

Mixture-of-Depths (MoD): Dynamic Compute Allocation in Transformers

Algorithmic Analysis: Non-Uniform FLOP Allocation, Top-K Routing, and 50% Latency Reductions

Published by Lesty Tech AI Systems Research • Date: October 21, 2026 • Lesty.tech Architectural Engineering Dossier

1. The Compute Inefficiency of Homogeneous Transformers

Standard autoregressive transformer models (GPT-4, Llama 3) apply a homogeneous computational graph: every token traverses every layer (all 32, 64, or 80 layers) and executes an identical number of floating-point operations (FLOPs). However, information theory dictates that natural language tokens possess highly variable entropy: simple tokens (commas, common conjunctions) require minimal compute, while complex reasoning steps demand deep algorithmic transformation.

2. Mathematical Formulation of Top-K Token Routing

Mixture-of-Depths (MoD, introduced by Google DeepMind) sets a static computational budget per layer: only a fraction of tokens (defined by dynamic capacity factor C , typically 12.5% to 50%) are permitted to execute self-attention and MLP blocks. For each layer l , a router predicts a scalar weight $R_l(x_i)$ for token x_i . The top- K tokens with the highest routing weights are passed into the transformer block, while remaining tokens bypass the layer entirely via an unmolested residual connection.

3. KV-Cache Compression and Memory Bandwidth Savings

Because skipped tokens do not participate in self-attention at layer l , they do not generate or append key-value vectors to that layer's KV-cache. This directly compresses KV-cache memory consumption by 50% to 70%, dramatically reducing memory bandwidth demand during autoregressive decoding and permitting larger batch sizes in production serving.

4. Training Dynamics and Co-design with Mixture-of-Experts (MoDE)

MoD models are trained end-to-end using auxiliary load-balancing losses to prevent router collapse. Furthermore, combining Mixture-of-Depths with Mixture-of-Experts (MoDE) creates an orthogonal routing space: MoD decides WHETHER a token receives compute, while MoE decides WHICH specialized expert processes it, achieving frontier-level model accuracy at a fraction of training and inference compute.

Key Engineering Metrics & Reference Specifications

Model Architecture	Compute per Token	Routing Mechanism	KV-Cache Growth Rate	Inference Speedup
Dense Transformer (Llama 3)	Fixed 100% FLOPs (Uniform)	None (Homogeneous)	1.0x (Every Token / Layer)	1.0x (Baseline)

Mixture-of-Experts (MoE)	Variable Expert Compute	Top-2 Expert Router	1.0x (Standard KV)	1.5x - 2.0x Throughput
Mixture-of-Depths (MoD)	Static Capacity Budget (C=0.5)	Top-K Layer Router	0.5x (50% Cache Skipping)	1.6x - 2.2x Latency Reduction
MoDE (Combined MoD + MoE)	Dynamic Depth & Expert	Hierarchical 2D Routing	0.4x - 0.5x Compressed	2.5x - 3.2x Efficiency

Engineering Notice: This whitepaper is an official technical companion resource released in conjunction with Lesty.tech research publications. All structural schematics, formulas, and empirical microarchitectural data are strictly verified against IEEE, JEDEC, USENIX, and ACM research standards.