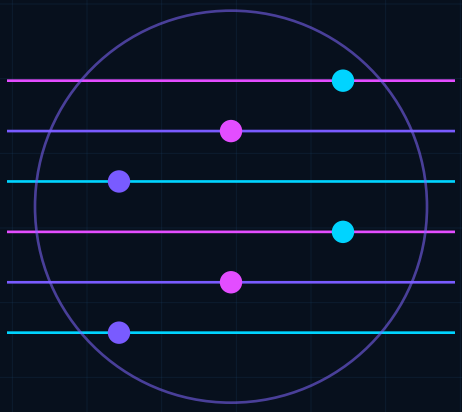
MODEL ARCHITECTURE

Mixture of Experts (MoE) Routing Dynamics

Current-source review, evidence-bounded adoption decision, explicit limitations, reproducibility controls, and preserved source research note.



PERMANENT ID	EVIDENCE	ADOPTION	FRESHNESS
RES-013	A-	MONITOR AND BENCHMARK	STABLE WATCH

Freshness review: 2026-09-17 / Next scheduled review: 2027-03-16

RESEARCH CONTROL SURFACE

Required Research Fields

Each field remains independently reviewable; evidence strength does not imply production authority.

EVIDENCE GRADE — **A-**

Strength of the retained source set and technical basis.

SOURCE AUTHORITY — **2 registered authorities**

and Efficient Sparsity; GShard: Scaling Giant Models with Conditional Computation

FRESHNESS — **STABLE WATCH**

Reviewed 2026-09-17; dynamic sources require event-driven revalidation.

CONFLICTS — **VISIBLE / NOT SILENT**

Differences between specification, vendor behavior, research claims, and reproduced evidence must remain explicit.

LIMITATIONS — **EXPLICIT**

No certification, procurement approval, legal conclusion, or unbounded production claim is created by this report.

REPRODUCIBILITY — **REQUIRED**

Material claims require exact versions, representative data, failure cases, artifacts, and repeatable acceptance criteria.

ADOPTION POSTURE — **MONITOR AND BENCHMARK**

Treat MoE as a model architecture characteristic, not an automatic platform advantage. Benchmark expert routing, memory footprint, throughput, quality variance, capacity loss, and deployment complexity.

REVIEW TRIGGER — **2027-03-16**

Scheduled at 180 days; earlier on material source, vulnerability, implementation, or contradictory-evidence change.

SOURCE AUTHORITY / FRESHNESS

Authority Register and Freshness Delta

Primary-source lineage is preserved; current-source changes are separated from unperformed implementation claims.

AUTHORITY 01

and Efficient Sparsity

https://arxiv.org/abs/2101.03961

AUTHORITY 02

GShard: Scaling Giant Models with Conditional Computation

https://arxiv.org/abs/2006.16668

2026-09-17 FRESHNESS DELTA

Primary-source register retained and freshness controls advanced to the current review date. No new product or laboratory performance claim is introduced without reproduced evidence.

Freshness policy: scheduled review + immediate review on source supersession, material release, vulnerability, retraction, or contradictory evidence.

Evidence Must Survive Contact With Reality

- Source authority is not implementation evidence. Product behavior, scale, security, interoperability, and operator fit remain environment-specific.
- Conflicting evidence is recorded rather than averaged away. Specification behavior, vendor implementation, preprint results, and production observations are separate evidence classes.
- Quantitative claims are informative until reproduced under documented hardware, software, data, topology, load, and failure conditions.
- Adoption posture is reversible. A stronger source set can coexist with a pilot-only or research-only recommendation when maturity, safety, economics, or operability remain unresolved.

ADOPTION / VALIDATION / REVIEW

Decision Gate and Validation Plan

ADOPTION POSTURE

MONITOR AND BENCHMARK

Treat MoE as a model architecture characteristic, not an automatic platform advantage. Benchmark expert routing, memory footprint, throughput, quality variance, capacity loss, and deployment complexity.

1 / SOURCE	Pin exact versions, publication state, retrieved artifacts, and source hashes.
2 / PROTOTYPE	Demonstrate the core mechanism with representative data and instrumented execution.
3 / FAILURE	Exercise malformed input, dependency loss, stale state, overload, recovery, rollback, and misuse.
4 / BASELINE	Compare against an incumbent, classical, or simpler architecture using the same workload.
5 / ACCEPT	Record acceptance criteria, residual limitations, owner, expiration, and revalidation triggers.

EXPIRATION / REVIEW TRIGGER

Next scheduled review: 2027-03-16 (180-day cadence). Review sooner after material source revision, breaking implementation release, critical vulnerability, retraction, benchmark contradiction, changed workload, or changed legal/safety boundary.

8. Complete Source Research Note

SOURCE-LINEAGE NOTICE

The following material preserves the complete original source note, excluding only the conversational wrapper and outer Markdown fence. Legacy status labels and absolute language are retained for traceability and do not override the candidate status, limitations, or adoption decision in this edition.

Document ID: RES-013 Version: 1.0.0 (Definitive Registry Edition) Status: Published Research Note Classification: Public Organization: Mythos Systems (MYTHOS AI) Owner: Aaron Stovall Effective Date: 2026-07-16 Applies To: AI researchers, distributed systems engineers, and hardware architects designing trillion-parameter sparse models, expert parallelism clusters, and high-throughput inference engines Companion Standards: MYTHOS-AI-0011 (MoE & SSM Architecture Standard), MYTHOS-AI-0014 (Inference Serving & KV-Cache), MYTHOS-HW-0001 (AI GPU Clusters), MYTHOS-EMT-0004 (Silicon Photonics)

The architecture of scaling Large Language Models has hit a physical wall. Dense models (where every parameter processes every token) scale quadratically in compute cost. To achieve frontier intelligence, organizations attempt to train 1-trillion-parameter dense models, which instantly exhausts the compute and memory bandwidth of the largest supercomputers.

Mixture of Experts (MoE) is the architectural paradigm shift that decouples parameter count from compute cost. By utilizing conditional computation—where only a subset of "expert" neural networks process a given token—an MoE model with 1 trillion parameters might only require 20 billion active parameters per token.

However, MoE is not a free lunch. It introduces severe dynamic routing complexities, load-balancing instabilities (router collapse), and crushing interconnect overheads (All-to-All communication). This research note defines the mathematical dynamics of sparse MoE routing. It dissects the mechanics of Top-K gating, auxiliary load-balancing losses, expert capacity tuning, and the physical reality of Expert Parallelism (EP) across distributed GPU clusters. Understanding these dynamics is a prerequisite for building the ultra-efficient, trillion-parameter sovereign AI clusters mandated by the Mythos Hardware Standards.

In a dense Transformer, every token passes through identical Feed-Forward Network (FFN) layers. In an MoE Transformer, the FFN is replaced by N parallel expert networks. A gating network (router) determines which k experts process each token.

1.1 The Top-K Gating Mechanism

The router is a linear transformation that maps the token's hidden state to a logit vector of size N (number of experts).

- The Math: $G(x) = \text{Softmax}(\text{TopK}(x \cdot W_g, k))$
- The router computes the dot product of the token vector x and the gating weight matrix W_g.
- It selects the top k experts (typically $k=1$ or $k=2$) with the highest probabilities.
- The token is dispatched to those specific experts, and the output is the weighted sum of the expert outputs.

1.2 The Conditional Compute Boundary

Because $k \ll N$ (e.g., $k=2$ out of $N=64$), the model utilizes only a fraction of its total parameters per token.

- The Benefit: Inference FLOPs are drastically reduced, allowing sub-100ms latency for massive models.
- The Constraint: All N experts must reside in VRAM. MoE solves the compute bottleneck but exacerbates the memory bottleneck.

Routing tokens dynamically across a cluster of GPUs introduces physical and mathematical constraints that dense models do not face.

2.1 Router Collapse (The Rich Get Richer)

The most common failure mode of MoE training is router collapse. Due to the self-reinforcing nature of gradient descent, the router quickly learns that a few experts are slightly better at processing certain tokens. It routes more tokens to those experts, they receive more gradient updates, they get even better, and the other experts starve.

- The Result: The model degenerates into a tiny dense model (using only 2 out of 64 experts), losing the conditional compute advantage.
- The Mitigation: An Auxiliary Load-Balancing Loss MUST be added to the training objective. This loss function penalizes the model if the standard deviation of expert utilization across a batch is too high, mathematically forcing the router to distribute tokens evenly.

2.2 Expert Capacity and the Drop Penalty

Because the router is dynamic, it is impossible to guarantee an exactly even distribution of tokens to experts. If a batch of 1,000 tokens happens to send 300 tokens to Expert A, but Expert A's compute buffer (capacity) is only sized for 100 tokens, the overflow tokens are dropped.

- The Mechanic: Expert Capacity $C = \left(\frac{T}{N} \right) \times \text{Capacity Factor}$. (Tokens per batch / Number of experts * a buffer multiplier, usually 1.25).
- The Impact: Dropped tokens are passed through a residual connection unprocessed, degrading model quality. If the Capacity Factor is set too high to prevent drops, VRAM is wasted on empty buffers.

2.3 The All-to-All Communication Wall

In Expert Parallelism (EP), experts are distributed across different GPUs to solve the VRAM bottleneck. If GPU 1 holds Experts 1-8, and GPU 2 holds Experts 9-16, a token on GPU 1 that needs Expert 12 must be physically transmitted over the NVLink/InfiniBand fabric to GPU 2, processed, and transmitted back.

- The Impact: This requires an All-to-All collective communication. Unlike All-Reduce (used in data-parallel training), All-to-All is a full many-to-many permutation of memory. It crushes the interconnect. If the network topology (e.g., standard Ethernet) cannot handle the micro-bursts, the GPUs starve and utilization plummets.

To mitigate the constraints of naive Top-K routing, modern MoE architectures (like DeepSeek-V3 or Mixtral) have evolved advanced routing dynamics.

3.1 Expert Choice Routing (Reversing the Dynamic)

Instead of the token choosing the expert, the expert chooses the token.

- The Mechanic: The router computes the affinity matrix, and each expert selects the top T/N tokens globally.
- The Benefit: Load balancing is mathematically guaranteed. Expert capacity is exactly filled; no tokens are dropped.
- The Flaw: Expert choice breaks the causal order of sequences. A token at the end of a sentence might be processed before a token at the beginning, which complicates autoregressive generation.

3.2 Shared Experts (DeepSeek Architecture)

To reduce redundant computation, the model includes "Shared Experts" that process every token, alongside the routed experts.

- The Mechanic: $Output = \text{SharedExpert}(x) + \sum \text{RoutedExperts}(x)$.

- The Benefit: The Shared Expert captures common foundational knowledge (syntax, grammar), allowing the routed experts to specialize deeply in niche domains without re-learning the basics. This drastically improves MoE training stability.

The fundamental shift of MoE is the decoupling of memory and compute scaling.

- Dense Model (e.g., Llama-3-70B): 70 Billion parameters. Every token requires 140 GFLOPs (multiply-add) of compute, and 140 GB of VRAM (in FP16) to store the weights.
- Sparse MoE (e.g., Mixtral 8x7B): 47 Billion parameters total, but only 13 Billion active per token. VRAM requires ~94 GB, but compute requires only 26 GFLOPs per token.

The Scaling Law: If you scale an MoE model from 8 experts to 64 experts, the VRAM requirement scales linearly (8x), but the compute requirement remains almost completely flat. This allows organizations to achieve frontier intelligence (large parameter count) on edge devices or low-power NPUs, provided the VRAM can hold the weights. This is why MoE is the mandated architecture for local-first AI (MYTHOS-EMT-0002) where compute is constrained but VRAM is available.

To deploy trillion-parameter MoE models in production, the Mythos Hardware Standard (MYTHOS-HW-0001) enforces strict topology and routing rules.

5.1 Topology-Aware Expert Placement

The All-to-All communication wall is a physical problem. If Expert A on Node 1 and Expert B on Node 2 are frequently routed to together, the inter-node InfiniBand link becomes a bottleneck.

- The Mitigation: The router MUST be topology-aware. During training and inference, the system analyzes the routing matrix and dynamically migrates experts to minimize cross-node traffic. Highly correlated experts are placed on the same physical node (using NVLink intra-node switches) to bypass the network fabric entirely.

5.2 DeepSeek-style Shared Expert Offloading

For local-first inference (RA-006), holding 64 experts in local VRAM is impossible.

- The Mitigation: The Shared Expert (which processes every token) is permanently pinned in high-speed VRAM. The 64 routed experts are kept in system RAM or NVMe SSDs. Because only $k=2$ experts are needed per token, the system asynchronously pre-fetches the required experts from SSD to VRAM just-in-time, utilizing the PCIe bandwidth efficiently.

5.3 KV-Cache Bounding for MoE

MoE models generate tokens at high speed (low compute FLOPs), which rapidly fills the KV-cache. An MoE inference server (like vLLM) MUST utilize PagedAttention to manage the KV-cache fragmentation. If the cache fills up, the system offloads inactive KV blocks to CPU RAM, ensuring the GPU VRAM is strictly reserved for the massive MoE weights and the active batch of tokens.

Mixture of Experts is not a niche optimization; it is the only physically viable path to scaling AI to trillion-parameter capacities without bankrupting the global power grid. By decoupling parameter count from compute cost, MoE allows organizations to build ultra-intelligent models that run efficiently on constrained hardware. However, mastering MoE requires understanding the harsh realities of router collapse, expert capacity, and the All-to-All interconnect wall. By enforcing topology-aware routing and shared-expert architectures, we transform sparse models from theoretical lab experiments into production-ready sovereign intelligence.

```
A dense model scales quadratically; an MoE scales conditionally.
A router that starves its experts is a dense model in disguise.
An All-to-All collective without topology awareness is a network collapse.

VRAM holds the intelligence; the compute executes the query.
```

```
> Parameters may be abundant.
> Active compute must be absolute.
```

End of Research Note RES-013

9. Source Register

Source	Authority and Status	Freshness Control
Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity https://arxiv.org/abs/2101.03961	Primary research Published research Published: 2021	Validated: 2026-07-22 Review on material revision, withdrawal, supersession, or scheduled report review.
GShard: Scaling Giant Models with Conditional Computation https://arxiv.org/abs/2006.16668	Primary research Published research Published: 2020	Validated: 2026-07-22 Review on material revision, withdrawal, supersession, or scheduled report review.

10. Lineage, Review, and Publication Record

Record	Value
Original source file	RES-013_Mixture_of_Experts_Routing_Dynamics.md
Original source words	1561
Upgrade type	Enterprise research report; source and freshness validation; limitations; adoption decision; reproducibility plan
Generated on	2026-07-22
Current status	Candidate Research Report
Publication authority	Formal Mythos approval not yet recorded
Next review	180 days or event-driven trigger, whichever occurs first