



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

DYNAMIC GPU COMPUTE MODE SWITCHING, MODEL SWITCHING AND MULTI-INFERENCE

Proven Single-GPU and Multi-GPU AI Execution, Concurrent Multi-Model Inference and Dynamic Compute Resource Allocation

AGENTIC AI SYSTEMS — TECHNICAL PUBLICATION

Author: Reddy (Eddie) Jamisola

Organisation: Agentic AI Systems

Document Number: 202610-AASPUB-0003

Publication Date: 20 October 2026

Location: Darwin, Northern Territory, Australia

Abstract

Agentic AI Systems has developed and demonstrated a dynamic AI compute architecture supporting GPU compute mode switching, AI model switching, multi-model execution and concurrent multi-inference. The architecture separates three dimensions: GPU allocation, number of AI models and number of simultaneous inference workloads. Demonstrated configurations encompass single-GPU single-model inference, multiple inferences against one model on one GPU, multiple models on one GPU, concurrent inferences across multiple models, collective multi-GPU inference and dynamic reconfiguration between collective and independent workloads. Proprietary implementation details are withheld.

1. Introduction

Fixed one-GPU/one-model deployments can limit flexibility as enterprise AI requirements change. Agentic AI Systems has developed and demonstrated adaptable arrangements that allow GPU resources, model assignments and concurrent inference workloads to vary. The engineering contribution is the integration of these capabilities in a flexible compute environment rather than the invention of GPU inference itself.

2. Three Independent Dimensions of AI Compute

GPU allocation describes how many GPU resources participate in a workload. Model allocation describes how many AI models operate within the assigned resources. Inference concurrency describes how many requests or processing tasks operate at once. One model does not necessarily imply one inference request, and one GPU does not necessarily imply one model.

3. Single-GPU, Single-Model, Single-Inference

1 GPU → 1 AI model → 1 inference. This dedicated arrangement establishes a baseline for controlled local inference and comparison with more complex configurations.

4. Single-GPU, Single-Model, Multi-Inference

1 GPU → 1 AI model → multiple concurrent inferences. A shared model can serve multiple active requests or AI agents using the same GPU resource. This supports multi-user services and concurrent business processing without dedicating a separate GPU to every request. Internal concurrency methods are proprietary.

5. Single-GPU, Multi-Model Operation

1 GPU → multiple AI models. A single GPU can support distinct model workloads, subject to available resources. Models may be active concurrently or managed through controlled transitions, depending on the configuration. This separates the model count from the GPU count.

6. Single-GPU, Multi-Model, Multi-Inference

1 GPU → multiple AI models → multiple concurrent inferences. Multiple models can support separate AI tasks while handling multiple active inference requests on the same GPU. This is distinct from simply storing or loading several models: it concerns model execution and inference concurrency.

7. Multi-GPU, Single-Model Inference

Multiple GPUs → 1 AI model. Participating GPU resources collectively support a single model workload. This is useful where one GPU alone is insufficient for the desired model configuration. Internal workload distribution and memory coordination are not disclosed.

8. Multi-GPU, Multi-Model and Multi-Inference

Multiple GPUs can support separate AI models, multiple requests to individual models, shared model execution across GPU groups and combinations of collective and independent inference workloads. The allocation can be adapted to changing operational needs.

9. Demonstrated GPU and Inference Configurations

Mode	GPU resources	Models / inference
A — Dedicated	1 GPU	1 model, 1 inference
B — Shared model	1 GPU	1 model, multiple inferences
C — Multi-model	1 GPU	Multiple models
D — Concurrent multi-model	1 GPU	Multiple models, multiple inferences
E — Collective	Multiple GPUs	1 model
F — Independent	Multiple GPUs	Multiple independent models
G — Distributed	Multiple GPUs	Multiple models and inferences
H — Consolidated	5 GPUs	1 shared model
I — Partitioned	3 GPUs + 2 GPUs	3 independent models + 1 shared model

10. Dynamic GPU Compute Mode Switching

Dynamic compute mode switching permits GPU resources to be reassigned between operating configurations. One example involves five GPUs collectively serving one model, followed by a partitioned configuration with three GPUs supporting independent models and two GPUs collectively serving another model. The proprietary switching and scheduling mechanisms are not disclosed.

11. Dynamic AI Model Switching

Model switching changes which AI models execute on available resources. It can operate independently or in combination with GPU reassignment, supporting task-specific inference and adaptable resource use.

12. Combined Switching and Multi-Inference

The architecture supports changes in GPU count per workload, models per GPU, concurrent inference requests, and division of GPUs between collective and independent model workloads. The defining capability is coordinated flexibility across these three dimensions.

13. Engineering Demonstration and Proven Development

Agentic AI Systems has conducted practical local inference and multi-GPU engineering work involving model execution, multi-model operation and compute configuration changes. The publication presents these as demonstrated engineering capabilities. Detailed test records, performance measurements, concurrency limits and switching times are intentionally not published; no numerical performance improvements are claimed.

14. Enterprise and Commercial Applications

Applications include enterprise automation, AI agent infrastructure, private local AI processing, flexible GPU utilisation, multi-user AI services, specialist models for distinct workflows and distributed AI operations. Multiple inference workloads can be supported without a permanently fixed one-GPU/one-model allocation.

15. Engineering Significance

GPU acceleration, concurrent inference serving and multi-GPU processing are established concepts. The engineering significance here is their integration into a flexible architecture supporting single-GPU multi-model execution, single-model multi-inference, multi-model multi-inference, collective multi-GPU execution and dynamic reassignment.

16. Intellectual Property Protection

Agentic AI Systems retains proprietary intellectual property relating to compute mode switching, model switching, concurrency coordination, resource allocation, memory management, orchestration, optimisation and internal integration. This public publication does not disclose source code, algorithms, network topology, configuration parameters or detailed test procedures.

17. Conclusion

Agentic AI Systems has developed and demonstrated an adaptable inference architecture in which GPU resources, AI model assignments and concurrent inference workloads can vary. One GPU can support multiple models, one model can serve multiple inference requests, and multiple GPUs can be consolidated or divided between independent workloads. The central engineering achievement is the coordinated flexibility of GPU allocation, model allocation and inference concurrency.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

DYNAMIC GPU COMPUTE MODE SWITCHING, MODEL SWITCHING AND MULTI-INFERENCE

Proven Single-GPU and Multi-GPU AI Execution, Concurrent Multi-Model Inference and Dynamic Compute Resource Allocation

AGENTIC AI SYSTEMS — TECHNICAL PUBLICATION

Author: Reddy (Eddie) Jamisola

Organisation: Agentic AI Systems

Document Number: 202610-AASPUB-0003

Publication Date: 20 October 2026

Location: Darwin, Northern Territory, Australia

Abstract

Agentic AI Systems has developed and demonstrated a dynamic AI compute architecture supporting GPU compute mode switching, AI model switching, multi-model execution and concurrent multi-inference. The architecture separates three dimensions: GPU allocation, number of AI models and number of simultaneous inference workloads. Demonstrated configurations encompass single-GPU single-model inference, multiple inferences against one model on one GPU, multiple models on one GPU, concurrent inferences across multiple models, collective multi-GPU inference and dynamic reconfiguration between collective and independent workloads. Proprietary implementation details are withheld.

1. Introduction

Fixed one-GPU/one-model deployments can limit flexibility as enterprise AI requirements change. Agentic AI Systems has developed and demonstrated adaptable arrangements that allow GPU resources, model assignments and concurrent inference workloads to vary. The engineering contribution is the integration of these capabilities in a flexible compute environment rather than the invention of GPU inference itself.

2. Three Independent Dimensions of AI Compute

GPU allocation describes how many GPU resources participate in a workload. Model allocation describes how many AI models operate within the assigned resources. Inference concurrency describes how many requests or processing tasks operate at once. One model does not necessarily imply one inference request, and one GPU does not necessarily imply one model.

3. Single-GPU, Single-Model, Single-Inference

1 GPU → 1 AI model → 1 inference. This dedicated arrangement establishes a baseline for controlled local inference and comparison with more complex configurations.

4. Single-GPU, Single-Model, Multi-Inference

1 GPU → 1 AI model → multiple concurrent inferences. A shared model can serve multiple active requests or AI agents using the same GPU resource. This supports multi-user services and concurrent business processing without dedicating a separate GPU to every request. Internal concurrency methods are proprietary.

5. Single-GPU, Multi-Model Operation

1 GPU → multiple AI models. A single GPU can support distinct model workloads, subject to available resources. Models may be active concurrently or managed through controlled transitions, depending on the configuration. This separates the model count from the GPU count.

6. Single-GPU, Multi-Model, Multi-Inference

1 GPU → multiple AI models → multiple concurrent inferences. Multiple models can support separate AI tasks while handling multiple active inference requests on the same GPU. This is distinct from simply storing or loading several models: it concerns model execution and inference concurrency.

7. Multi-GPU, Single-Model Inference

Multiple GPUs → 1 AI model. Participating GPU resources collectively support a single model workload. This is useful where one GPU alone is insufficient for the desired model configuration. Internal workload distribution and memory coordination are not disclosed.

8. Multi-GPU, Multi-Model and Multi-Inference

Multiple GPUs can support separate AI models, multiple requests to individual models, shared model execution across GPU groups and combinations of collective and independent inference workloads. The allocation can be adapted to changing operational needs.

9. Demonstrated GPU and Inference Configurations

Mode	GPU resources	Models / inference
A — Dedicated	1 GPU	1 model, 1 inference
B — Shared model	1 GPU	1 model, multiple inferences
C — Multi-model	1 GPU	Multiple models
D — Concurrent multi-model	1 GPU	Multiple models, multiple inferences
E — Collective	Multiple GPUs	1 model
F — Independent	Multiple GPUs	Multiple independent models
G — Distributed	Multiple GPUs	Multiple models and inferences
H — Consolidated	5 GPUs	1 shared model
I — Partitioned	3 GPUs + 2 GPUs	3 independent models + 1 shared model

10. Dynamic GPU Compute Mode Switching

Dynamic compute mode switching permits GPU resources to be reassigned between operating configurations. One example involves five GPUs collectively serving one model, followed by a partitioned configuration with three GPUs supporting independent models and two GPUs collectively serving another model. The proprietary switching and scheduling mechanisms are not disclosed.

11. Dynamic AI Model Switching

Model switching changes which AI models execute on available resources. It can operate independently or in combination with GPU reassignment, supporting task-specific inference and adaptable resource use.

12. Combined Switching and Multi-Inference

The architecture supports changes in GPU count per workload, models per GPU, concurrent inference requests, and division of GPUs between collective and independent model workloads. The defining capability is coordinated flexibility across these three dimensions.

13. Engineering Demonstration and Proven Development

Agentic AI Systems has conducted practical local inference and multi-GPU engineering work involving model execution, multi-model operation and compute configuration changes. The publication presents these as demonstrated engineering capabilities. Detailed test records, performance measurements, concurrency limits and switching times are intentionally not published; no numerical performance improvements are claimed.

14. Enterprise and Commercial Applications

Applications include enterprise automation, AI agent infrastructure, private local AI processing, flexible GPU utilisation, multi-user AI services, specialist models for distinct workflows and distributed AI operations. Multiple inference workloads can be supported without a permanently fixed one-GPU/one-model allocation.

15. Engineering Significance

GPU acceleration, concurrent inference serving and multi-GPU processing are established concepts. The engineering significance here is their integration into a flexible architecture supporting single-GPU multi-model execution, single-model multi-inference, multi-model multi-inference, collective multi-GPU execution and dynamic reassignment.

16. Intellectual Property Protection

Agentic AI Systems retains proprietary intellectual property relating to compute mode switching, model switching, concurrency coordination, resource allocation, memory management, orchestration, optimisation and internal integration. This public publication does not disclose source code, algorithms, network topology, configuration parameters or detailed test procedures.

17. Conclusion

Agentic AI Systems has developed and demonstrated an adaptable inference architecture in which GPU resources, AI model assignments and concurrent inference workloads can vary. One GPU can support multiple models, one model can serve multiple inference requests, and multiple GPUs can be consolidated or divided between independent workloads. The central engineering achievement is the coordinated flexibility of GPU allocation, model allocation and inference concurrency.

AGENTIC AI SYSTEMS — Intelligence. Automation. Advantage.

© 2026 Agentic AI Systems. All rights reserved.