InfraAI Center at The University of Texas at Austin

Infrastructure for and by Generative AI

Generative AI is changing the structure of computing infrastructure itself. The dominant workloads now span training, post-training, long-context inference, retrieval, multimodal pipelines, and increasingly long-running agentic workflows. At the same time, the underlying platforms are becoming much more heterogeneous: GPUs and other accelerators coexist with CPUs, DPUs and SmartNICs; HBM, DRAM, CXL-attached memory, SSDs, and object stores form deep state hierarchies; and communication can traverse on-package links, scale-up fabrics, datacenter networks, and wide-area paths. Models and workflows evolve far faster than the infrastructure that serves them. A software stack designed around one device, one memory tier, or one fixed execution graph quickly becomes inefficient or obsolete.

The InfraAI Center at UT Austin is organized around a simple premise: future AI infrastructure must embrace this heterogeneity rather than hide it and must incorporate co-design across layers that have traditionally been optimized in isolation. Where computation runs affects communication; where AI state resides affects scheduling; model choice changes compute, memory, and network demand; and agent workflow structure changes which resources are most critical at any instant. InfraAI develops abstractions and runtime mechanisms that let these choices be made together and enables them to evolve as hardware, workloads, and operating conditions change.

InfraAI builds infrastructure for generative AI, and uses generative AI and learning to improve the infrastructure itself.

Pillar 1: Nimble AI Stacks

Our first pillar develops flexible AI stacks that can adapt across heterogeneous compute, memory, storage, and network resources, and across rapidly changing model and workflow structures. The goal is to expose stable semantics to applications while allowing the system to decide late how computation is placed, how state is represented and moved, and how kernels and communication are specialized for the available platform. Recent work spans disaggregated memory for LLM inference, persistent and multi-turn AI state, fabric-aware execution, compute-communication co-design, adaptive inference, RAG serving, and runtimes for dynamic agentic workflows.

Pillar 2: AI-Assisted Infrastructure

Our second pillar asks how infrastructure itself can become adaptive. The configuration space of modern systems is too large and too dynamic for manually engineered rules to cover every workload, topology, accelerator, and failure mode. InfraAI is developing AI-assisted techniques that synthesize and refine resource-management policies, kernels, and control mechanisms from objectives and runtime data. These techniques are paired with richer state representations, constrained interfaces, verification, guardrails, rollback, and recovery so that adaptation can be deployed safely in production systems.

The two pillars reinforce each other. A nimble stack exposes the knobs and semantic boundaries that make adaptation possible; AI-assisted control decides how those knobs should be used for a particular workload and infrastructure instance. This joint agenda is the basis for our recent research directions with companies spanning cloud platforms, enterprise infrastructure, networking, telecom, fabrics, storage, and memory.

Pillar 1: Nimble AI Stacks for Heterogeneous Infrastructure

Generative-AI systems increasingly operate over a continuum of compute and state rather than a single server or accelerator. InfraAI is creating the software abstractions needed to make that continuum programmable and efficient. Three themes organize this work.

Portable AI state across memory and storage

AI state is now a first-class systems object. Model weights, optimizer state, KV and prefix caches, retrieval indexes, checkpoints, agent memory, and intermediate artifacts differ in latency sensitivity, lifetime, mutability, sharing, persistence, and recomputation cost. Yet today each framework often encodes its own assumptions about whether such state belongs in GPU memory, host memory, local SSD, or remote storage. InfraAI is developing interfaces that describe the semantics of AI state independently of its physical placement. The runtime can then decide when to keep state in HBM, move it to DRAM or CXL-attached memory, spill it to NVMe or object storage, compress it, prefetch it, or reconstruct it. Work such as SYMPHONY and our multi-turn inference research shows how workload-level hints can make disaggregated memory practical rather than treating all state movement as a generic cache problem.

Fabric-aware execution and cross-stack co-design

Compute, communication, and memory placement are tightly coupled in modern AI workloads. InfraAI is exploring an OS-like control layer for heterogeneous and composable AI infrastructure, together with a fabric-aware programming model that gives software a common way to name and reason about computation and state across GPUs, CPUs, DPUs/SmartNICs, pooled memory, and network fabrics. Rather than independently choosing a kernel, collective, placement, and data path, the system can optimize them jointly for topology, workload shape, and device capabilities. At the mechanism level, our compute-communication co-design work synthesizes or selects fused implementations whose best form depends on the exact accelerator and interconnect. At the control-plane level, the same principle governs placement, sharing, tiering, migration, routing, and admission across disaggregated resources.

Dynamic generative-AI and agent runtimes

The workload itself is also heterogeneous. A single service may mix conventional inference, long-context requests, RAG, reasoning, multimodal processing, and agent workflows containing tools and external services. These workloads create different bottlenecks and their structure can change during execution. InfraAI is developing runtimes that manage the full execution rather than optimizing one model call at a time. Patchwork provides end-to-end orchestration for RAG pipelines; adaptive attention work changes model execution under load; and Ventis/Nalar exposes dynamic agent and tool calls as runtime-visible futures so scheduling, routing, migration, and state placement can be coordinated across an evolving workflow. New work extends this model across opaque agent harnesses and toward safely revising long-running workflows while old work remains in flight.

A common abstraction: stable intent, adaptive realization

Across these efforts, the application should express what matters: the identity and requirements of AI state, the dependencies among pieces of computation, as well as performance and correctness objectives. The infrastructure should retain freedom over how those requirements are realized on the available platform. Separating intent from realization allows the same stack to evolve with new accelerators, memory technologies, fabrics, and model architectures instead of being rebuilt for each generation.

Pillar 2: AI-Assisted Infrastructure

The same heterogeneity that motivates nimble stacks also makes infrastructure control dramatically harder. A scheduler, memory manager, network controller, or kernel tuner may face thousands of interacting choices whose consequences depend on the workload, hardware generation, topology, and current system state. InfraAI is developing a new control architecture in which AI helps construct and continually improve these mechanisms, while the systems substrate constrains and validates what AI is allowed to do.

AI-assisted policy and mechanism synthesis

Instead of asking an LLM to generate arbitrary systems code, our approach exposes structured decision interfaces around the core choice made by a policy. The system supplies runtime features, state, safety boundaries, and an objective; AI-assisted search or synthesis produces the decision rule encapsulated by that interface. This creates a search space expressive enough to specialize to a workload or platform while excluding broad classes of integration errors by construction. We are applying this idea to resource-management heuristics, scheduling, placement, and lower-level mechanism synthesis. Similar techniques can search over fused compute-communication kernels or other cross-layer implementations where the best mechanism is instance dependent.

State representation and closed-loop adaptation

Adaptive control requires a faithful view of system state. Conventional infrastructure often relies on hand-crafted instantaneous counters or moving averages that expose only a fragment of the conditions driving performance. InfraAI research develops richer system-wide representations, structured runtime memory, fine-grained telemetry, and trace-based reasoning so controllers can distinguish transient symptoms from persistent changes and coordinate decisions across components. The control loop combines online sensing with runtime feedback and replay: observe the current system, choose an action, measure its effect, and refine the policy as the workload or environment evolves.

Safety, verification, and lifecycle management

Infrastructure control has consequences that are very different from ordinary content generation. An adaptive policy can overload a resource, violate an SLO, or leave the system in a state from which recovery is difficult. InfraAI therefore treats verification and guardrails as part of the control architecture. We are developing constrained action spaces, quantitative certificates, runtime checks, provenance, rollback, and recovery mechanisms that allow learning-based policies to operate inside explicit safety boundaries. The long-term goal is infrastructure that can improve continuously while preserving the invariants required by operators and applications.

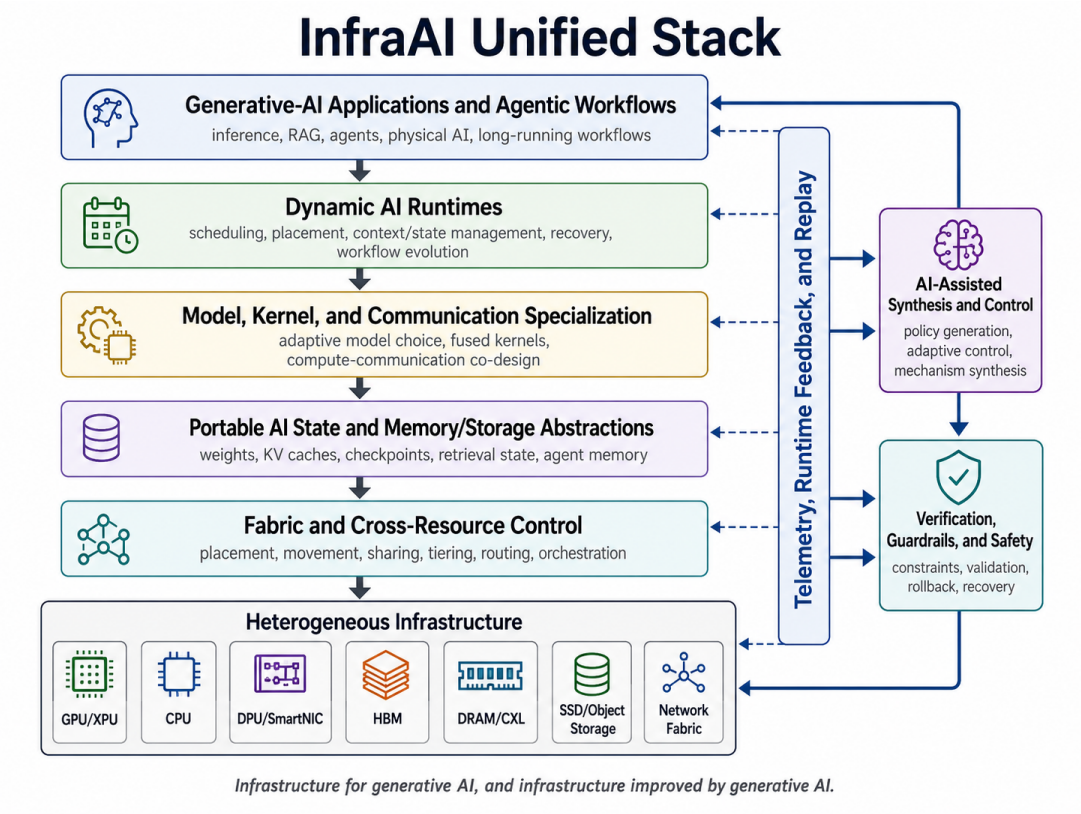
LDOS: the flagship InfraAI project

The Learning Directed Operating System (LDOS) is a flagship realization of this pillar and is supported by a prestigious NSF Expedition in Computing, a $12 million, five-year effort running from 2024 through 2029. LDOS is developing a next-generation operating system that uses machine learning to adapt to dynamic and heterogeneous computing environments. Its research spans richer system state and observation, learned and synthesized policies, cross-component coordination, safe learning, runtime guardrails, and the tooling needed to train, deploy, monitor, and update learned control mechanisms.

Although LDOS begins with the operating system, the abstractions are deliberately broader. The same problems arise in cloud control planes, AI fabrics, edge and telecom systems, storage and memory hierarchies, and autonomous platforms: the environment changes, static policies age, different controllers interact, and adaptation must remain safe. LDOS therefore serves as both a deep systems project and a foundation for the wider InfraAI vision of self-improving infrastructure.

A Unified InfraAI Stack

The figure below summarizes how the two pillars meet. The vertical stack exposes stable abstractions from applications down to heterogeneous hardware. Runtime telemetry and replay feed AI-assisted synthesis and control, while verification and guardrails constrain how those decisions are applied back across the stack. This architecture lets InfraAI attack partner problems at different layers without fragmenting the research agenda: memory and storage projects, fabric and networking projects, cloud orchestration, dynamic inference, and agent runtimes are different views of the same cross-stack system.



Infrastructure for generative AI, and infrastructure improved by generative AI.

Why UT Austin

InfraAI brings together UT Austin strengths in operating systems, networking, distributed systems, architecture, cloud computing, machine-learning systems, generative AI, and formal methods. The center is reinforced by the LDOS Expedition, the Institute for Foundations of Machine Learning, the Center for Generative AI, and the Texas Advanced Computing Center. This combination allows us to work across the full path from new algorithms and models to kernels, runtimes, fabrics, memory systems, and large-scale deployment.

Current InfraAI industry partners

Our work is strengthened by industry partners that support the center and engage with the broader UT Austin systems and AI community.

Meta Google cisco amazon NUTANIX Everpure formerly Pure Storage