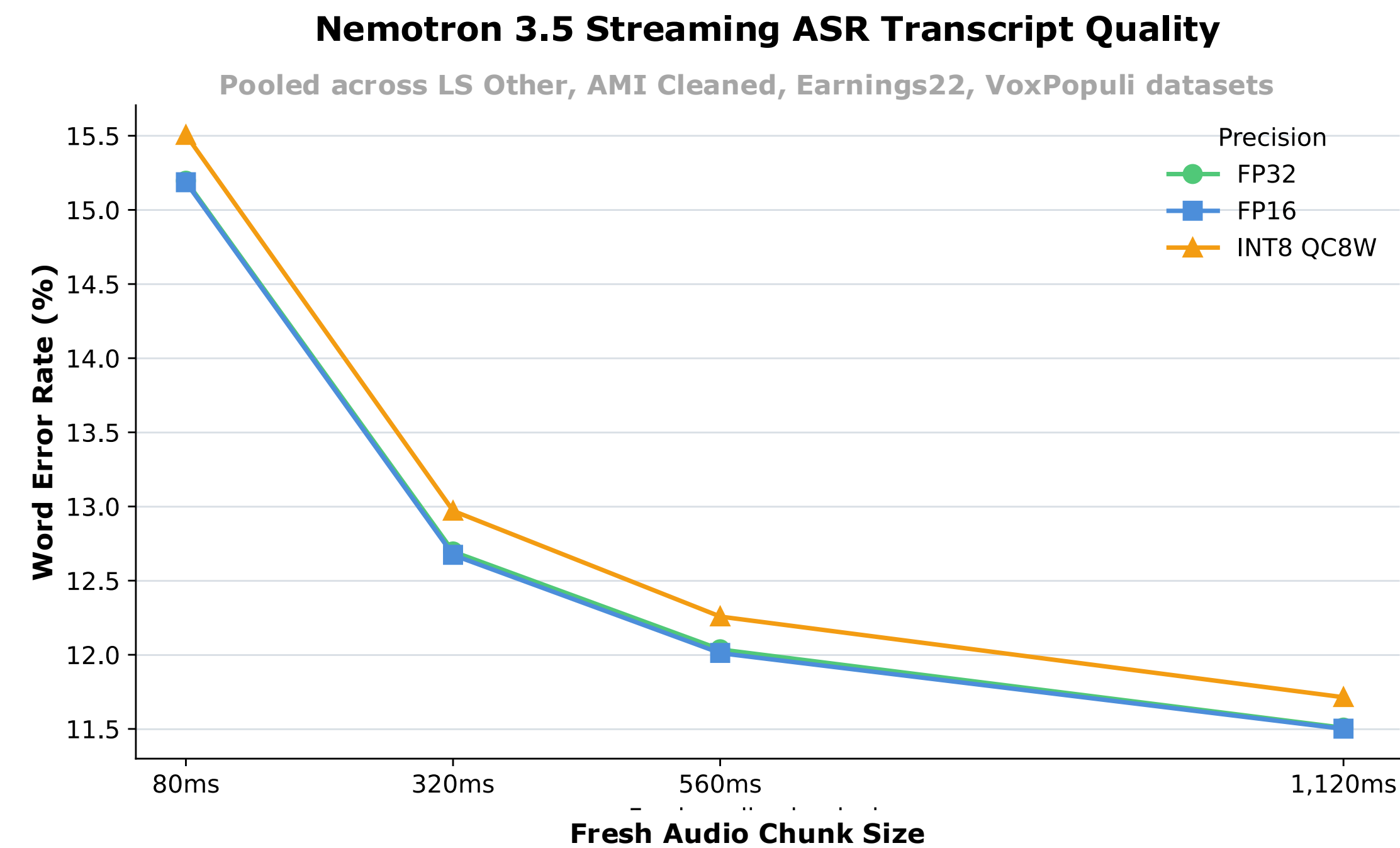
1. From PyTorch to CPU Instructions

Tech Stack



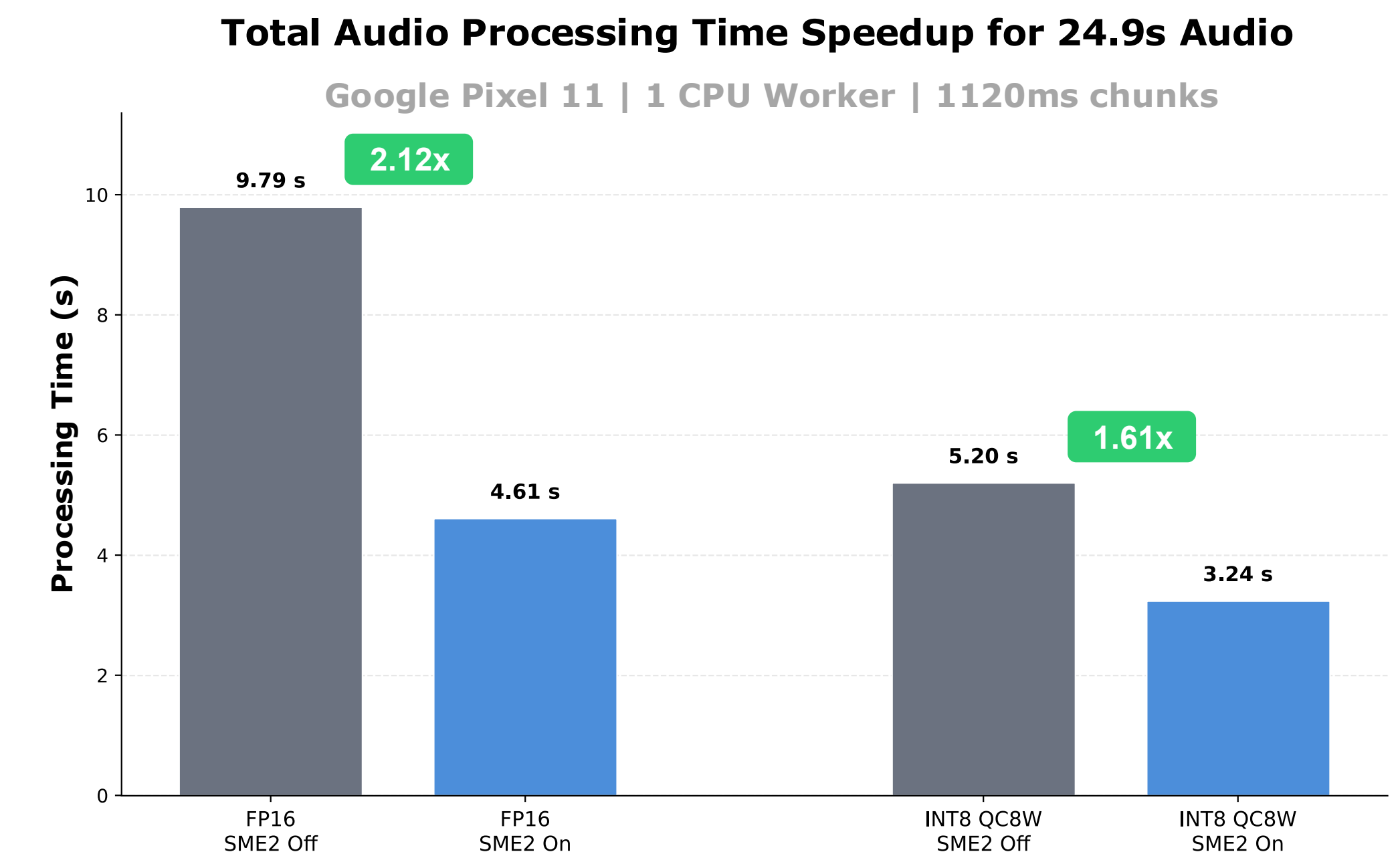
Pipeline: ExecuTorch connects PyTorch models to optimized Arm CPU kernels.

2. Chunk size and transcript quality



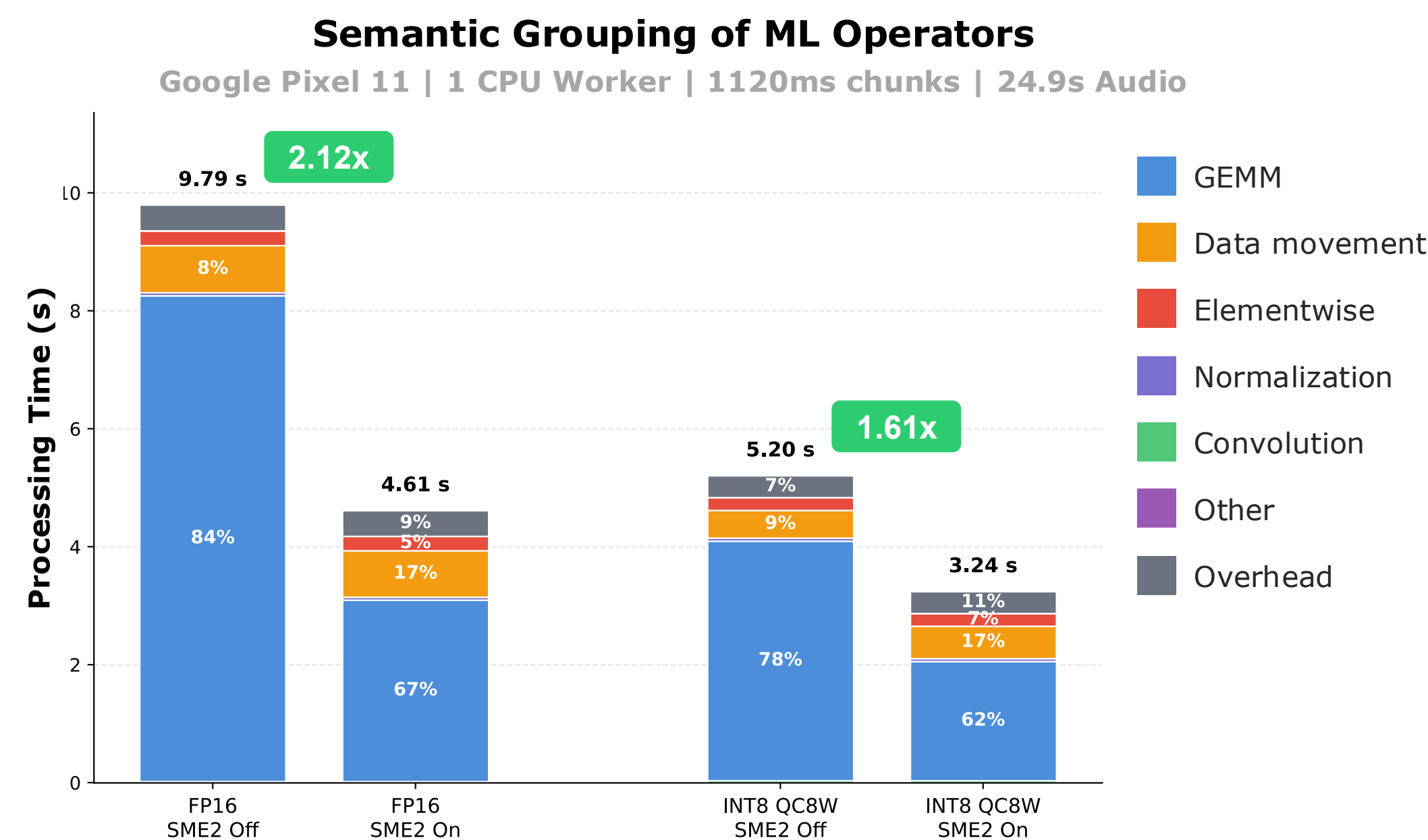
Quality: Longer chunks improve final-transcript accuracy.

3. SME2 audio processing speedup



Latency: SME2 cuts total processing time across precision modes.

4. Find the bottleneck with ETDump



Bottleneck: GEMM drives most of the speedup. Data movement remains.

5. Which kernels ran?

Visible in Arm Streamline

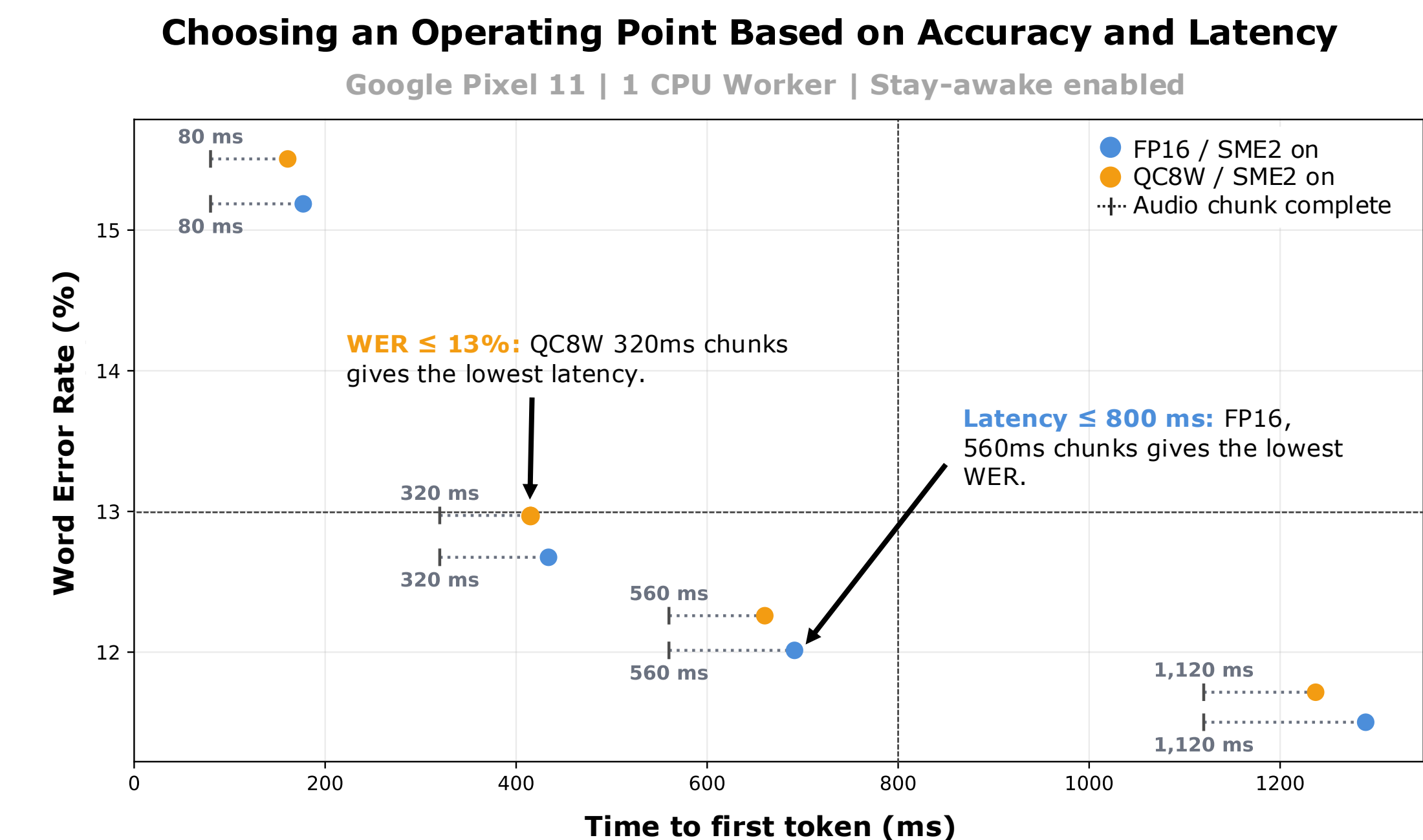
Precision	SME2	GEMM
FP16	True	KleidiAI FP16 SME2 kai_kernel_matmul_clamp_f16_f16p2vlx2_f16p2vlx2_2vlx2vl_sme2_mopa
FP16	False	XNNPACK FP16 NEON xnn_f16_gemm_minmax_ukernel_6x16__asm_aarch64_neonfp16arith_cortex_a75
QC8W	True	KleidiAI quantized SME2 kai_kernel_matmul_clamp_f32_qai8dpx1vlx4_qsi8cxp4vlx4_1vlx4vl_sme2_mopa KleidiAI FP32 SME2 kai_kernel_matmul_clamp_f32_f32p2vlx1_f32p2vlx1biasf32_sme2_mopa
QC8W	False	KleidiAI NEON I8MM kai_run_matmul_clamp_f32_qai8dpx4x8_qsi8cxp4x8_16x4_neon_i8mm

FMOPA
 Floating-point Outer Product and Accumulate

SMOPA
 Signed Integer Sum of Outer Products and Accumulate

Evidence: Native samples confirm SME2 kernels ran; some work uses other paths.

6. Accuracy and Latency Tradeoffs



Decision: Choose chunk size and precision together to meet your targets.