Hypergraphs: Facilitating High-Order Modeling of the Computing Continuum

Dragi Kimovski
dragi.kimovski@aau.at
University of Klagenfurt
Klagenfurt, Austria

ABSTRACT

As contemporary computing infrastructures evolve to include diverse architectures beyond traditional von Neumann models, the limitations of classical graph-based infrastructure and application modelling become apparent, particularly in the context of the computing continuum and its interactions with Internet of Things (IoT) applications.

Hypergraphs prove instrumental in overcoming this obstacle by enabling the representation of computing resources and data sources irrespective of scale. This allows the identification of new relationships and hidden properties, supporting the creation of a federated, sustainable, cognitive computing continuum with shared intelligence.

The paper introduces the HyperContinuum conceptual platform, which provides resource and applications management algorithms for distributed applications in conjunction with next-generation computing continuum infrastructures based on novel von Neumann computer architectures. The HyperContinuum platform outlines high-order hypergraph applications representation, sustainability optimization for von Neumann architectures, automated cognition through federated learning for IoT application execution, and adaptive computing continuum resources provisioning.

KEYWORDS

Hypergraphs, Computing Continuum, Optimisation

ACM Reference Format:

Dragi Kimovski. 2024. Hypergraphs: Facilitating High-Order Modeling of the Computing Continuum. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24 Companion), May 7–11, 2024, London, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3629527.3651423

1 INTRODUCTION

The digital representation of the physical universe is a complex task that requires understanding the concept of morphism first. Morphism is a way to describe how different parts of a shape or structure relate to each other mathematically. It was first introduced by the French mathematician Henri Poincare in 1895 [1]. More than seventy years later, in 1968, the American mathematician Haskell Curry and logician William Alvin Howard applied this concept to computer science. They described the Curry-Howard correspondence, which shows that the proof of a computer system and the model of computation are the same kind of mathematical object [2]. This means we can model computer programs and systems as directed graphs commonly used today to represent infrastructures, data, and applications.

However, graphs have limitations when it comes to modelling modern *computing continuum* infrastructures and their interactions with Internet of Things (IoT) applications involving millions of data sources, as they can not express the scale of the data sources or application instances.

The problem of using graphs for modeling of distributed systems is further aggravated as researchers have recently integrated novel computing architectures in the computing continuum beyond the ones based on the traditional stored program von Neumann model, where the programs and data are stored in a single operating memory [3]. These novel architectures use different production processes, processing implementation, data representations and distributed memory models, known as non-Von Neumann architectures. They range from power-efficient single-board Artificial Intelligence (AI) accelerators to Quantum and Neuromorphic computers [4, 5]. While these architectures hold great potential for revolutionizing data processing and analysis in healthcare, transportation, and entertainment, integrating them with established cloud and edge computing paradigms remains challenging due to significant architectural heterogeneity, data representation, communication, and limited modelling tools. Unfortunately, extending the computing continuum with non-Von Neumann architectures causes multiple difficulties in application and infrastructure modelling, resource provisioning and execution optimisation [6].

To address the complexity of the computing continuum and the integration of non-Von Neumann architectures, we discuss the concept of hypergraphs, powerful mathematical objects generalising graphs [7], as potent tools for modelling. In hypergraphs, hyperedges can connect any number of hypervertices. Hypergraphs are more expressive than pair-wise classical graphs, allowing us to model the computing continuum and extreme-scale applications as mathematical objects with higher-order, high-dimensional relations. They can represent computing resources and data sources, regardless of their scale. Therefore, we can identify new relationships between resources and data sources by abstracting from the scale. For example, we can use a hypervertex to represent a set of computing continuum resources and connect it with a hyperedge to another hypervertex representing various data sources, regardless of their scale. By leveraging hypergraphs, we can expose previously unknown relations between the resources and identify hidden properties of the applications. To illustrate the benefits of

ICPE '24 Companion, May 7–11, 2024, London, United Kingdom

ACM ISBN 979-8-4007-0445-1/24/05.
https://doi.org/10.1145/3629527.3651423

these modelling approaches, let us discuss an example, in which a distributed application composed of three components interconnected in a specific topology is deterministically deployed on a given computing continuum infrastructure, which, in continuation, is connected to multiple sensing devices that monitor the environment. This implies that for modelling the system using classical graphs, we have to consider the application, infrastructure and sensor graphs in isolation and only afterwards to identify their interactions manually. On the other hand, we can model the application, infrastructure, sensing devices, and environment using hypergraphs as hyperedges. The hypervertices model the interactions between these hyperedges. This allows us to move away from the fixed size of the application and infrastructure and models them nevertheless of the scale (how many components, instances, infrastructures and sensing devices are available).

Therefore, the paper proposes the HyperContinuum conceptual platform for sustainable and scalable distributed applications processing over computing continuum infrastructures based on a hypergraph (HG) representation of the data, environment, infrastructure, and applications.

The paper has four sections. We first survey the related work in Section 2. Afterwards, we present the proposed conceptual architecture in Section 3 and Section 4 concludes the paper.

2 RELATED WORK

This section details the state-of-the-art optimisation of container orchestration systems.

2.1 Hypergraph applications modelling and sustainability analysis

The current application analysis approaches rely on pair-wise ordinary graphs or state machine representations to model the applications and the infrastructure below [7]. This limits their application for highly adaptive and heterogeneous systems, such as the computing continuum. From a performance point of view, the classical workflow and hardware-specific optimization approaches brought significant improvements in distributed applications execution, specifically in performance, energy management, and financial cost [8]. Unfortunately, these approaches primarily support large data centers with a relatively homogeneous set of resources with static topologies and performance profiles. They lack functionality for supporting complex application workflows, which can change the structure and conditional execution branches based on the input parameters and workload. Concretely, ordinary workflows and dynamics (i.e., changing the number and content of vertices and edges) lead to high variability in computational needs [9]. Therefore, the hypergraph and hyperworkflow models can be intelligently transformed into ordinary workflows for improved performance prediction. They demonstrate they can enable conditional algorithm/execution branch selection and advanced auto-scaling techniques to ensure better performance. Furthermore, energy consumption is a primary component of a computing infrastructure's total cost of ownership. Power consumption and thermal dissipation limit the achievable peak performance with lower cost.

2.2 Hyperworkflow optimisation and cognition with federated learning

Federated learning techniques where multiple decentralised devices or nodes collaborate to train a shared model while keeping data local to the devices [10]. It can be used to optimise hyperworkflows by improving the accuracy and efficiency of the model while ensuring the privacy and security of data. In hyperworkflows optimization, federated learning can be applied in several ways. One approach is to use federated learning to optimise decision-making processes in hyperworkflows, such as determining the next best task or predicting execution outcomes. Another approach is to use federated learning to improve the performance of machine learning models used in workflows. State-of-the-art methods in federated learning for workflow optimization involve advanced techniques such as federated transfer learning and federated reinforcement learning. These methods address challenges such as communication overhead, data heterogeneity, and model convergence in federated learning systems. Existing workflow management systems such as Pegasus [11], Apollo [12], and Askalon are centralised systems and either do not support machine learning (ML)-based workflow execution optimization or utilise conventional and centralized ML approaches. In such systems, ML systems are centralised, and workers periodically send local updates about the workload to a set of parameter servers, such as Tensorflow and traditional federated learning systems [13].

2.3 Overlay infrastructure provisioning

The workloads in the distributed computing continuum are often machine learning and data-intensive and have high requirements for performance; deployment strategies, e.g., offloading, computing close to data, and parallelizing the heavy tasks, are required due to the constraints of capacity, time constraints, and energy [14]. Therefore, infrastructure provisioning and deployment planning algorithms have been proposed based on critical paths, graph decomposition, and potential data traffic [15]. Most of the early work is based on data workflows with a deterministic performance model of the components on a set of given computing nodes. The data processing or machine learning workloads heavily depend on the volume and availability of the data, which results in new challenges to apply those existing approaches [16]. Machine learning-based infrastructure provisioning and deployment management, e.g., reinforcement learning, has been proposed in the past years; however, those approaches face challenges of low robustness when the workload patterns change.

3 CONCEPTUAL ARCHITECTURE

This section presents a conceptual architecture of the HyperContinuum framework.

The HyperContinuum high-level framework involves two conceptual layers displayed in Figure 1, creating an automated, sustainable loop for managing distributed applications as hypergraphs over the computing continuum with non-Von Neumann architectures:

- Hypergraph cognition layer covering the creation, optimisation and analysis of the hypergraphs;

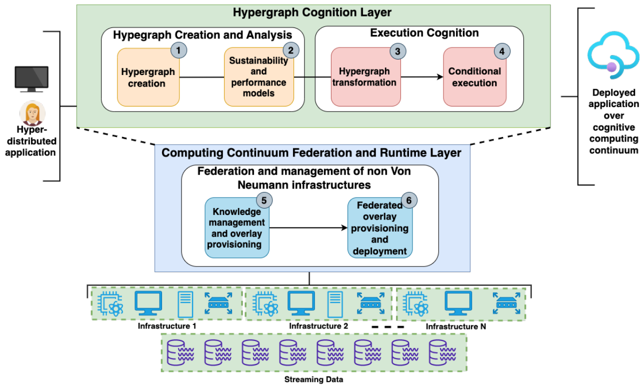

Figure 1: Conceptual life cycle of the HyperContinuum framework.

- Federation and management of non-Von Neumann infrastructures layer that focuses on provisioning computing continuum resources as logical overlays.

The *hypergraph cognition layer* facilitates creating, analysing, and optimising the applications' hyperworkflows comprising performance and energy models and optimization techniques. Hypergraph creation with sustainability and performance analysis analyses the distributed application's higher-order interactions with the environment and infrastructure and represents it as the hypergraph (See Figure 2a and b). Unlike classical approaches for managing distributed applications, which model the applications in isolation from the infrastructure and environment, we generalize the applications and the infrastructure as hypergraphs [17]. The hypergraphs allow us to model the distributed application components and all multi-dimensional interactions, including the interactions with the environment. This enables us to use a more realistic representation of the application and its interaction with the environment. Thereafter, as depicted in Figure 2c, when the data patterns and directions of the interactions between the application and environment are known, the hypergraph is transformed into a directed acyclic hypergraph (DAH). During the transformation from HG to DAH multiple sustainability metrics are considered, including energy wastage estimation by creating benchmarks for energy and performance modelling of non-Von Neumann architectures. The execution cognition, including the hypergraph transformation and conditional execution, utilises an intelligent distributed approach for cognitive optimization of the application DAH considering the possibility for conditional execution of the application branches based on external factors, such as users' location and cached results [18]. Furthermore, it applies an intelligent data distribution algorithm to only store the information of the applications and users on trusted storage infrastructures, thus complying with the data security standards [19]. Based on the workload, environmental parameters, input data, and scale of the systems and application, the system can transform the relevant part of the DAH to a directed acyclic graph (DAG) specifically tailored for the given execution depicted in Figure 2d.

The computing continuum *federation and management of non-Von Neumann infrastructures* layer focuses on the execution aspect

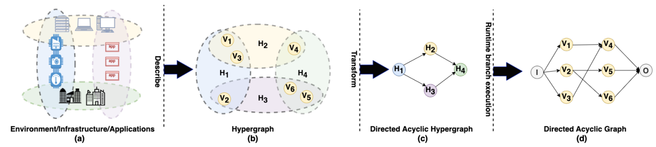

Figure 2: Transformation from hypergraph to ordinary graph.

of the DAH by providing automated configuration and provisioning of interoperable infrastructures and deployment. Federated infrastructure knowledge management and overlay provisioning enable provisioning over multiple computing continuum systems. It enables the creation of an interoperable resources overlay that includes heterogeneous hardware resources, including non-Von Neumann hardware, across multiple computing continuum infrastructures by implementing novel infrastructure knowledge management algorithms and approaches [20]. In addition, the layer manages the deployment of the given branches of the DAH as DAG, making it interoperable with any existing system. Lastly, it identifies suitable storage sites to create a virtualized distributed data federation and provides continuous infrastructure and later application monitoring [21].

4 CONCLUSION

This paper introduces a novel conceptual framework for modelling of the computing continuum and IoT applications as hypergraphs, with a primary focus on non-Von Neumann systems. The limitations of classical graph-based modeling in accommodating diverse architectures beyond traditional von Neumann models are highlighted. The utilization of hypergraphs emerges as a crucial solution, allowing for the representation of computing resources and data sources at any scale. This innovation facilitates the identification of new relationships and hidden properties, laying the foundation for the development of a federated and sustainable computing continuum.

The HyperContinuum conceptual platform describes novel concepts for resource and applications management algorithms tailored for distributed applications within next-generation computing continuum infrastructures based on novel von Neumann computer architectures. The platform introduces the concepts for high-order hypergraph applications representation, sustainability optimization for von Neumann architectures, automated cognition through federated learning for IoT application execution, and adaptive computing continuum resources provisioning. Overall, the HyperContinuum platform not only discusses the challenges posed by diverse computing infrastructures but also sets the stage for the future development of intelligent and sustainable systems. The paper underscores the importance of embracing hypergraph-based models in shaping the next era of computing, marking a significant step towards creating a more efficient and interconnected computing environment.

REFERENCES

[1] Jiří Adámek, Horst Herrlich, and George Strecker. *Abstract and concrete categories.* Wiley-Interscience, 1990.

[2] C-HL Ong and Charles A Stewart. A curry-howard foundation for functional computation with control. In *Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 215–227, 1997.

[3] Antara Ganguly, Rajeev Muralidhar, and Virendra Singh. Towards energy efficient non-von neumann architectures for deep learning. In *20th international symposium on quality electronic design (ISQED)*, pages 335–342. IEEE, 2019.

[4] National Academies of Sciences. Quantum computing: progress and prospects. 2019.

[5] Dragi Kimovski, Roland Mathá, Josef Hammer, Narges Mehran, Hermann Hellwagner, and Radu Prodan. Cloud, fog, or edge: Where to compute? *IEEE Internet Computing*, 25(4):30–36, 2021.

[6] Reza Farahani, Dragi Kimovski, Sashko Ristov, Alexandru Iosup, and Radu Prodan. Towards sustainable serverless processing of massive graphs on the computing continuum. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*, pages 221–226, 2023.

[7] Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete applied mathematics*, 42(2-3):177–201, 1993.

[8] Vincenzo De Maio and Dragi Kimovski. Multi-objective scheduling of extreme data scientific workflows in fog. *Future Generation Computer Systems*, 106:171–184, 2020.

[9] Björn B Brandenburg, John M Calandrino, and James H Anderson. On the scalability of real-time scheduling algorithms on multicore platforms: A case study. In *2008 Real-Time Systems Symposium*, pages 157–169. IEEE, 2008.

[10] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

[11] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira Da Silva, Miron Livny, et al. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, 46:17–35, 2015.

[12] Fedor Smirnov, Behnaz Pourmohseni, and Thomas Fahringer. Apollo: Modular and distributed runtime system for serverless function compositions on cloud, edge, and iot resources. In *Proceedings of the 1st Workshop on High Performance Serverless Computing*, pages 5–8, 2020.

[13] Guillem Ramirez-Gargallo, Marta Garcia-Gasulla, and Filippo Mantovani. Tensorflow on state-of-the-art hpc clusters: a machine learning use case. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 526–533. IEEE, 2019.

[14] Dumitru Roman, Radu Prodan, Nikolay Nikolov, Ahmet Soylu, Mihhail Matskin, Andrea Marrella, Dragi Kimovski, Brian Elvesæter, Anthony Simonet-Boulogne, Giannis Ledakis, et al. Big data pipelines on the computing continuum: Tapping the dark data. *Computer*, 55(11):74–84, 2022.

[15] Thang Le Duc, Rafael García Leiva, Paolo Casari, and Per-Olov Östberg. Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–39, 2019.

[16] Ali Shahidinejad and Mostafa Ghobaei-Arani. Joint computation offloading and resource provisioning for e dge-cloud computing environment: A machine learning-based approach. *Software: Practice and Experience*, 50(12):2212–2230, 2020.

[17] João Paulo A Almeida. Model-driven design of distributed applications. In *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops: OTM Confederated International Workshops and Posters, GADA, JTRES, MIOS, WORM, WOSE, PhDS, and INTEROP 2004, Agia Napa, Cyprus, October 25-29, 2004. Proceedings*, pages 854–865. Springer, 2004.

[18] José Carlos Fonseca, Vincent Nélis, Gurulingesh Raravi, and Luís Miguel Pinho. A multi-dag model for real-time parallel applications with conditional execution. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 1925–1932, 2015.

[19] Lori M Kaufman. Data security in the world of cloud computing. *IEEE Security & Privacy*, 7(4):61–64, 2009.

[20] Michael Maurer, Ivona Brandic, and Rizos Sakellariou. Adaptive resource configuration for cloud infrastructure management. *Future Generation Computer Systems*, 29(2):472–487, 2013.

[21] Polona Štefanič and Vlado Stankovski. Multi-criteria decision-making approach for container-based cloud applications: the switch and entice workbenches. *Tehnički vjesnik*, 27(3):1006–1013, 2020.