HIGH PERFORMANCE COMPUTING PROJECT: GPU IMPLEMENTATIONS AND OPTIMIZATIONS

PRESENTED BY:

IBRAHIM AWAIS

TAHA

MUHAMMAD RAHAT SHAFI

V1: SEQUENTIAL IMPLEMENTATION

- Baseline version provided

- Simple CPU implementation

- Used for performance comparison with other versions

V2: NAIVE GPU IMPLEMENTATION

- Ported the sequential version to GPU

- Basic parallelization using CUDA kernels

- No optimization applied

- Provides initial speedup over CPU

V3: OPTIMIZED GPU IMPLEMENTATION

- Launch Configuration: Optimized thread/block sizing
- Occupancy: Maximized resource utilization per SM
- Communication Optimization: Reduced memory transfers
- Memory Hierarchy Optimization: Used shared memory effectively

V4: TENSOR CORE OPTIMIZATION

- Built upon V3 optimizations

- Utilized Tensor Cores for matrix operations

- Used __half data types and "wmma"

- Achieved significant performance boost

V5: OPENACC OPTIMIZATION

- SIMPLIFIED PARALLELISM USING OPENACC DIRECTIVES

- IMPROVED PORTABILITY AND MAINTAINABILITY

- COMPILER-MANAGED OPTIMIZATIONS

- BALANCED EASE OF DEVELOPMENT WITH PERFORMANCE

THE END

ANY QUESTIONS?