# CFMoE: Cache-Friendly Mixture of Experts Model Serving System on Edge Devices

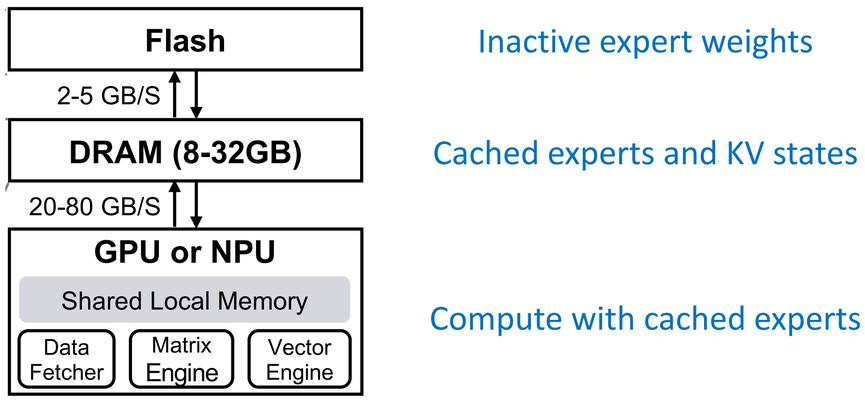
Haoyu Zhang[12]†, Hongyi Wang[13]†, Jianwei Zheng[1]*
Yu Li[1], Tianyang Jiang[1], Zhenhua Li[3], Ning Jia[1]

[1] Huawei Technologies Co., Ltd [2] Peking University [3] Tsinghua University

† Equal contribution * Corresponding author

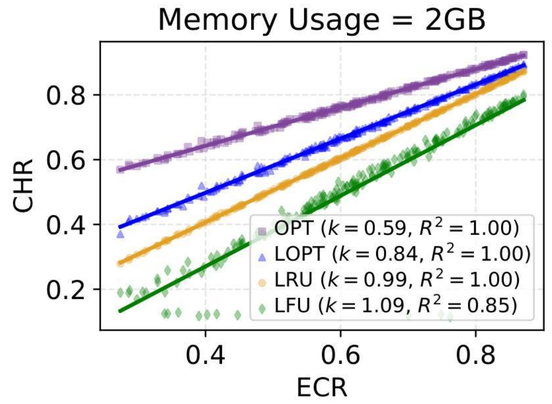
## Background & Motivation

- On-device MoE serving faces a **severe memory bottleneck**
- **Expert offloading** mitigates this: keep inactive experts in Flash and load activated experts into DRAM
- Poor locality → frequent Flash–DRAM I/O → slower inference




## Routing locality limits cache hits

- **Load-balancing auxiliary loss** for cloud expert parallelism promotes even token assignments across experts within a batch
- This can reduce **consecutive-token expert overlap** and cache hits
- Even optimal eviction remains limited by the model's activation locality

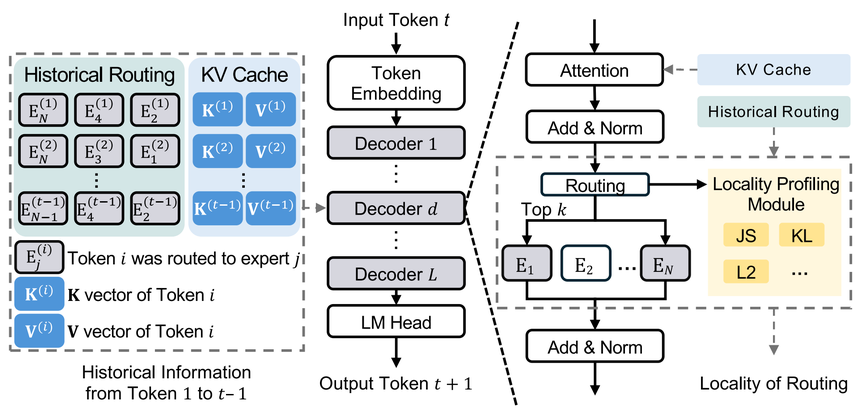


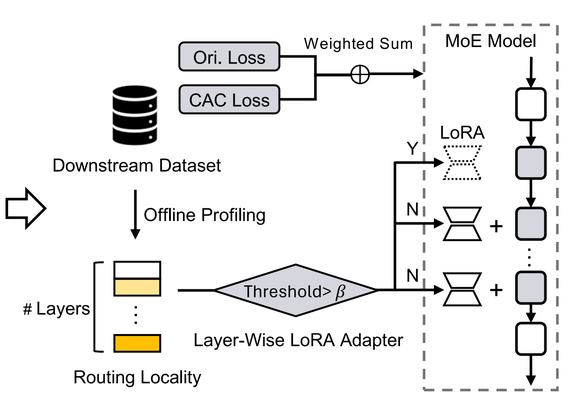

ECR: expert consistency rate CHR: cache hit rate

## Data-Aware Fine-Tuning (DAFT)

Only LoRA adapters are trainable

- **Cache Affinity Constraint (CAC)** aligns adjacent-token routing with a JS loss
- Profile layer locality and insert **LoRA adapters** only into low-locality layers
- Optimize **task loss + CAC loss** with the backbone model frozen

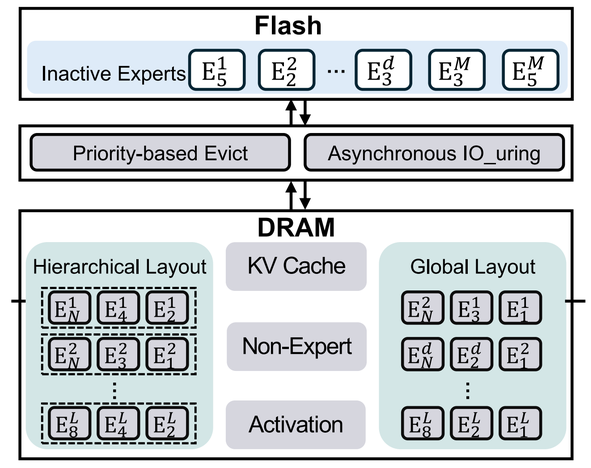



## PD Disaggregated Cache Management

Phase-aware layouts on one edge device

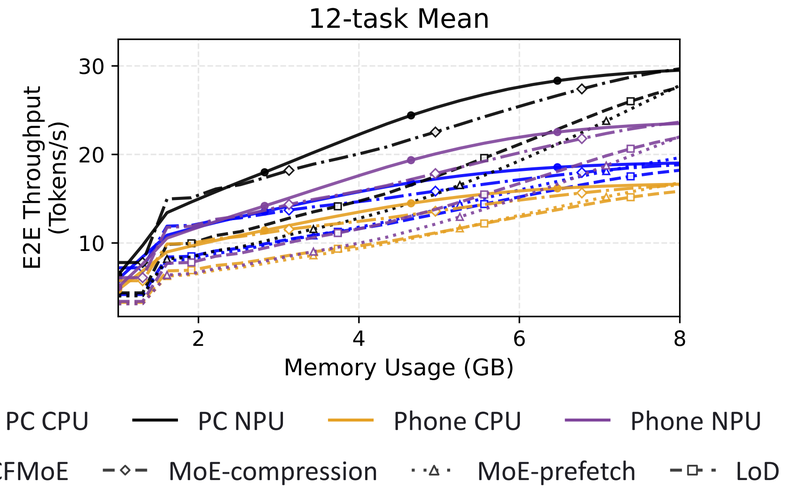
- **Prefill:** use hierarchical, layer-wise cache partitions for batched inputs
- **Decoding:** share a global expert cache across layers to reuse idle space
- **Priority-based eviction** + **asynchronous I/O** overlap transfers with computation



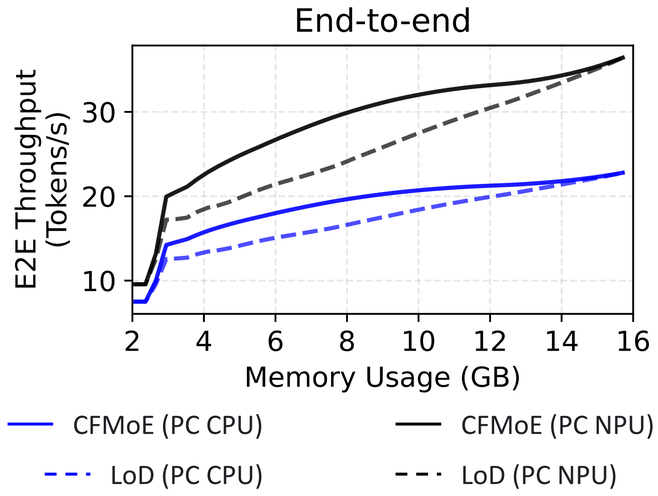

## Evaluation & Results

- **Models:** DeepSeek-MoE-16B and Qwen3-30B-A3B with INT4 expert weights
- **Datasets:** 12 benchmarks for Q&A (3), math (4) and commonsense (5)
- **Devices:** Ultra 5 125H PC (16 GB LPDDR5), Kirin 9020 phone (12 GB RAM), CPU / NPU
- **Budgets:** DeepSeek-MoE-16B 2 / 4 / 6 / 8 GB (PC + phone), Qwen3-30B-A3B 4 / 8 / 12 / 16 GB (PC only)

### DeepSeek-MoE-16B




### Qwen3-30B-A3B




**Model precision & average memory saving**

| Model | LoD / Pref. | CFMoE | Memory saving |
|---|---|---|---|
| DeepSeek-MoE-16B | 0.526 | **0.529** | **32.9%** |
| Qwen3-30B-A3B | 0.691 | **0.689** | **24.2%** |

Mean score: ROUGE-L / accuracy. Memory saving: average vs LoD at matched throughput.

ACM MobiCom 2026 Paper #803

Code & sample evaluation data are available at https://cfmoe.github.io