# PUBLIC SUBMISSION

**Received:** May 29, 2025
**Tracking No.** mba-7ba7-1fxq
**Comments Due:** May 28, 2025 **Submission Type:** API

**Docket:** NSF-2025-OGC-0001
NITRD_FRDOC_0001

**Comment On:** NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

**Document:** NSF-2025-OGC-0001-DRAFT-0336
Comment on FR Doc # 2025-07332

## Submitter Information

**Government Agency Type:** State
**Government Agency:** University of Washington

## General Comment

See attached file(s)

## Attachments

NSF-AI-US-RFI-UW

**Title:** Disaggregated Foundation Models: Breaking Monolithic One-size-fits-all Models and Rebuilding from the Pieces
**Authors**: Luis Ceze, Baris Kasikci, Stephanie Wang, Luke Zettlemoyer,

**Organization:** University of Washington, Paul G. Allen School of Computer Science & Engineering



**Introduction.** The United States must lead not only in scaling today's AI systems but also in inventing the next generation of AI architectures. While transformers have revolutionized AI and accelerated the possibility of artificial general intelligence, continuing to scale our current foundation models is no longer sufficient. Continued leadership in AI demands a fundamental shift in AI systems, one that requires redesigning the entire stack, from the hardware to the largest scale model training and the end application.

Previous gains in foundation models have come about from tight codesign of AI and systems, and GPUs in particular. Transformers have succeeded in part because they were designed to scale using parallel and distributed GPUs. Meanwhile, new software systems for training and inference have been built specifically for transformers, making deployments orders of magnitude more efficient.

However, a critical inflection point is approaching that calls for a fundamentally new approach. Single-GPU performance is plateauing, making **scale-out** with multiple accelerators more critical. But scale-out also has its limits. The cost of continually building high-performance AI datacenters with the latest hardware is already prohibitive; further gains will require more efficient use of existing and therefore **heterogeneous** hardware. At large enough scale, physical limitations also mean that **communication** overheads will dominate, especially given current model architectures that rely heavily on synchronous execution. Finally, **data scarcity** is an existential problem for future scaling of current **monolithic** foundation models. Together, these trends point to the need for a new architecture for foundation models, one that is codesigned with the physical infrastructure of the future.

We propose to break open the monolithic architecture used by current foundation models and to rebuild the components into a **disaggregated foundation model**. A

disaggregated foundation model is one that dynamically and selectively activates different components within the model, with the goal of using exactly the right amount of compute to produce an accurate prediction for a given input. This is a rethinking of the foundation model architecture that is designed from the ground up for modularity, flexible sparsity, and massive scale. By breaking up the monolithic architecture used by current foundation models, we unlock new opportunities for leveraging arbitrary sparsity within the model to further improve model capabilities and efficiency while using the same data. Meanwhile, by designing a model whose individual components can be highly specialized, we will unlock new opportunities in systems efficiency by fully leveraging disaggregated hardware.

**Opportunities.** There is a wide gap between current AI systems and models and our vision of a disaggregated foundation model. Current foundation models are monolithic and support only limited forms of model sparsity, using the same amount of compute for all inputs. This is both inefficient, for simple inputs that don't require complex reasoning, and insufficient, for harder inputs that do. Current AI deployments, especially proprietary ones, are also tightly coupled to a specific and usually homogeneous physical infrastructure, with limited ability to adapt. Finally, inference is proving to be a critical component of further training improvements, but remains an under-explored avenue, especially as systems support for executing inference and training together lags.

We propose a research program centered on the disaggregated foundation model to address these gaps. The key problem is in codesign of the disaggregated foundation model and the underlying system stack. For example, how should the model incorporate current system load and topology to decide how to process a given input? How can we design the model components to encourage specialization and therefore gain efficiency through custom systems and hardware? How can we close the loop between the model and the system, using the model itself to improve system efficiency? These questions will require a concerted redesign of every layer of the AI systems stack:

**Algorithms for AI:** The disaggregated foundation model is designed to support arbitrary model sparsity, by learning how to use the optimal amount of compute to process a given input. This requires a rethinking of the foundation model architecture that takes into account both system characteristics and available algorithms. We lay out several directions for future disaggregated foundation models:

1. *Model sparsity*: While current models have leveraged sparsity with impressive gains, the extent is still limited, with a fixed number of layers, experts, and experts to activate, all of which are typically homogeneous. Pushing model sparsity to the extreme, we envision a foundation model that can use arbitrary control flow to choose a submodel to activate. Furthermore, individual

components of the submodel may be heterogeneous, allowing the AI system to right-size its compute to its input.

2. *Multimodality and agentic workloads:* The disaggregated foundation model is a unique fit for application domains that require a composition of reasoning methods. These applications could be supported by a single advanced model that is designed to adapt to the specific workload characteristics. Meanwhile, these directions require deep exploration in designing single models that can efficiently handle arbitrary patterns between different components handling different modalities and tasks.
3. *Mixing training and inference*: Gains from simple scaling for pretraining have slowed, requiring more advanced techniques that blur the boundary between training and inference, such as reinforcement learning for LLMs, test-time scaling, and continual learning. We propose exploring these techniques within the lens of a disaggregated foundation model, which unlocks new opportunities for mixing training and inference in diverse ways, e.g., by fine-tuning some components while keeping others frozen.
4. *Closing the loop to the system*: A disaggregated foundation model deployment must be tightly integrated with its particular cluster topology to achieve maximum efficiency. We envision models that are aware of system characteristics such as physical topology and current load, and can reconfigure themselves to improve system efficiency. For example, the model can specify the pieces and configurations needed to determine the optimal resources to provision.

**Programming systems for AI:** Our ability to continue scaling model capabilities is closely tied to our ability to generate, test, and iterate on new ideas. Thus, programmability for high-performance AI systems is a critical component of developing the disaggregated foundation model. Today's programming systems require high development effort from experts to maximize efficiency for new hardware and new models. Reducing this effort can be done along several axes:

1. *Compilers and hardware portability*: As single-GPU performance plateaus, hardware vendors must introduce new accelerators and specialized units within GPUs to continue hardware gains. This poses a significant challenge to programming for AI systems, as single GPUs become more complex and heterogeneous. Meanwhile, the disaggregated foundation model will introduce new operators that must be accelerated on diverse hardware. Investment in compiler infrastructure is necessary to bridge this gap.
2. *System controllability:* The most efficient systems for current foundation models are built monolithically and highly specialized to the transformer architecture. This makes it difficult to modify these systems and reuse their innovations

towards future model architectures. Exploring alternative system designs that allow users more control over system execution is thus critical for realizing the disaggregated foundation model.

3. *Debuggability, testing, and observability:* Debugging and testing is notoriously difficult for ML models, which are stochastic by nature and which require parallel execution for performance. These problems will become even more urgent with the disaggregated foundation model, as the execution path itself may be nondeterministic, along with the final outputs. Future programming systems for AI must consider debuggability and testing as first-class citizens.

**Distributed systems for AI**: Scale-out is a critical part of the disaggregated foundation model. We envision a model whose components can be trained individually and in concert, with dynamic and asynchronous switching between components during training and inference time. For efficiency, the underlying infrastructure must also be flexible, with the ability to scale elastically. This is a significant departure from today's distributed systems for AI, which are typically static, synchronous, tightly coupled to a particular cluster topology, and have limited ability to handle failures. Thus, there are several opportunities for rethinking the scale-out layer:

1. *Distributed training paradigms*: Virtually all current models are trained centrally, which favors just a handful of entities that have the necessary resources and data, and synchronous algorithms, which limits scaling efficiency. A disaggregated foundation model allows for new distributed training paradigms, including federated learning and fully decentralized training. By investing in these paradigms now, we promote openness and resilience for future AI infrastructure. A key challenge lies in scaling asynchronous distributed training algorithms, which can be more efficient than synchronous algorithms but also harder to use and control.
2. *Disaggregated inference:* Similar to training, current inference systems are built on synchronous model execution, executing an entire model's forward pass on one batch of inputs. This leads to under-utilization for heterogeneous models that have different resource requirements for different components within the model. A fully disaggregated inference system would maximize hardware utilization by leveraging dynamic switching between heterogeneous components, placing and scheduling different inputs according to current load.
3. *Codesign for heterogeneous networks:* Future networks for AI systems will be increasingly heterogeneous. Today's systems already comprise multiple options for intra-node accelerator-accelerator and CPU-accelerator links, cross-node switches, and in some cases even cross-datacenter networks. Meanwhile, today's communication primitives for AI systems assume homogeneity and tight synchronization between devices. Mitigating communication bottlenecks at scale

requires more flexible and efficient communication mechanisms specialized to the underlying network.

4. *Flexible and fault-tolerant infrastructure*: Current distributed infrastructure for ML is tightly coupled to an end application, e.g., distributed inference for monolithic transformers, and a particular cluster topology. This is a poor fit for the future AI stack that must rely on disaggregated architectures, both at the model and the hardware level, to further improve capabilities. Meanwhile, the line between training and inference is blurring, with advanced workloads requiring a tight loop between the two. This calls for the need to build a flexible distributed infrastructure that can support both training and inference together. Meanwhile, this infrastructure must also be designed with future hardware in mind, with the ability to scale up and down, handle failures smoothly, and provide optimal routing decisions and communication mechanisms for increasingly heterogeneous networks.

**Hardware systems for AI:** We envision disaggregated foundation models to unlock unprecedented opportunities for hardware-model co-design by breaking the rigid coupling between monolithic model architectures and homogeneous accelerator hardware. Instead of designing a single model to run uniformly across general-purpose GPUs, disaggregated models enable heterogeneous compute environments where different model components—such as attention modules, expert layers, embedding lookups, or memory-augmented modules—can be mapped to distinct hardware units best suited to their computational characteristics. Specifically, we envision that the shift in foundation model architectures will permit a rethinking of AI hardware systems along the following axes:

1. *Specialized Hardware Units:* Each component of a disaggregated model can be assigned to specialized accelerators: for instance, high-throughput tensor cores for dense matrix operations, memory-optimized processors or processing-in-memory (PIM) elements for retrieval or generation-heavy modules, and power-efficient inference cores for lightweight filtering components such as router and gating mechanisms. These accelerators can vary in numerical precision (e.g., FP8 for early-stage processing, FP16/BF16 for dense compute, INT8 for inference) or memory capacity, depending on the stage and sparsity of computation.
2. *Composable/Elastic Hardware Resources*: By treating model components as independent computational units, hardware can be organized into poolable, disaggregated units, where compute resources are allocated just-in-time based on model needs and system constraints. This model allows compute, memory,

and network resources to be independently provisioned and scaled, enabling more efficient reuse of hardware.

3. *Efficient Inference Paths*: Disaggregated models will allow seamless early exits, creating opportunities to design efficient inference paths for simpler inputs. These paths could be mapped to in-memory computing units, or run on edge-optimized cores, conserving energy and bandwidth.
4. *Model-Aware Communication Fabrics*: Current synchronous transformer architectures are bottlenecked by communication at scale. Disaggregated architectures allow hierarchical compute graphs that align with the physical network topology (e.g., intra-node vs inter-node links), enabling topology-aware model partitioning and communication-efficient training/inference across different locality domains, racks, and datacenters.

**Co-Design Across the AI Stack.** Achieving the vision of disaggregated foundation models requires not only innovation within each layer of the AI stack—algorithms, programming systems, distributed infrastructure, and hardware—but also deep co-design across these layers. Algorithmic advances in sparsity, modularity, and dynamic routing must expose meaningful abstractions that programming systems can compile and optimize for disaggregated hardware. Programming environments must in turn support introspection and control hooks that allow distributed systems to dynamically reconfigure execution paths based on workload demands, system load, and network topology. These systems must propagate performance and utilization feedback upward to inform routing and adaptation policies within the model itself. At the hardware level, emerging accelerator heterogeneity and network asymmetries must be surfaced through APIs and runtimes that enable model-aware scheduling and data placement. In this tightly coupled stack, disaggregated models do not merely sit atop infrastructure but they actively reshape it through control mechanisms and workload-aware reconfiguration. Co-design ensures that improvements in one layer do not get bottlenecked by mismatched abstractions or rigid interfaces in another, allowing progress in scalable, efficient, and resilient AI systems.

**Conclusion.** To maintain AI leadership, the United States must not only scale existing AI models but also pioneer the next generation of AI systems that are modular, adaptable, and optimized across the stack. The shift from homogeneous and monolithic AI systems to a new generation of heterogeneous and disaggregated foundation models represents a transformative change in how we design, deploy, and scale AI systems. This opens the door to more efficient, scalable, and specialized computation, aligning model behavior with hardware capabilities and system dynamics. Realizing this vision demands a strategic, coordinated investment in algorithms, programming models, distributed infrastructure, and hardware design, anchored by a commitment to co-design

across these layers. By prioritizing this direction, the Federal government can catalyze breakthroughs that redefine scalability, unlock new applications, and retain strategic advantage in the global AI race.