Accelerating Linear System Solvers using Mixed-Precision Iterative Refinement

Mustafa Albahrani

CS380 - GPU and GPGPU Programming
Prof. Markus Hadwiger

December 14, 2025

The Problem: Speed vs. Accuracy

- **Scientific Computing** relies on Double Precision (FP64).
- **Hardware Trend**: Modern GPUs (FP16/TF32) are optimized for AI.

NVIDIA A100 Specs:

- FP64 Peak: **19.5 TFLOPS**
- FP16 Tensor Peak: **312 TFLOPS** (**16x Faster**)

Question: Can we solve $Ax = b$ at FP16 speeds but retain FP64 accuracy?

Solution: Mixed-Precision Iterative Refinement

Core Idea: "Guess fast, check carefully."

1. **Factorize**: $LU \approx A$ in FP16/TF32 (Fastest part)
2. **Solve**: $x_0 \approx U^{-1}L^{-1}b$ (Low accuracy)
3. **Refine Loop**:
   - Compute residual $r = b - Ax_k$ in **FP64**
   - Solve correction $Ad = r$ in FP32
   - Update $x_{k+1} = x_k + d$ in **FP64**

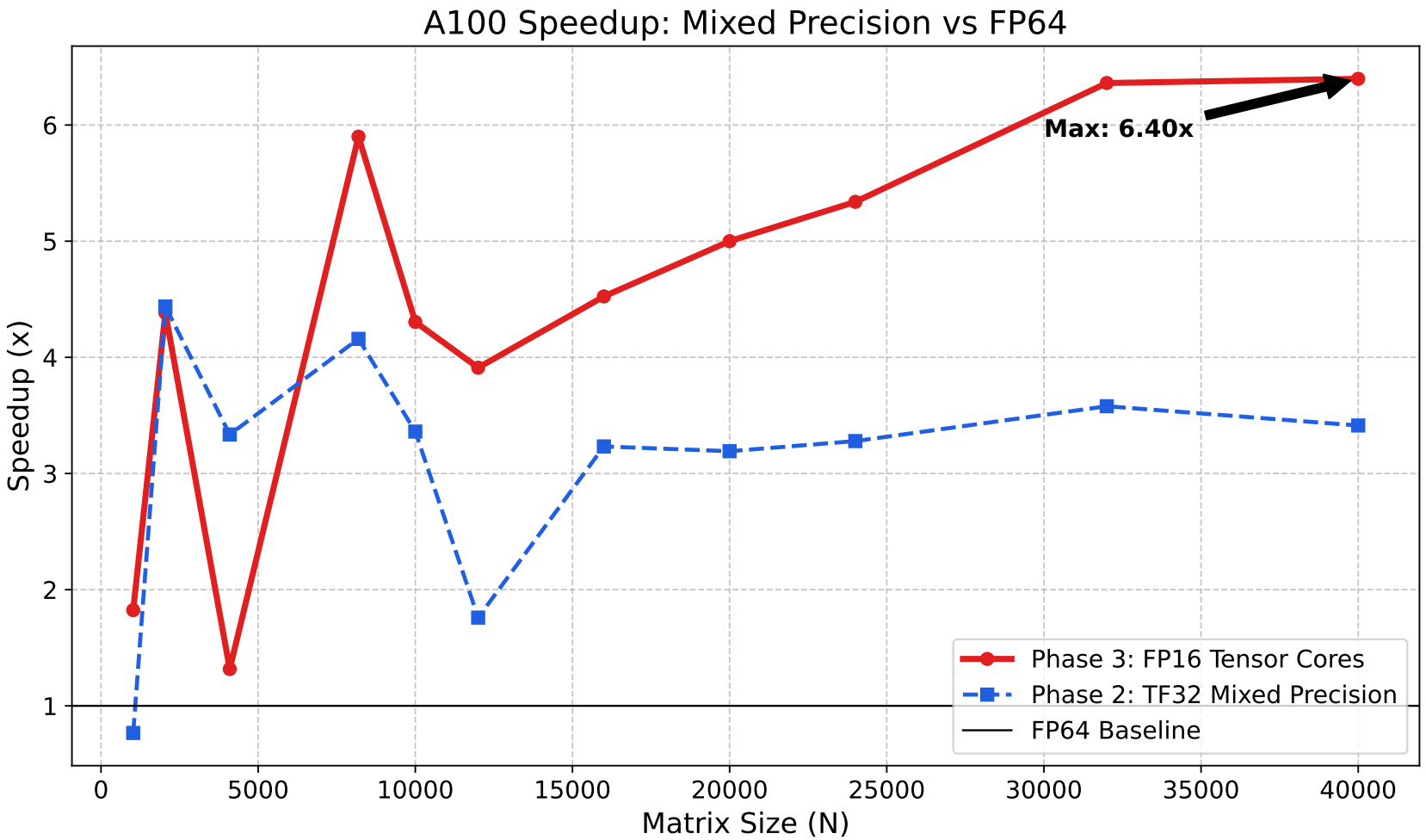
Phase 2: The "Free Lunch" (TF32)

- Implemented using standard `cusolverDnSgetrf` (FP32).
- **Discovery:** On A100, standard FP32 functions automatically use **TF32 Tensor Cores**.
- **Result:** $\approx 3.0\times$ Speedup over FP64.
- **Limitation:** Good, but we can go faster (FP16).

Phase 3: Manual Block LU (Main Contribution)

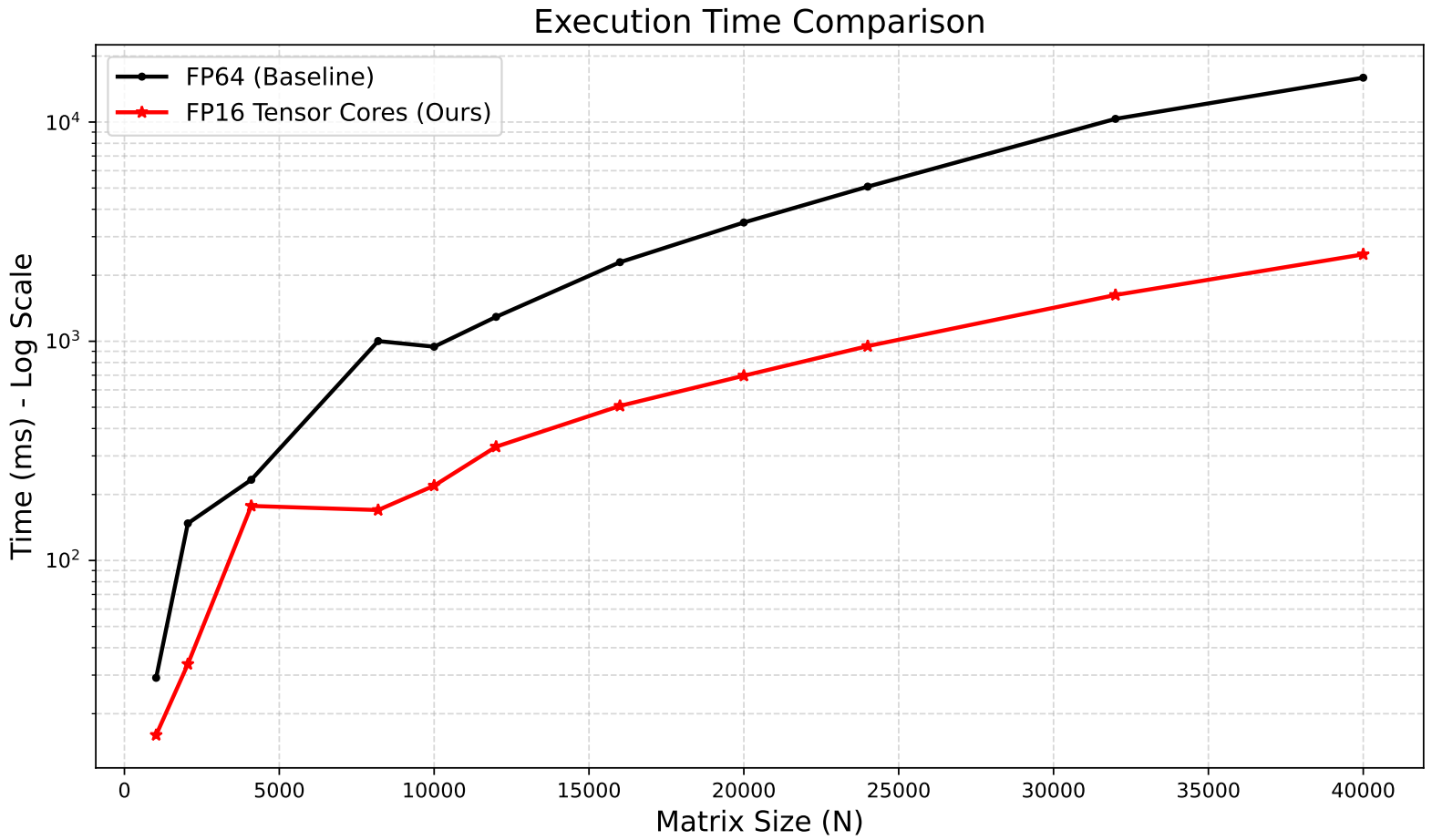
- **Implementation:** Custom Block LU Solver.
- **Key Techniques:**
  - Explicitly use `cublasGemmEx` with `CUDA_R_16F`.
  - Force `CUBLAS_TENSOR_OP_MATH` mode.
  - **Stability Fix:** "Tall Panel" factorization with Global Pivoting to prevent NaNs in FP16.

Results: The Speedup



Max Speedup: **6.40x** at $N = 40,000$.

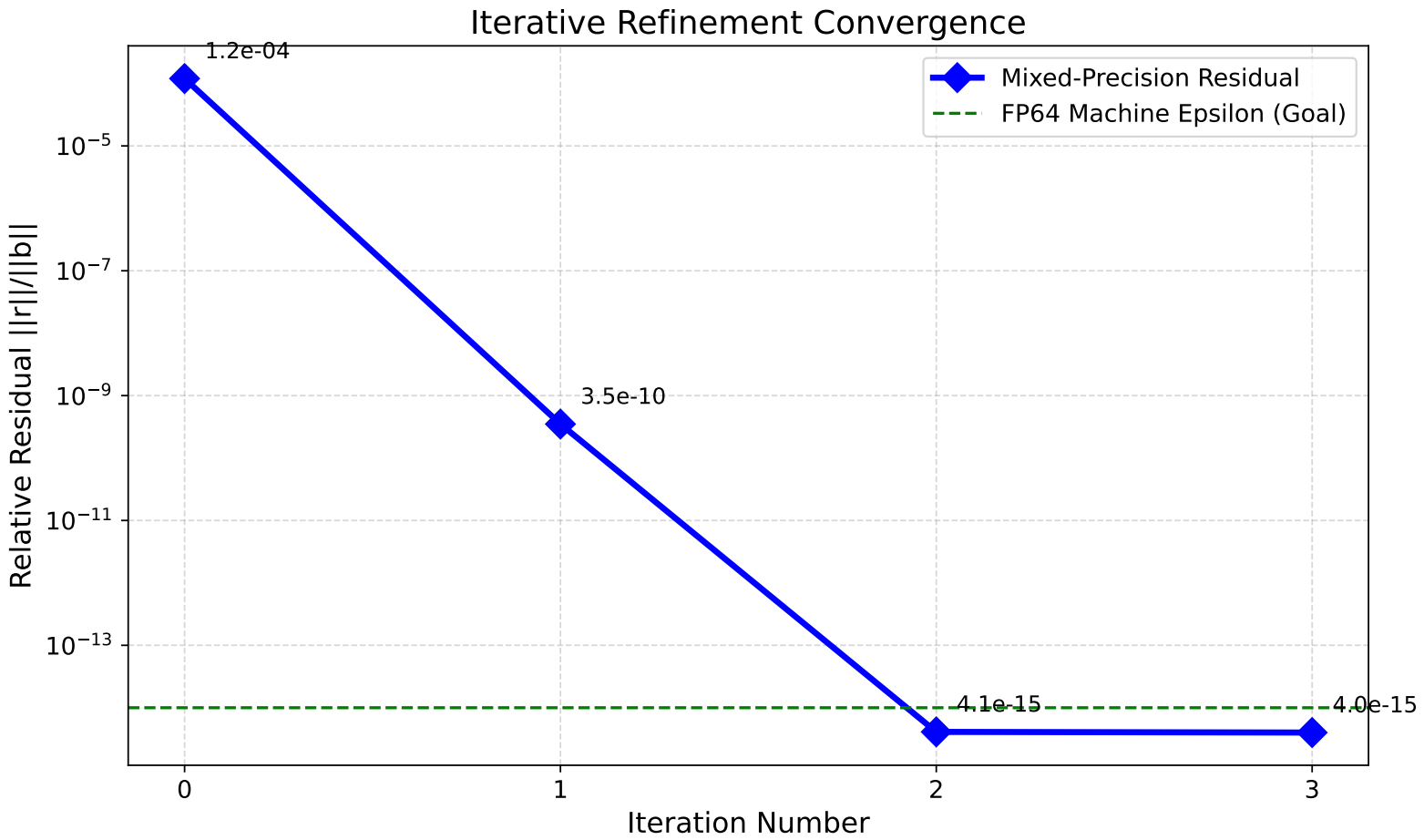
Performance Breakdown (Log Scale)



Note the order-of-magnitude difference in slope.

Convergence: The "Refinement" Magic

- Does using FP16 ruin accuracy? **No.**
- **Graph Below:** Shows residual drop per iteration ($10^{-4} \rightarrow 10^{-15}$).

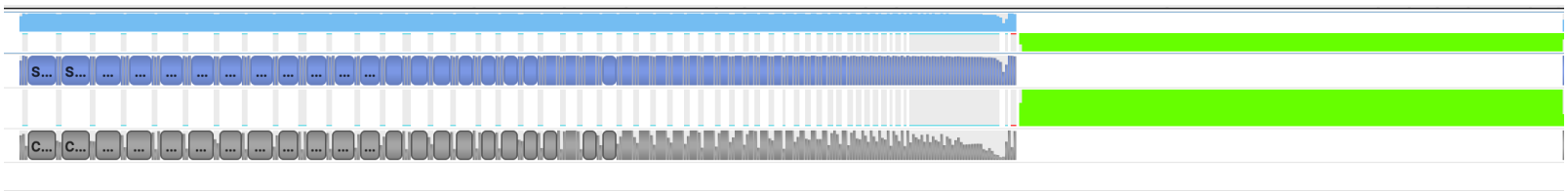


Recovers **Full FP64 Accuracy** in just 2-3 iterations.

Hardware Verification (Nsight Compute)

Tensor Core Usage:

- To demonstrate we hit hardware limits, we used Nsight.
- **Result:** 90% Compute Utilization.
- **Timeline:** Profile dominated by massive GEMM kernels.



Conclusion

- Successfully implemented a Mixed-Precision Solver from scratch.
- **Performance:** 6.4x Speedup over optimized FP64.
- **Accuracy:** Maintained 10^{-14} error (Full Double Precision).
- **Impact:** Validates FP16 Tensor Cores for high-precision scientific workloads.

Thank You!

Question?