Novel Scaling Laws for MoE Architectures: A Theoretical Analysis

Wesley Medford

February 25, 2025

Abstract

This paper presents fundamental scaling laws governing the relationship between expert granularity, cache efficiency, and bus bandwidth in Mixture-of-Experts (MoE) model architectures. Through rigorous mathematical analysis, we demonstrate that increasing expert count while decreasing individual expert size can theoretically lead to exponentially improved cache efficiency, even under bandwidth-constrained scenarios. This relationship is governed by specific scaling laws that we derive. The theoretical framework suggests that models with smaller but more numerous experts could achieve superior performance while significantly reducing memory requirements, potentially enabling efficient deployment of trillion-parameter models without requiring full VRAM residency.

1 Introduction

The rapid advancement of large language models has driven hardware requirements to unprecedented levels, particularly in VRAM capacity. While MoE architectures improve parameter efficiency through sparse activation, conventional deployment assumes model parameters must reside entirely in accelerator memory. This assumption has led to costly solutions involving GPU meshing or reduced-quality edge deployments.

Recent work by Skliar et al. [5] demonstrated that cache-aware routing strategies can significantly improve MoE inference efficiency. Building on these insights, this paper presents a theoretical framework that predicts and explains the relationship between expert granularity and system performance.

1.1 Core Hypothesis

This work proposes that the relationship between expert count, size, and cache efficiency follows specific scaling laws that can be theoretically derived. The central hypothesis is that increasing the number of experts while reducing their individual size leads to exponentially improved cache efficiency, governed by the relationship:

$$\text{Performance Gain} \propto \frac{N_{total}^{\alpha}}{s_{exp}^{\beta}} \tag{1}$$

where N_{total} is the total number of experts, s_{exp} is the size of each expert, and α, β are architecture-specific constants derived in this work.

1.2 Key Contributions

1. Derivation of fundamental scaling laws governing the relationship between expert granularity and cache efficiency
2. Mathematical framework for predicting MoE model performance under varying hardware constraints
3. Theoretical analysis of the interaction between model architecture and hardware capabilities
4. Implementation methodology framework for modern hardware systems

2 Theoretical Framework

2.1 Fundamental Scaling Laws

Building on the work of He [3] regarding expert scaling in MoE models, we derive a comprehensive set of scaling laws that govern the relationship between expert configuration and system performance.

For a given layer with C_{exp} experts needed per forward pass, we define:

- N as the number of experts per layer
- s_{exp} as the size of each expert in bytes
- bus as the available bus bandwidth (bytes/second)
- t_{exp} as the time to process a single expert
- P_{cached} as the fraction of total experts present in cache
- P_{miss} as the probability of a cache miss

The fundamental relationship between these parameters can be expressed as:

$$P_{miss} = (1 - P_{cached})^{C_{exp}} \tag{2}$$

To account for expert loading time, the effective miss penalty is defined as:

$$x_{miss} = P_{miss} \times \left(\frac{s_{exp}}{bus}\right) \tag{3}$$

2.2 Cache Efficiency Analysis and Scaling Constants

The probability of finding necessary experts in cache improves exponentially with increased expert count, following:

$$P_{hit} = 1 - (1 - P_{cached})^{C_{exp}} \tag{4}$$

This relationship leads to the first key scaling law:

$$\text{Cache Efficiency} \propto \exp(k \cdot N_{total}) \tag{5}$$

where k is a system-specific constant determined by memory hierarchy latencies.

The constants α and β in the scaling law arise from two key relationships:

1. α **(Expert Count Scaling Factor):**

$$\alpha = \log_2\!\left(\frac{t_{cache\_miss}}{t_{cache\_hit}}\right)$$

   Represents how performance scales with increased expert count. Derived from cache coherency overhead, where t_{cache_miss} and t_{cache_hit} are the respective latencies.

2. β **(Expert Size Penalty Factor):**

$$\beta = \frac{\log(bus_{bandwidth})}{\log(cache_{bandwidth})}$$

   Captures how larger experts impact bandwidth utilization. Calculated from memory system characteristics, reflecting the penalty of moving larger experts through the memory hierarchy.

These constants can be theoretically predicted for a given architecture by analyzing:

- Memory hierarchy latencies
- Bus bandwidths between different memory tiers
- Cache coherency protocol overhead
- Memory controller queuing characteristics

2.3 Bandwidth Utilization Model

Following ProMOE's findings [6], bandwidth utilization is modeled as:

$$\text{Effective Bandwidth} = bus \cdot (1 - P_{miss}) + \text{cache bandwidth} \cdot P_{hit} \tag{6}$$

This leads to the second key scaling law:

$$\text{Throughput} = \frac{N_{active}}{t_{proc} + x_{miss}} \tag{7}$$

where N_{active} is the number of active experts per forward pass.

3 System Architecture Considerations

3.1 Hardware Considerations

The theoretical framework applies to modern accelerator architectures with:

- High-bandwidth GPU memory
- Lower-bandwidth CPU memory
- High-speed interconnects for coherent memory access

3.2 Model Architecture Implications

The theoretical framework suggests several architectural considerations:

- Trade-offs between expert count and size
- Impact of layer count on overall system performance
- Memory hierarchy utilization patterns

3.3 Integration with Existing Systems

The framework can be integrated with existing cache-aware routing strategies through:

1. Stride-based prefetching for expert parameters
2. Chunked prefetching for bandwidth optimization
3. Early preemption for critical path optimization

4 Theoretical Predictions

4.1 Expected System Behavior

The theoretical framework predicts several key behaviors:

- **Cache Efficiency:**
  - Exponential improvement in cache hit rates as expert count increases
  - Inverse relationship between expert size and system performance
  - Memory hierarchy utilization patterns
- **Performance Characteristics:**
  - Throughput scaling with expert count and size
  - Latency implications of cache efficiency
  - Bandwidth utilization patterns

4.2 Hardware Interaction Predictions

The framework predicts specific interaction patterns with modern hardware:

- **Memory Hierarchy Utilization:**
  - Optimal cache utilization patterns
  - Memory tier transition behaviors
  - Bandwidth utilization characteristics
- **System-Level Effects:**
  - Memory coherency impact
  - Interconnect utilization patterns
  - Overall system efficiency characteristics

5 Discussion and Future Work

The theoretical framework provides several key insights:

1. **Theoretical Implications:**
   - Mathematical basis for expert scaling decisions
   - Bandwidth utilization optimization strategies
   - Architecture-specific performance predictions
2. **Practical Applications:**
   - Guidelines for expert count/size trade-offs
   - Optimization strategies for different hardware configurations
   - Memory hierarchy design recommendations
3. **Future Directions:**
   - Comprehensive empirical validation of the theoretical framework using modern hardware architectures
   - Experimental verification of scaling laws across different expert configurations
   - Real-world performance measurements and comparison with theoretical predictions
   - Investigation of dynamic expert sizing
   - Integration with emerging memory technologies
   - Development of reference implementations to validate theoretical claims

6 Conclusion

This work establishes fundamental scaling laws governing the relationship between expert granularity and system performance in MoE architectures. The theoretical framework provides a foundation for understanding and optimizing model deployment across different hardware configurations. While the mathematical relationships derived suggest significant potential for improving performance through expert count/size trade-offs, empirical validation of these theoretical predictions represents a crucial next step. This work opens up exciting opportunities for future research and practical implementation, particularly in validating and refining these scaling laws through real-world experimentation.

References

[1] Dai, D., Deng, C., et al. (2024). DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv:2401.06066*.

[2] Eliseev, A., & Mazur, D. (2023). Fast inference of mixture-of-experts language models with offloading. *arXiv:2312.17238*.

[3] He, X. O. (2024). Mixture of a million experts. *arXiv:2407.04153*.

[4] Kurtic, E., Marques, A., et al. (2024). Give me BF16 or give me death? Accuracy-performance trade-offs in LLM quantization. *arXiv:2411.02355*.

[5] Skliar, A., van Rozendaal, T., et al. (2024). Mixture of cache-conditional experts for efficient mobile device inference. *arXiv:2412.00099*.

[6] Song, X., Liu, Z., et al. (2024). ProMoE: Fast MoE-based LLM Serving using Proactive Caching. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.