



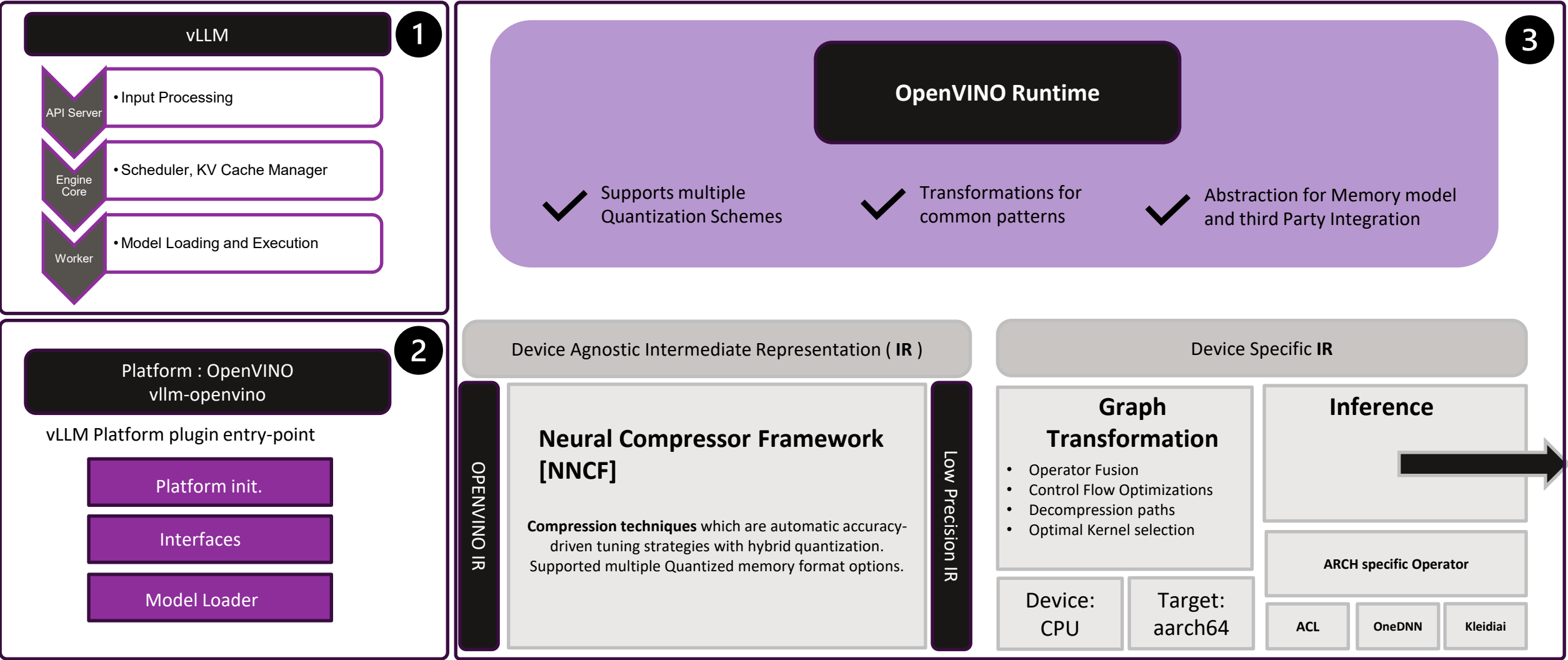
**CONFERENCE**

—NORTH AMERICA 2026—

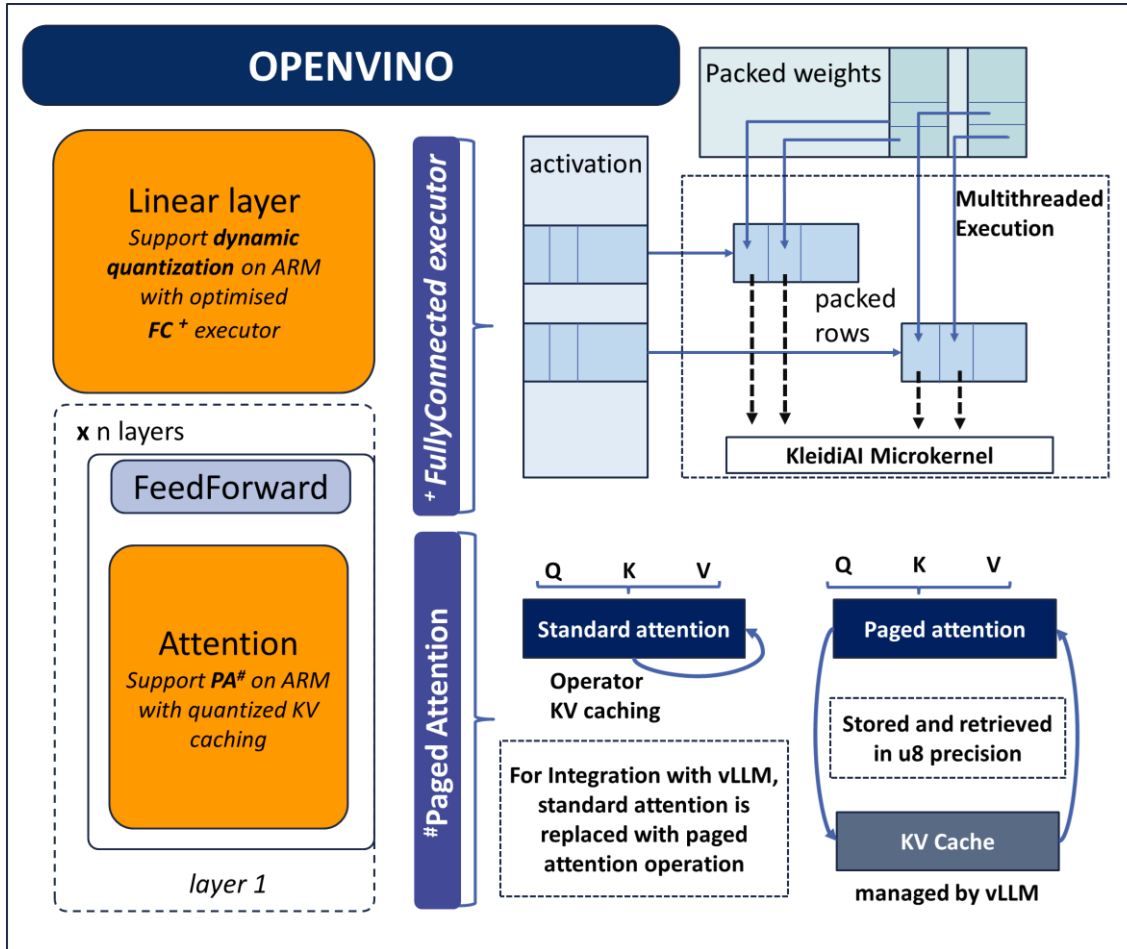
# Efficient MoE LLM Inference on Arm with vLLM and OpenVINO

N Maajid Khan, Ashwin Sekhar, Abhishek Jain  
Fujitsu Research of India

# vLLM - OpenVINO Ecosystem



# OpenVINO Optimizations for ARM for LLM Inference



## Linear layers

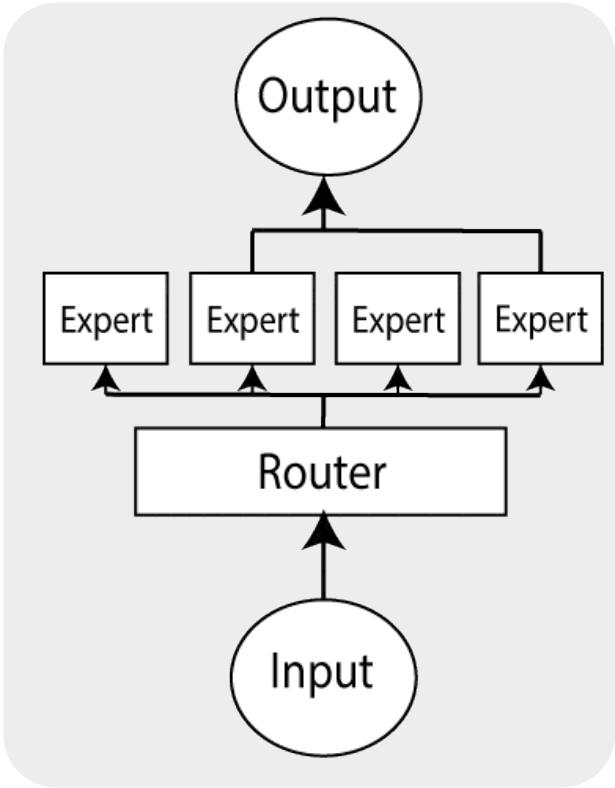
- ✓ Kernel Selection\* Based on Hardware support
- ✓ Supports Low precision weight formats INT8 / 4
- ✓ Fused Packing and Execution

## Attention Layers

- ✓ Vectorized with **NEON/SVE**
- ✓ Supports Paged-Attention implementation
- ✓ Supports KV caches in INT8



# Mixture of Experts (MoE) Inference on CPUs



### Benefit on CPU

Sparse Activations
 Memory Capacity

### Opportunities

Model development
 
 Performance tuning

Abstraction

Modularity

Python

Conditional Constructs

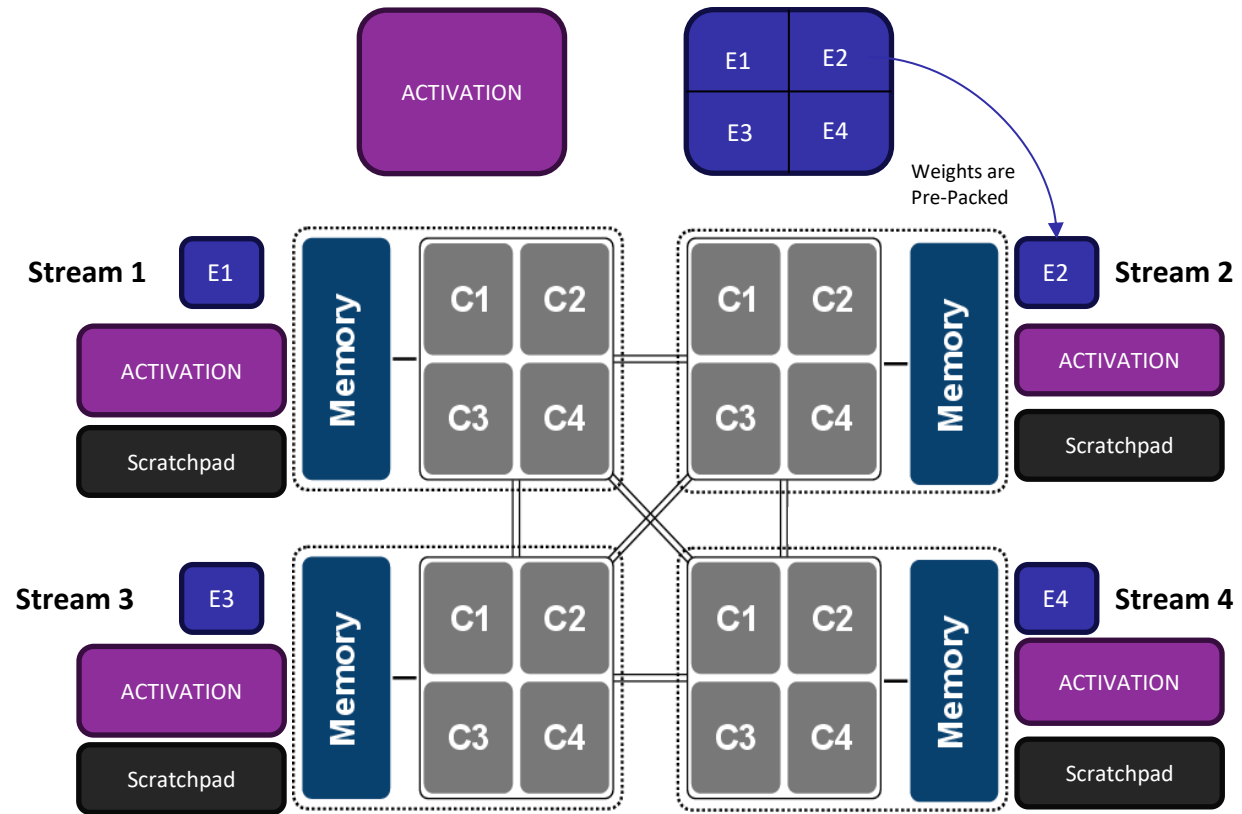
Pattern matching

Op Fusion

Cache Locality

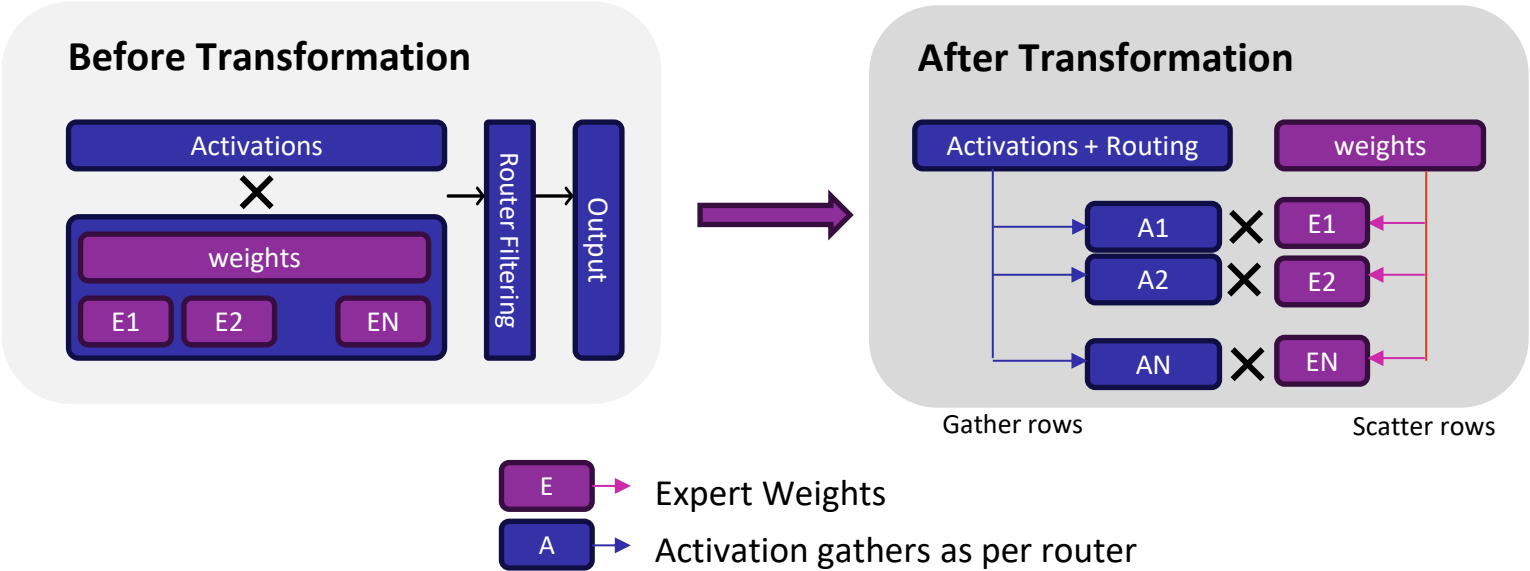
Constant Folding

# Executor Design in OpenVINO with NUMA streams



- OpenVINO CPU Plugin**  
 Detect CPU topology and organize available cores
- Stream Allocation**  
 Identify NUMA nodes → assign streams → Pin to Core Groups (CG)
- Memory Manager**  
 Abstraction that internally manages memory partitions per core group
- Fully-Connected Executor**  
 Manages Weight Pre-packing and splitting across NUMA (CG)

# GatherMatmul Fused operation for MoE



- Independent Expert Prepacking:** Each expert's weights are prepacked independently, enabling efficient execution without repacking weights at runtime.
- Routing-Aware Execution:** The executor uses routing information to pack activations for cache-friendly GEMM execution

## Results:

**Machine :** AWS Graviton 3E  
**Model :** GPT-OSS 20B  
**Software :** vLLM with OpenVINO backend

|            | Before Transformation | After Transformation* |
|------------|-----------------------|-----------------------|
| Throughput | 40 TPS                | 250 TPS               |
| 3.6x       | TTFT Reduction        | 6.2x TPOT Reduction   |



**THANK YOU !!**