



PROFILE



Linkedin
LINKEDIN



AGENTIC AI SYSTEMS

AGENTIC AI SYSTEMS | TECHNICAL PUBLICATION 202610-AASPUB-0005

DF Nexus: Unified Agentic AI Workflows Across Edge, Home Lab, Distributed GPU and Cloud Infrastructure

A Hardware-Adaptive Architecture Integrating CPU, Integrated GPU, Discrete GPU, Small Language Models, Private AI Networking, Multi-GPU Execution, Elastic Compute and Frontier AI Services

Author: Reddy (Eddie) Jamisola Organisation: Agentic AI Systems

Publication Date: 20 October 2026 Technology: DF Nexus

Classification: Applied AI Systems Engineering

Abstract

This technical publication presents the foundational architecture of DF Nexus, a distributed agentic artificial intelligence technology developed by Agentic AI Systems. DF Nexus integrates edge computing, home laboratory infrastructure, distributed AI nodes and cloud computing into a coordinated network of agentic systems.

The architecture recognises that edge devices may use CPU-only processing, CPU with an integrated GPU (iGPU), or CPU with a discrete GPU (dGPU). Smartphones, laptops, mini PCs, desktops and other compatible devices participate according to their actual hardware capabilities. Small language models, including approximately 1-billion-parameter (1B) and 4-billion-parameter (4B) models, can execute in CPU system RAM where suitable. Speech-to-text (STT) and text-to-speech (TTS) enable voice interaction on supported processing hardware.

A single GPU can support multi-stage agentic workflows through compute-mode switching, model switching and multi-inference. Multi-GPU nodes extend capacity and concurrency. WireGuard and Tailscale can connect authorised systems over encrypted private networks; cloud virtual private servers (VPS) can coordinate persistent services. Remote GPU nodes, elastic cloud GPU resources and approved frontier-model APIs extend execution when local resources are insufficient.

The synergy between hardware-adaptive edge AI, home lab infrastructure and cloud computing is the technological foundation of DF Nexus.

1. Introduction

Agentic AI coordinates models, tools, tasks and execution resources toward an objective. It does not intrinsically require a particular GPU, operating system, server or cloud provider. DF Nexus applies this principle across heterogeneous hardware: an edge device can initiate work, a CPU-based model can classify it, a home-lab GPU can execute inference, and a cloud VPS can maintain workflow state. The logical workflow is separated from the physical location of execution.

2. DF Nexus: Foundational Technology

DF Nexus is the distributed agentic computing architecture developed by Agentic AI Systems. Its three principal infrastructure environments are edge AI (smartphones, laptops and CPU/iGPU/dGPU devices), home lab and local AI infrastructure (CPU servers, consumer GPUs, multi-GPU nodes, storage and persistent memory), and cloud infrastructure (VPS orchestration, private connectivity, elastic GPU resources and authorised frontier-model APIs). They cooperate through secure communication, workload orchestration and dynamic resource selection.

3. Hardware-Adaptive Edge Computing

DF Nexus recognises CPU-only, CPU+iGPU and CPU+dGPU as three practical edge-compute configurations. Actual capabilities depend on memory, model formats, drivers and runtime support. Smartphones commonly use a system-on-chip combining CPU and integrated graphics, often with neural accelerators; laptops may have integrated and/or discrete graphics.

Edge configuration	Typical resources	Agentic role
CPU-only	CPU + system RAM	Small-model inference, routing, compatible STT/TTS
CPU + iGPU	CPU + shared-memory graphics	Local inference and supported graphics acceleration
CPU + dGPU	CPU + dedicated GPU/VRAM	Accelerated inference, model switching and multi-inference

4. Smartphones as DF Nexus Edge Nodes

Smartphones may provide speech capture, lightweight local inference, task classification, user interaction and result presentation. Supported mobile CPUs, integrated graphics and neural accelerators can execute appropriately sized models. When a workload exceeds available resources, the phone can delegate an authorised request to another DF Nexus node. Battery, connectivity, OS restrictions, permissions and thermal limits constrain local execution.

5. Laptops and Mini PCs as DF Nexus Edge Nodes

Laptops and mini PCs may operate CPU-only, CPU+iGPU or CPU+dGPU configurations. CPU-only systems can execute quantised small models in RAM. Compatible iGPUs may accelerate inference using shared memory. Discrete GPUs may execute larger or faster models within VRAM limits. Where multiple processing devices exist, DF Nexus can select an appropriate supported execution path.

6. Small Language Models on CPU and System RAM

Approximately 1B and 4B parameter models may support lightweight reasoning, intent classification, extraction, tool selection, workflow routing and structured output. Quantisation can reduce memory requirements. Practical speed and quality depend on processor, RAM, context length, model architecture and runtime. These models can coordinate requests to larger GPU-hosted or API models rather than attempting every task locally.

7. Integrated GPU Acceleration

An iGPU commonly shares system memory with the CPU, though memory architectures vary. Supported inference runtimes can use integrated graphics for suitable AI workloads. Acceleration and memory limits depend on GPU architecture, bandwidth, drivers, model format and framework support; an iGPU is not guaranteed to support every accelerator workload.

8. Discrete GPU Acceleration at the Edge

A discrete GPU typically supplies dedicated graphics compute and VRAM. In DF Nexus, CPU+dGPU edge nodes may host accelerated language models, specialised inference and shared workloads. Model size, context and concurrency are constrained by available VRAM, compute capability and software support.

9. STT and TTS Across CPU, iGPU and dGPU

Speech-to-text converts spoken input to text; text-to-speech converts responses to audio. Depending on model and runtime, either component may operate on CPU and system RAM or benefit from compatible GPU acceleration. An illustrative flow is Microphone → STT → Small Local Agent → DF Nexus Resource Selection → AI Execution → TTS → Speaker. The execution stage may be local or remote.

10. Single-GPU Agentic Workflows

A single physical GPU can execute multiple agentic functions through compute-mode switching (changing the active GPU workload or supported execution configuration), model switching (selecting and replacing models across stages), and multi-inference (handling multiple requests via supported concurrent, batched or interleaved execution). One GPU may perform extraction, checking, reasoning, validation and reporting using separate or shared models. Models can be loaded sequentially or coexist when resources allow. One GPU does not necessarily mean one model, one agent or one workflow.

11. Multi-GPU AI Nodes

Multiple GPUs may jointly serve one model, independently host separate models, or operate in groups. A five-GPU node could allocate all five to a shared-model workload or allocate three to independent models and two jointly to another model, where supported. Dynamic reassignment depends on orchestration, interconnects, model partitioning, runtimes and available memory.

12. Distributed AI Nodes

Independent nodes can specialise in language-model inference, document extraction, speech processing, image generation, validation, memory or orchestration. An authorised edge agent can request work from a remote GPU node without hosting its larger models. Node selection can consider availability, capability, latency and policy.

13. WireGuard and Tailscale Private Networking

WireGuard provides encrypted VPN connectivity; Tailscale builds managed private networking on WireGuard with device identity and access-control capabilities. Either may connect edge devices, home-lab systems, remote GPU nodes and cloud VPS instances across physical locations. Private transport reduces exposure of internal AI endpoints but does not replace application authentication, authorisation or audit logging.

14. VLAN Segmentation and Private AI Networks

Local VLANs provide Layer 2 network segmentation for AI compute, management, storage and other services where networking equipment supports them. WireGuard and Tailscale generally provide routed encrypted Layer 3 overlays, not VLANs. DF Nexus can combine VLAN segmentation with VPN overlays and application access controls, forming a logical private network of authorised agentic systems.

15. Cloud VPS as a DF Nexus Coordination Layer

A CPU/RAM-based VPS can host private connectivity, APIs, identity services, workflow scheduling, task queues, resource discovery, monitoring, persistent memory and audit logs without a dedicated GPU. It may operate as a VPN hub or a peer in a mesh topology. Direct authorised node-to-node execution can avoid forcing every inference request through the VPS.

16. Unified Edge-to-Home-Lab-to-Cloud Architecture

DF Nexus links Edge (CPU / CPU+iGPU / CPU+dGPU) → Private WireGuard or Tailscale Connectivity → DF Nexus Coordination and VPS Services → Home Lab and Distributed CPU/GPU Nodes → Elastic Cloud GPU or Authorised Frontier APIs. This is a logical architecture rather than a mandatory linear data path. The orchestration layer selects authorised resources by capability, cost, latency, data sensitivity and availability.

17. Elastic GPU Compute

When local compute or VRAM is insufficient, DF Nexus may route work to an existing remote GPU node or provision a cloud GPU instance. Elastic capacity may host larger language models, intensive reasoning, media generation or peak concurrency. Provisioning, secure connectivity, billing controls and resource release depend on the chosen service. Elastic compute is a selectable execution tier, not a permanent requirement for edge participation.

18. Frontier AI Model Integration

Approved provider APIs can complement self-hosted models. The architecture can select OpenAI GPT models via API, Anthropic Claude, DeepSeek models and Qwen-based deployments where offered and authorised. A consumer ChatGPT subscription does not itself confer OpenAI API access. Public model APIs do not automatically join the private VPN and require separate authentication, cost governance and data-handling controls.

Provider / ecosystem	Example models	Potential role
OpenAI	GPT via API	Reasoning, language, tool-enabled tasks
Anthropic	Claude	Analysis, reasoning, drafting
DeepSeek	DeepSeek family	Language and reasoning
Alibaba Qwen	Qwen family	Multilingual and structured tasks

19. Example DF Nexus Agentic Workflow

A smartphone captures speech; a compatible local STT component transcribes it; a 1B or 4B local model interprets the request. DF Nexus selects CPU, iGPU or local dGPU execution for suitable tasks. A demanding stage can be sent over private networking to a home-lab GPU node using model switching or multi-inference. If necessary and permitted, an elastic cloud GPU or frontier-model API supplies further capability. Results return for validation, persistence and optional local TTS. Each external transfer is governed by authorisation and data policy.

20. Dynamic Model and Compute Selection

Resource selection can consider model capability, CPU performance, system RAM, iGPU support, dGPU availability, VRAM, expected latency, network conditions, energy, price and security. Local small models can handle routine routing; a single GPU can run multi-stage tasks; multi-GPU systems can increase capacity; elastic resources and APIs provide selective escalation. Logical workflow state remains independent of execution hardware.

21. Persistent Memory and Workflow Continuity

Persistent memory can retain authorised organisational knowledge, evidence, task state, validation outcomes and reusable results across model sessions and nodes. This complements DF Nexus private networking: the network transports authorised requests, while memory services preserve knowledge and workflow continuity. Implementations should enforce permissions, retention and audit requirements.

22. Security and Governance

DF Nexus deployments require authenticated APIs, encrypted transport, least-privilege access, credential protection, logging and restrictions on external actions and data movement. A VPN is not a substitute for application-layer controls. Sensitive information must not be sent to cloud GPU providers or frontier-model APIs without authorisation. Consequential operations may require human approval.

23. Reliability and Operational Resilience

Node outages, network interruptions, model failures and exhausted compute capacity can be handled with appropriate retries, alternate resources, checkpointing, controlled reassignment and human escalation. Durable workflow state can reduce repeated work and support resumption across execution locations. Idempotency and safeguards are important when tasks cause external side effects.

24. Cost and Energy Efficiency

CPU-only and iGPU devices may handle lightweight work, discrete GPUs may execute intensive local inference, and multi-GPU systems can provide model capacity and concurrency. Elastic GPUs and frontier APIs can be reserved for workloads that justify additional cost. Actual efficiency depends on model quality, utilisation, energy consumption, network overhead, workload duration and provider pricing.

25. Relationship to Previous Agentic AI Systems Publications

DF Nexus unifies earlier engineering themes: Publication 0001 local AI and Microsoft 365 integration; Publication 0002 DF Memory and persistent knowledge; Publication 0003 dynamic GPU compute-mode switching, model switching and multi-inference; and Publication 0004 consumer GPU power-curve and thermal management. Publication 0005 establishes their architectural relationship across edge, home-lab, distributed and cloud environments.

26. Conclusion

DF Nexus is the foundational distributed agentic computing architecture of Agentic AI Systems. Its defining principle is the coordinated synergy of hardware-adaptive edge devices, home-lab AI infrastructure and cloud services. Smartphones, laptops and other nodes can participate through CPU-only, CPU+iGPU or CPU+dGPU configurations; small 1B and 4B models and compatible STT/TTS provide lightweight local capabilities. A single GPU can support agentic workflows through mode switching, model switching and multi-inference, while multi-GPU nodes and distributed infrastructure extend capacity.

WireGuard and Tailscale can provide encrypted private connectivity, VLANs can add local segmentation, and a VPS can maintain orchestration and persistent services. Elastic GPU compute and authorised frontier-model APIs extend execution when necessary. DF Nexus is not dependent on one model, GPU, server or cloud provider: it is a coordinated network of agentic intelligence spanning edge, home lab and cloud.

© 2026 Agentic AI Systems — Intelligence. Automation. Advantage.

Document ID: 202610-AASPUB-0005 | Author: Reddy (Eddie) Jamisola

This publication describes architectural principles and applicable engineering configurations. Performance, compatibility, security and scalability depend on deployed hardware, software, models, networking and operating conditions. Not every configuration is represented as independently benchmarked in a single deployment.