A Novel Compound AI Model for 6G Networks in 3D Continuum

Milos Gravara
Distributed Systems Group
TU Wien
m.gravara@dsg.tuwien.ac.at

Andrija Stanisic
Distributed Systems Group
TU Wien
a.stanisic@dsg.tuwien.ac.at

Stefan Nastic
Distributed Systems Group
TU Wien
s.nastic@dsg.tuwien.ac.at

***Abstract*—The 3D continuum presents a complex environment that spans the terrestrial, aerial and space domains, with 6G networks serving as a key enabling technology. Current AI approaches for network management rely on monolithic models that fail to capture cross-domain interactions, lack adaptability, and demand prohibitive computational resources. This paper presents a formal model of Compound AI systems, introducing a novel tripartite framework that decomposes complex tasks into specialized, interoperable modules. The proposed modular architecture provides essential capabilities to address the unique challenges of 6G networks in the 3D continuum, where heterogeneous components require coordinated, yet distributed, intelligence. This approach introduces a fundamental trade-off between model and system performance, which must be carefully addressed. Furthermore, we identify key challenges faced by Compound AI systems within 6G networks operating in the 3D continuum, including cross-domain resource orchestration, adaptation to dynamic topologies, and the maintenance of consistent AI service quality across heterogeneous environments.**

***Index Terms*—3D Continuum, Compound AI, 6G Networks**

I. Introduction

The integrated 3D continuum represents a fundamental shift from traditional ground-based infrastructure, seamlessly connecting terrestrial, aerial, and space-based components into a unified ecosystem. 6G networks serve as one of the key enabling technologies for this continuum, facilitating interactions between ground networks, high-altitude platforms, unmanned aerial vehicles, and satellite constellations [1]. This 3D continuum extends beyond mere connectivity to encompass distributed compute and storage resources across heterogeneous domains, creating a fabric that supports ubiquitous digital services regardless of geographical location or altitude. Through 6G's orchestration capabilities, the system promises unprecedented global coverage, enhanced reliability, and improved throughput across all segments of this multidimensional environment.

However, managing this 3D environment presents significant challenges due to the high mobility, heterogeneity, and dynamic nature of its components, particularly satellite nodes that continuously reshape network topologies in real time [2]. Although current AI-assisted approaches to network management predominantly rely on monolithic models, these face critical limitations within the 3D continuum context. Single-model strategies typically specialize in optimizing specific domains without effectively capturing cross-domain interactions, lack the adaptability required for rapidly changing network conditions, and often require substantial computational resources unavailable across all segments of the network. Furthermore, training and operating such large monolithic models incurs prohibitive costs in terms of computation, energy, and maintenance. System performance metrics such as latency and throughput also suffer significantly when these models attempt to handle large-scale, heterogeneous network environments, creating bottlenecks that undermine real-time decision-making capabilities.

Compound AI systems [4], [5], [6] offer distinct advantages over these current state-of-the-art approaches. By distributing intelligence across the network, Compound AI systems enable localized decision making while maintaining global coordination, efficiently utilize heterogeneous computational resources, and adapt to the unique characteristics of each domain. This paper presents Compound AI systems to address the key management challenges of the 6G networks in the 3D continuum, demonstrating how this modular yet integrated approach can enable autonomous adaptation while maintaining consistent service quality and operational efficiency across terrestrial, aerial, and space domains.

II. Overview of Compound AI Systems

A. Towards system design approach to AI

Rather than relying on a single model to handle all aspects of a complex problem, Compound AI systems decompose tasks into manageable sub-components. This modular approach enables developers to leverage specialized models or tools for different sub-tasks and control information flow. For example, lightweight models can make real-time decisions at terrestrial base stations, while different specialized models simultaneously optimize aerial and satellite resource allocation. Supporting infrastructure components such as vector databases for similarity search, orchestration frameworks for module coordination, API gateways for external tool access, and monitoring systems for quality assurance create a rich ecosystem of components to build upon. This diversity allows the system to function effectively even when parts have varying computational capacities or connectivity constraints. Similarly, routing inputs to different-sized models based on task complexity and overall system load can optimize computational resources

while maintaining quality of service across heterogeneous network segments. As network conditions evolve, individual AI components can be updated independently, preserving adaptability without requiring retraining of the entire system.

B. Example Compound AI System and Main Design Principles

To better understand the concept of Compound AI systems, we turn to the following explanatory example.

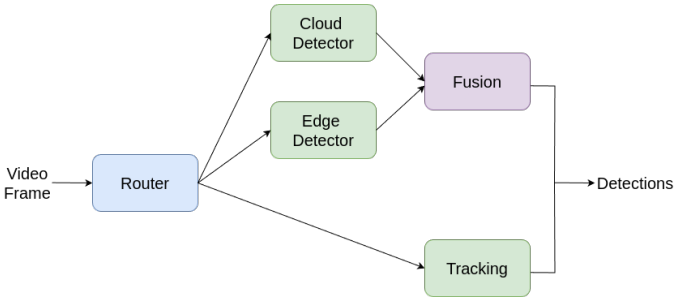

Fig. 1. VATE: A Compound AI System for Edge-Cloud Object Detection and Tracking. [7]

Fig 1 illustrates a Compound AI system for edge-cloud collaborative object detection and tracking. It demonstrates how multiple specialized modules collaborate in a unified system. At the edge, a lightweight detector identifies potential objects, while a tracking module maintains object persistence across frames. On the cloud side, a more powerful detector offers greater accuracy for challenging cases. A fusion module combines detections from both edge and cloud sources to create an improved understanding of the scene. Furthermore, a dedicated orchestrator module makes intelligent decisions about when to process input locally, when to offload to the cloud, and when to rely on tracking rather than detection, balancing between computational efficiency and accuracy.

With these considerations, we define Compound AI systems as systems composed of specialized, interoperable modules that collectively address complex AI tasks. Each module performs a distinct function and interacts via well-defined defined interfaces. The following key characteristics define the structural and functional principles that underpin Compound AI systems, enabling them to address complex tasks in a scalable and efficient manner:

- **Modularity** – Separate parts of Compound AI systems can be developed, tested and maintained independently while minimizing impacts on the overall system. This represents a separation of concerns that facilitates parallel development and iterative improvement by specialized teams [8].
- **Adaptability** – The modular design of Compound AI systems enables rapid adaptation to new requirements or changing conditions by allowing individual components to be replaced, enhanced, or reconfigured without rebuilding the entire system.
- **Abstraction** – Internal complexities of modules are hidden behind well-defined interfaces, ensuring that changes to a module's implementation don't affect other components. This creates a clear separation between what a module does and how it accomplishes its task.
- **Interaction-defined Architecture** – The architecture of a Compound AI system is fundamentally defined by how modules interact with each other. These interactions establish the data flow through the system and determine how information is processed, transformed and utilized across modules.
- **Cost-Effectiveness** - Designing with cost-effectiveness in mind ensures that resource utilization, energy consumption, and computational requirements are optimized. This principle is crucial not only for reducing operational expenses but also for enabling an architecture where individual components can be updated or replaced without necessitating a complete retraining of the entire system

III. Compound AI System Model

A. Definition and Core Components

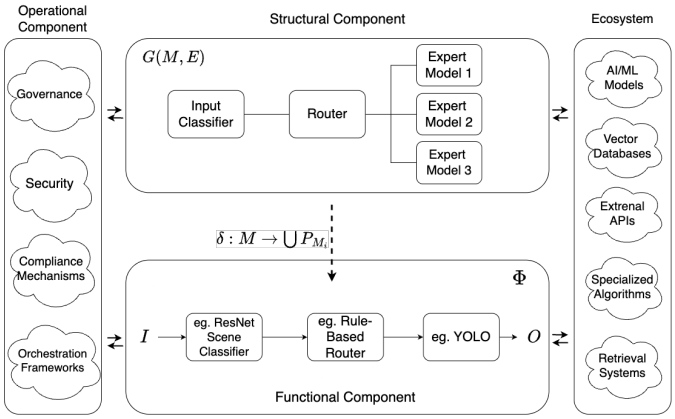

Fig. 2. Tripartite model for Compound AI systems

In Fig 2 we present a model and reference architecture for Compound AI system. This reference architecture offers a conceptual framework through which to analyze, conceptualize, and implement Compound AI (CAI) systems as multimodular, interactive, and adaptive entities. These systems are composed from a comprehensive ecosystem encompassing various AI/ML models, vector databases, retrieval mechanisms, external APIs, and domain-specific algorithmic solutions.

We define a Compound AI System formally as a triple:

$$\text{CAI} = (S, F, O)$$

Where:

- S represents the **Structural** component
- F represents the **Functional** component
- O represents the **Operational** component

While each component serves a distinct role, they interact in important ways:

1) The Structural component (S) defines what modules exist and how they connect, constraining the possible implementations in the Functional component (F).
2) The Functional component (F) realizes the abstract architecture defined in S through concrete implementations, potentially informing structural changes based on implementation constraints.
3) The Operational component (O) monitors and maintains S and F, providing feedback for optimization and adaptation.

The tripartite model naturally supports key system design principles such as modularity, adaptability, and abstraction. By clearly separating architectural, behavioral, and operational concerns, it enables more systematic approaches to the development, deployment, and maintenance of complex, multi-model AI systems.

B. Structural Component (S)

The Structural component defines the high-level architecture of the system, specifying what each module does and how information flows between modules. Formally, we represent this as a directed graph:

$$S = G(M, E)$$

Where:

- $M = \{M_1, M_2, \ldots, M_n\}$ is the set of modules
- $E \subseteq M \times M$ is the set of connections between modules

For any two modules M_i and M_j, an edge $E(M_i, M_j)$ indicates that the output of module M_i serves as an input to module M_j.

Each module M_i is characterized by its input space I_i and output space O_i, which define the types and formats of data the module can accept and produce. The Structural component thus establishes a "schema" for the Compound AI system, defining the roles and relationships of its constituent parts without specifying their implementation details.

C. Functional Component (F)

The Functional component defines how the system behaves by mapping the abstract architecture to concrete implementations. It encompasses:

$$F = (P, \Phi, \delta)$$

Where:

- $P = \{P_1, P_2, \ldots, P_n\}$ is a implementation set where each P_i is an implementation pool for module M_i
- Φ is the composition function that determines the overall system behavior
- δ is the mapping function that selects specific implementations from pools

For each module M_i, the implementation pool $P_i = \{f_{i1}, f_{i2}, \ldots, f_{ik}\}$ contains multiple possible implementations that fulfill the same functional role but may differ in their performance characteristics, resource requirements, or other properties.

The mapping function $\delta : M \rightarrow \bigcup P_i$ selects a specific implementation for each module, such that $\delta(M_i) \in P_i$.

The composition function Φ integrates these implementations according to the structural blueprint to produce the overall system behavior:

$$\Phi : I \rightarrow O$$

Where I is the input space and O is the output space of the entire system.

Critically, this function ensures that the compound system's behavior matches what would be expected from a monolithic model:

$$y = \Phi(x) \quad \text{where } x \in I,\ y \in O$$

The composition function can be formally defined in terms of the graph execution:

$$\Phi(x) = \Psi(G(M, E), \delta(M_1), \delta(M_2), \ldots, \delta(M_n), x)$$

Where Ψ is a graph execution function that propagates inputs through the implementation graph according to the connection pattern defined in S.

D. Operational Component (O)

The Operational component encompasses the infrastructure and processes that enable, maintain, and optimize the running system. This includes:

$$O = (Mon, Sec, Gov, Orch)$$

Where:

- Mon represents monitoring and observability systems
- Sec represents security and compliance mechanisms
- Gov represents governance frameworks and policies
- $Orch$ represents orchestration and resource management

The Operational component serves as the foundation that supports both the Structural and Functional components, providing:

1) **System monitoring** that tracks performance, resource utilization, and failure modes
2) **Security controls** that protect system integrity and data privacy
3) **Governance mechanisms** that ensure compliance with regulations and ethical standards
4) **Orchestration tools** that manage deployment, scaling, and resource allocation

This component parallels DevOps and MLOps practices in software engineering but extends them to address the unique challenges of compound AI systems.

IV. Open Challenges for Compound AI in 6G Networks for 3D Continuum

To realize our vision of Compound AI, the following challenges need to be addressed.

A. Cross-Domain Resource Orchestration

Orchestrating Compound AI resources across the 3D continuum faces unique constraints that existing AI solutions fail to address. Terrestrial components can leverage high computational capacity but are limited in coverage, aerial platform-based components faces energy and computational constraints, while satellite-hosted systems can provide wide coverage but with significant computational and latency limitations. Current AI orchestration approaches treat these domains separately, creating inefficiencies at domain boundaries. Compound AI for 6G networks requires intelligent decomposition and coordination mechanisms that can distribute AI tasks optimally across these diverse domains while accounting for their unique characteristics and computational limitations.

B. Adaptation to Dynamic Network Topologies

The constantly evolving network topologies of the 3D continuum challenge traditional deployment strategies. As satellite constellations orbit, aerial platforms move, and terrestrial demand shifts, Compound AI systems must continuously reconfigure themselves to maintain performance. Current AI composition algorithms struggle with this dynamism, leading to suboptimal configurations where component distribution becomes misaligned with actual network conditions. Compound AI for 6G networks requires adaptive systems capable of predicting topology changes and proactively reconfiguring component distribution and communication patterns, thus maintaining overall model performance while optimizing system performance metrics.

C. Maintaining AI Service Consistency

Delivering consistent AI service quality across the 3D continuum presents significant technical challenges. As AI requests and data transition between terrestrial, aerial, and space segments, maintaining continuity of AI inference quality becomes increasingly difficult. Current approaches typically react to AI service degradations after they occur, particularly at domain boundaries where computational resources vary dramatically. Compound AI for 6G networks requires predictive capabilities that can anticipate performance variations across the continuum and implement proactive measures in order to maintain service level objectives despite the inherent heterogeneity and dynamic nature of the underlying 3D infrastructure.

D. Balancing the Trade-offs

Compound AI systems present a spectrum of performance trade-offs that must be carefully navigated [3]. In scenarios demanding enhanced reasoning capabilities, introducing additional specialized components to a large monolithic model can improve model performance (accuracy, precision, recall, etc.) but may degrade system performance through increased latency, higher energy consumption, and greater computational requirements. In contrast, when deploying AI on the edge, Compound AI approaches aim to maintain model performance comparable to monolithic solutions while significantly improving system performance by reducing latency, energy consumption, and computational demands to accommodate resource-constrained environments. These inherent trade-offs create significant research challenges for Compound AI systems in the 3D continuum, where components range from powerful data centers to resource-limited edge devices. Future research must develop adaptive frameworks that can dynamically reconfigure Compound AI systems based on changing network conditions, available resources, and application requirements. Successfully addressing these challenges will enable Compound AI systems to flexibly balance model and system performance according to deployment contexts, ensuring that the benefits outweigh the complexities they introduce.

V. Conclusion

In this paper, we introduced the concept of Compound AI systems and presented a formal tripartite model that captures their structural, functional, and operational dimensions. By breaking down complex tasks into specialized, interoperable modules, Compound AI systems offer a scalable and adaptable alternative to traditional monolithic AI architectures. We explored how this general framework can address the unique challenges of 6G networks in the 3D continuum, highlighting the advantages of modularity and distributed intelligence in such complex scenarios.

Looking ahead, our future work will focus on applying this general Compound AI system model to real-world 6G use cases and other dynamic, distributed environments. This includes implementing adaptive orchestration strategies, developing proactive reconfiguration mechanisms, and validating system performance across varying network and resource conditions. By doing so, we aim to demonstrate how Compound AI systems can be effectively deployed to meet the demands of next-generation networks and beyond.

References

[1] Massod Khorsandi Bahare, Anastasius Gavras, Marco Gramaglia, John Cosmas, Xi Li, Ömer Bulakci, Arifur Rahman, Alexandros Kostopoulos, Agapi Mesodiakaki, Dimitris Tsolkas, Mårten Ericson, Mauro Boldi, Mikko Uusitalo, Mir Ghoraishi, and Patrik Rugeland. The 6g architecture landscape - european perspective, 2023. Working paper, Version v1.

[2] D. Bhattacherjee, W. Aqeel, I. N. Bozkurt, A. Aguirre, B. Chandrasekaran, P. B. Godfrey, G. Laughlin, B. Maggs, and A. Singla. Gearing up for the 21st century space race, 2018.

[3] Gohar Irfan Chaudhry, Esha Choukse, Íñigo Goiri, Rodrigo Fonseca, Adam Belay, and Ricardo Bianchini. Towards resource-efficient compound ai systems, 2025.

[4] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Matei Zaharia, James Zou, and Ion Stoica. Optimizing model selection for compound ai systems, 2025.

[5] Jared Quincy Davis, Boris Hanin, Lingjiao Chen, Peter Bailis, Ion Stoica, and Matei Zaharia. Networks of networks: Complexity class principles applied to compound ai systems design, 2024.

[6] Swayambhoo Jain, Ravi Raju, Bo Li, Zoltan Csaki, Jonathan Li, Kaizhao Liang, Guoyao Feng, Urmish Thakkar, Anand Sampat, Raghu Prabhakar, and Sumati Jairath. Composition of experts: A modular compound ai system leveraging large language models, 2024.

[7] Maximilian Maresch and Stefan Nastic. Vate: Edge-cloud system for object detection in real-time video streams. In *2024 IEEE 8th International Conference on Fog and Edge Computing (ICFEC)*, pages 27–34, 2024.

[8] Helena Zhang, Jakobi Haskell, and Yosef Frost. Flow state: Humans enabling ai systems to program themselves, 2025.