LLMIR System Architecture

Application Layer (vLLM / SGLang)

Frontend Converters

- PyTorch Models
- vLLM Integration
- SGLang Integration

MLIR Optimization Pipeline

- LLM Dialect Operations
- KV Cache Optimization
- Multi-precision Passes

LLMIR Compiler Core

LLM Dialect

- PagedKVCacheType
- ShardedTensorType
- QuantizedTensorType
- llm.paged_attention

Optimization Passes

- KV Cache Fusion
- Quantization-Aware
- Parallelization
- Memory Optimization

Backend Code Generators

CUDA Backend | CPU Backend | ROCm Backend

Execution Layer (CUDA / ROCm / CPU)