SERRANO: Transparent Application Deployment in a Secure, Accelerated and Cognitive Cloud Continuum

Aristotelis Kretsis[1], Panagiotis Kokkinos[1], Polyzois Soumplis[1], Juan Jose Vegas Olmos[2], Marcell Fehér[3], Márton Sipos[3]
Daniel E. Lucani[3], Dmitry Khabi[4], Dimosthenis Masouros[5], Kostas Siozios[5], Paraskevas Bourgos[6], Sofia Tsekeridou[6],
Ferad Zyulkyarov[7], Efstathios Karanastasis[8], Efthymios Chondrogiannis[8], Vassiliki Andronikou[8], Aitor Fernandez Gomez[9],
Silviu Panica[10], Gabriel Iuhasz[10], Anastassios Nanos[11], Charalampos Chalios[11] and Manos Varvarigos[1]
[1] Institute of Communication & Computer Systems, School of Electrical & Computer Engin., NTUA, Greece Email: vmanos@central.ntua.gr
[2] NVIDIA Corporation, Israel, Email: juanj@nvidia.com
[3] Chocolate Cloud ApS, Denmark, Email: daniel@chocolate-cloud.cc
[4] High Performance Computing Center, Universitaet Stuttgart, Germany, Email: khabi@hlrs.de
[5] Aristotle University of Thessaloniki, Greece, Email: ksiop@auth.gr
[6] Intrasoft International SA, Luxembourg, Email: Sofia.tsekeridou@intrasoft-intl.com
[7] Inbestme Europe Agencia de Valores S.A., Spain, Email: ferad.zyulkyarov@inbestme.com
[8] Innovation Acts Limited, Cyprus, Email: v.andronikou@innov-acts.com
[9] Ideko S. Coop, Spain, Email: afgomez@ideko.es
[10] Universitatea de Vest din Timişoara, Romania, Email: silviu.panica@e-uvt.ro
[11] Nubificus Ltd, United Kingdom, Email: ananos@nubificus.co.uk

***Abstract*—We are witnessing a wave of emerging cloud computing technologies and services that empower advanced applications from different vertical sectors, with diverse requirements. These trends give rise to a number of fundamental challenges that relate to the application deployment, the support of heterogeneous infrastructures and the provided security. In this setting, the SERRANO project steps in to define an intent-based paradigm of operating federated infrastructures consisting of edge, cloud and HPC resources, which will be realized through the SERRANO platform. Applications' high-level requirements will be translated to infrastructure-aware configuration parameters. SERRANO orchestration will then provide adaptive and efficient access to secure by design and accelerated resources. In this way, SERRANO will support cloud-native applications and services towards the cloud continuum.**

***Keywords*—cloud computing, edge computing, HPC, security, hardware acceleration, orchestration**

I. Introduction

Centralized cloud computing infrastructures are currently handling most of the processing and storage requirements of applications from different vertical sectors. Cloud computing constitutes a key component of modern economy in enabling the design and realization of novel digital services. In addition, there is a movement from top-down-designed architectures that apply centralized resource control, towards federations of loosely coupled autonomous or semi-autonomous edge computing systems, managed by multiple independent actors that are self-organized in a distributed manner. Edge computing offers computation and storage at the very edge of the network where data is produced, and has recently emerged as a way to reduce latency and limit the load that is carried to higher layers of the infrastructure hierarchy. High-performance computing (HPC), historically older than both cloud and edge computing, is based on on-premise high-end supercomputers that provides enormous capacity for computationally intense and exascale data analysis tasks and acts in a way that is complementary to that of cloud and edge computing.

The co-existence and use of the aforementioned computing paradigms gives rise to a number of fundamental challenges that relate to the application deployment, the support of heterogeneous infrastructures and the provided security. Inline with the above, the SERRANO project [1] proposes **an intent-driven** paradigm of **federated infrastructures**, with edge, cloud and HPC resources (Fig. 1), supporting a *develop once, deploy everywhere* approach, to be realized through the SERRANO platform.

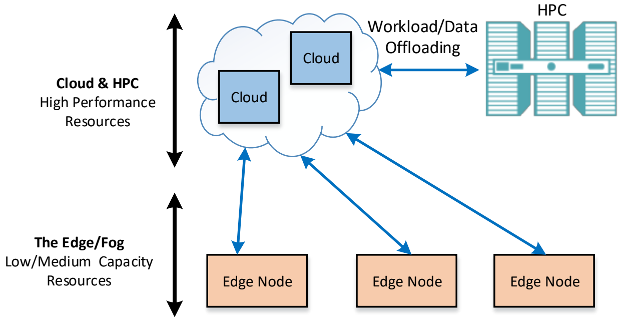

Fig. 1 Federated computing and storage infrastructures consisting of Edge, Cloud, and HPC resources.

SERRANO will provide advancements on several fronts, incorporating and improving several key technologies. In particular, SERRANO platform will provide: (i) security and privacy by design in distributed computing and storage infrastructures, (ii) application security and low-latency in multi-tenant environments, (iii) hardware acceleration and energy efficiency in developing, deploying and managing data-intensive applications, (iv) transparent application deployment across seamlessly integrated heterogeneous computing resources in edge, cloud and HPC, and (v) data-driven orchestration of network, computation and storage resources as well as of the applications themselves.

At the top level, SERRANO will create an abstraction layer that automates the process of application deploying functionality across the various computing technologies. This layer will be part of an infrastructure-agnostic automation process that translates applications' high-level requirements to infrastructure-aware configuration parameters. The SERRANO platform will automatically determine the most appropriate (computing, storage, networking) resources of the cloud continuum to be used by an application, and then transparently deploy workloads and coordinate the related data movement. A *sense, discern, infer, decide, and act* continuous control loop will run over an infinite time horizon to adjust resources and migrate the tasks, using feedback regarding the application's and the resources' state (Fig. 2). Service assurance mechanisms based on artificial intelligence and machine learning techniques will facilitate the autonomous adaptation and management of the deployed services and resources. These mechanisms will be dynamically triggered by a data-driven cloud computing and network telemetry framework that collects and analyses telemetry data across the distributed edge/cloud/HPC infrastructure.

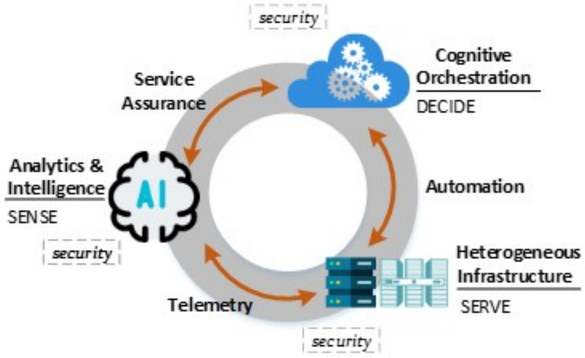

Fig. 2 The SERRANO platform, utilizing edge, cloud and HPC resources and empowering the everything as a service notion towards the cloud continuum.

The SERRANO platform will also develop hardware and software-based mechanisms that provide security, privacy and multi-tenancy by design. In this way, applications and users will be able to maintain control over integrity and privacy of their data when relying on publicly shared edge and cloud infrastructures. Also, SERRANO will capitalize on the benefits offered by hardware accelerators, used to execute prototype tasks that arise often in applications, coupled with novel transprecision computing mechanisms to exploit the accuracy versus resource usage tradeoff. The latter techniques will enable the dynamic adaptation of the computations' precision, based on application requirements, further improving the overall performance and energy efficiency of the infrastructures.

In the rest of the paper, we present the SERRANO concept, together with the research advancements that will be implemented throughout the duration of the project.

II. The SERRANO Concept And Architecture

A. Concept

SERRANO envisages the creation of an abstraction layer that will fully exploit the available resources and automate their use. SERRANO's abstraction mechanisms will support cloud-native application and services in reaping the benefits of the entire cloud continuum.

This abstraction layer will be part of an infrastructure agnostic automation process that will translate applications' high-level requirements to infrastructure-specific configuration parameters. It will further promote interoperability through the implementation of common description models for cloud and edge nodes, thus enabling services to run on top of heterogeneous clouds (involving different fabrics and geographic locations) and multi-technology edges.

Also, SERRANO will develop security and trustworthiness mechanisms, operating at multiple layers, to provide by design privacy and resiliency against security threats. SERRANO will also serve mission-critical applications through hardware and software acceleration on both the cloud and edge segment. In particular, concepts from approximate computing will be incorporated, offloading workload to application-agnostic configurable hardware accelerated kernels and reducing cloud and edge nodes' energy consumption and applications' execution delay.

The realization of SERRANO's ambition requires the efficient and seamless orchestration of the available heterogeneous resources. The overall orchestration will be performed in a cognitive, automated, integrated and holistic manner, enabling the continuous optimized allocation of the available heterogeneous resources. The orchestration mechanisms will leverage the advanced telemetry framework and AI/ML techniques to successfully react to unpredictable changes at the resource layer.

B. Architecture

Fig. 3 depicts the general SERRANO layered architecture along with its main functional components. The Resource Layer consists of heterogeneous edge, cloud and HPC computational and storage resources that encompass the SERRANO hardware innovations and the developed software mechanisms and frameworks. That exceptional unification of highly diverse resources provides the SERRANO platform the ability to cater for application and user constraints, while calibrating the configuration of available resources. Part of the resource layer are the network resources, whose performance is constantly and dynamically monitored.

The task of resource exposure to the upper layers is assigned to the Infrastructure Abstraction Layer that abstracts the peculiarities regarding the management and interaction of the individual resources. It also provides a modular design that ensures the extension of the infrastructure through the immediate integration of new hardware and software platforms at the Resource Layer. Integration with the low-level resources is achieved via the Unified Resource API and the developed orchestration drivers that provide the required mechanisms and resource profiles to enable efficient and transparent deployment of services across the heterogeneous Resource Layer. The Secure Infrastructure Layer contains all the mechanisms required to enable the secure and trustworthy sharing and access of the edge, cloud and HPC resources over heterogeneous networks. It abstracts the required actions for each individual resource type, operating as a security access broker that guarantees and enforces privacy and security requirements on data, through the Distributed Secure Storage API. Across the SERRANO ecosystem resides the Infrastructure, Platform and Application Telemetry stack that tracks and logs metrics across the infrastructure and deployed applications. Its purpose is to fuel the platform and applications with data from all levels, enabling the implementation of data-driven and cognitive automation mechanisms.

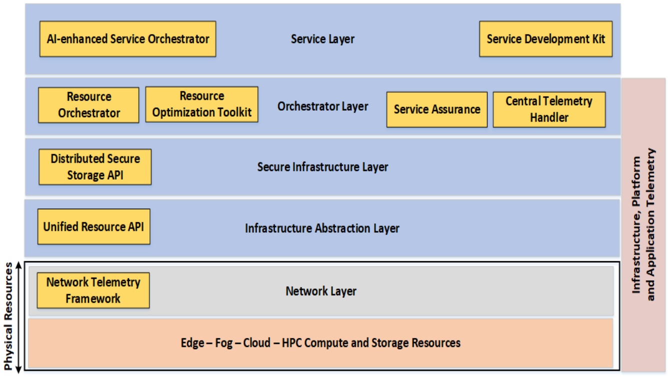
AI-enhanced Service Orchestrator
Service Layer
Service Development Kit
Resource Orchestrator
Resource Optimization Toolkit
Orchestrator Layer
Service Assurance
Central Telemetry Handler
Distributed Secure Storage API
Secure Infrastructure Layer
Unified Resource API
Infrastructure Abstraction Layer
Network Telemetry Framework
Network Layer
Edge – Fog – Cloud – HPC Compute and Storage Resources
Physical Resources
Infrastructure, Platform and Application Telemetry

Fig. 3 The SERRANO Overall Layered Architecture

The Orchestration Layer ensures efficient service orchestration and resource management in the disaggregated and heterogeneous SERRANO infrastructure. Taking as input from the Service Layer the application high-level requirements and a candidate set of appropriate resource configurations, the central Resource Orchestrator allocates the resources and decides their proper configuration so as to fulfill the individual tasks constraints, while it also automatically coordinates the necessary supplemental actions (e.g., transfer of required data). The Resource Optimization Toolkit provides network-aware joint computational and storage resource allocation and service placement algorithms, leveraging optimization and AI/MI techniques, with special emphasis on energy consumption, robustness and latency. The Central Telemetry Handler, which is the root in SERRANO's telemetry hierarchy, collects and analyses the monitoring information to provide the other components with useful information and insights on the infrastructure's current state. The Central Service Assurance component manages the runtime lifecycle of each service deployment across the SERRANO heterogeneous infrastructure. Based on the services' specific needs, the infrastructure's current state and notifications received from the Central Telemetry Handler, it can automatically trigger, proactively and reactively, re-optimization actions to maintain the required performance level.

Finally, the Service Layer contains the AI-enhanced Service Orchestrator that analyses the applications in order to automatically determine the most appropriate platform type for their deployment, translating their high-level requirements into specific infrastructure related operational constraints and orchestration objectives. The SERRANO SDK facilitates the rapid development and deployment of innovative applications that fully leverage the provided innovations.

III. Provision of Security and Privacy By Design

SERRANO will enhance the security in multi-tenant environments by providing specific solutions for the low-level software stack to ensure non-interference and controlled data access among different and possibly concurrently running applications. To mitigate the security issues that arise in these environments, SERRANO tackles two major issues: (i) minimize attack surface and (ii) ensure trusted execution. The attack surface minimization will be based on an efficient and secure mechanism for applications to be executed as part of a self-contained machine image, keeping only the needed dependencies, and reducing the execution and exposure of trusted/kernel code to the minimum possible. To this end, SERRANO will build on the solo5 architecture [2] to minimize the kernel code being accessed from user-space applications and Virtual Machines images/Unikernels [3]. To ensure trusted execution, SERRANO will introduce a strict security attestation mechanism at the lowest level of the software stack (VMM and OS-glue to the application) that will leverage hardware extensions and software mechanisms to ensure trusted boot and execution while keeping data access controlled to authenticated parties. Finally, SERRANO will develop novel AI/ML-based methods, as part of the overall telemetry framework, which will correlate the monitoring information in order to accurately identify and address the continuously evolving network attacks in a scalable and dynamic manner.

SERRANO will also leverage skyflok.com service that provides secure, GDPR-compliant, privacy-aware data storage and sharing, extending it with new security features and integrating cloud, edge and HPC resources. This service already supports AES-256 encryption for protecting data at rest and network coding [4], for slicing data, creating random linear combinations as an added encryption mechanism, generating added redundancy to compensate for losses, and distributes coded fragments of each file to a number of cloud providers and locations around the globe. SERRANO will expand the skyflok.com service capabilities to edge devices and will develop intelligent and autonomous mechanisms for optimal resource allocation based on multiple criteria (e.g. security, user mobility, latency requirements). These enhancements will enable self-managed and trusted load sharing across the disaggregated infrastructure.

Recently, the Non-Volatile Memory Express (NVMe) over TCP [5] has been gaining ground for many data-driven businesses as an efficient way for storage disaggregation due to its ability to transport NVMe over a standard IP network. The Transport Layer Security (TLS) is the de-facto standard for encrypting TCP traffic in data centers, but when applied to NVMEoTCP it requires copying data between TCP and the NVME layers, thereby reducing performance and increasing CPU load. To address this, SERRANO will accelerate and offload from CPU (based on powerful Mellanox Bluefield-2 NIC) both the NVMEoTCP and TLS processes by following a novel three-fold approach: (i) removing copy overhead, (ii) accelerating the calculation/validation overhead for the

NVMEoTCP CRC32 signature, and (iii) providing TLS symmetric acceleration that will hugely remove overheads associated with the encryption process. Moreover, SERRANO targets to bring the self-encrypting drive concept at the Just a Bunch of Flash (JBOF) platforms, which are considered as a primary storing option for the future cloud infrastructures, by using the NVMe over Fabrics (NVMe-oF) network protocol and NVMe flash as the physical drives. SERRANO will develop in-line AES-XTS encryption of data from the network before storing to JBOF, based on Mellanox's BlueField-2 system-on-chip units. This approach will lead to a scalable secure storage system, enabling massive storage deletion by crypto-key handling (rather than data handling) that will be able to support secure exchange of massive data among CPUs, processing blocks and storage blocks.

IV. Hardware and Software Acceleration

A significant proportion of the tasks that run on heterogeneous computing infrastructures involve data intensive applications that are programmed using traditional frameworks and widely employ data management technologies. In these data process environments, we can identify certain extremely intensive computationally tasks that are common across several applications. SERRANO approach is to implement these prototype tasks that arise often in applications at the hardware level, as customized intellectual property (IP) accelerators. To further capitalize on this idea, SERRANO will provide hardware accelerators with varying error characteristics, based on the applications' requirements, to further improve the overall performance and energy efficiency of heterogeneous infrastructures. The proposed solution relies on a repository (library) with several customizable templates of approximate kernels per IP kernel that trade-off accuracy with energy consumption. SERRANO is implementing the appropriate run-time execution controller (Fig. 4), based on the necessary enhancements to the REMAP [6] framework that will automatically select and configure (fine-tune) the most suitable approximate kernel according to the application's workload. The derived synthesis framework will enable the generation of multi-level approximate hardware accelerators that satisfy a given error bound. Hence, SERRANO will enable dynamic and input-driven approximations for real-time analysis of very large data volumes.

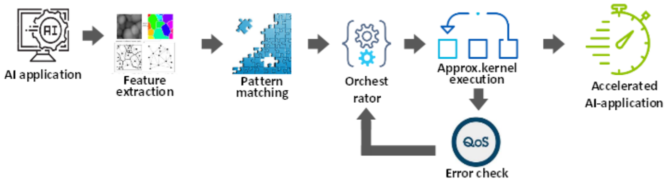

Fig. 4 The runtime execution controller for improving Performance/Watt metric while satisfying workload's accuracy based on REMAP framework.

The inherent parallelism found in hardware accelerators can provide flexibility in building, deploying and managing data-intensive applications. This requires, however, holistic SDK frameworks for rapid prototyping solutions, such as Hardware-in-the-Loop (HiL), to enable the simulation, debugging and verification of hardware/software (HW/SW) co-designed computational kernels that are substantially faster than current state of the art. SERRANO extends the Plug&Chip API [7] for rapid prototyping to make it aware of the different types of hardware accelerators and approximate kernels. In this framework, hardware architects could model part of the system's functionality on a higher level of abstraction and then employ High-Level Synthesis (HLS) tools in order to map these functionalities to silicon. Hence, developers will get an early start on their work, long before the hardware platform is finalized.

To ensure isolation of workload execution in the cloud and at the edge, virtualization techniques are being employed. Adding an extra layer of abstraction complicates the stack significantly, leading to resources under-utilization, complicated time/space sharing of hardware, etc. To alleviate this, paravirtual devices expose very simple APIs to applications running as VM instances. Instead of exposing the full API of a hardware accelerator, e.g. GPU, SERRANO will develop a generic hardware acceleration paravirtual device in order to provide coarse-grained hardware acceleration semantics. The paravirtual device will expose a simple, function-based API for popular (prototype) operators, such as image inference, k-means clustering, cryptographic operations, etc, that can be accelerated by any available hardware device on the host, using generic frameworks such as Tensorflow, PyTorch, etc. The advantages of this approach are two-fold. First, we enable hardware acceleration for serverless functions on the edge, while preserving strong isolation semantics between tenants. Second, we relieve the user from the need to utilize complex frameworks and complicated software stacks.

SERRANO introduces novel transprecision computing mechanisms to enable the dynamic adaptation of the precision of computations performed in the edge, thus reducing computational requirements, energy consumption, memory usage, bandwidth utilization and experienced latency. Also, appropriate AI/ML methods will be developed to provide analytic insights for the dynamic adaptation of the required precision level. Moreover, the delivered gains can be further maximized by strategically applying in a coordinated manner approximate computations and approximate data transfer at the edge. In this case, SERRANO aims to assess the approximations' uncertainties by validating the above methodologies through HPC simulations of much higher precision and by developing appropriate Verification, Validation and Uncertainty Quantification (VVUQ) methods. The implemented VVUQ framework enables the identification of uncertainties and the validation of results for various levels of input precision. These mechanisms can be coupled with AI/ML techniques to estimate the influence of the selected accuracy for the input data, in order to optimally fine-tune the dynamic adaptation of compute and transfer approximations. Since, the efficient (in terms of time and energy) transfer of very large data volumes from and to HPC platforms is a critical factor that determines also the ability to efficiently offload tasks at HPC, SERRANO will study the exploitation of these innovations in the HPC domain.

V. Application Deployment

Cloud-native applications can be divided into a concrete set of components-microservices that use different data and are deployed independently on distributed heterogeneous resources. Hence, new abstractions are needed to support a *develop once, deploy everywhere* paradigm that allows developers to focus solely on business logic.

SERRANO is developing the ARDIA (A Resource reference model for Data-Intensive Applications) modelling framework for capturing heterogeneous data, infrastructure and network deployments. ARDIA provides the abstractions and common description models required in order to promote resource interoperability and composition, and the transparent deployment and mobility of applications' workload and data across the entire computing continuum. These abstractions also facilitate the automatic and transparent combining of hardware and software resources across multiple locations based on the applications' needs. ARDIA follows a three-distinct dimensions design. The first is the lifecycle dimension that captures different stages that heterogeneous resources undergo, including creation, operation, evolution, dissolution, and metamorphosis. The second is the environment dimension that covers the structure of deployments and their relationships, the components or resources, the processes within an application and the behavior with regards to principles, policies or rules. The third dimension is the intent, which covers SERRANO's support for higher levels of abstraction in service definitions that are translated to automated and proactive adjustments based on service requirements.

Moreover, a lifecycle methodology (Fig. 5) has been adopted to absolve the developers from the burden of resource allocation, application monitoring and scaling.

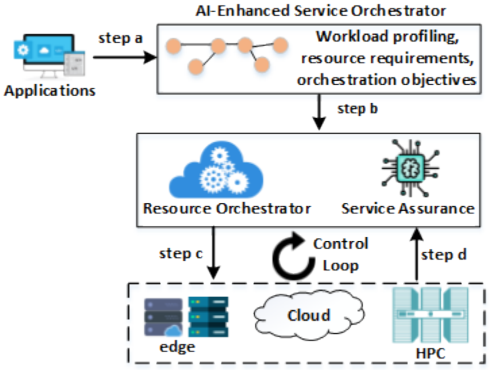

Fig. 5 Lifecycle workflow for SERRANO applications

Initially, users provide their applications along with a high-level infrastructure agnostic (application intent) description of their requirements (step a). Next, SERRANO through the AI-enhanced Service Orchestrator based on neural networks and intelligent AI/ML forecasting methods (including deep reinforcement learning techniques) performs the application profiling and the decomposition of the high-level requirements into specific service goals (step b). This uses multiple optimization criteria, forming a hierarchy of objectives, by breaking service goals into individual infrastructure-specific resource and performance requirements. For the communication with the Resource Orchestrator we are developing the ARDIA API, part of the overall SDK. Then, in step c, the actual allocation of resources to application's microservices takes place using cognitive orchestration mechanisms that aim to satisfy the orchestration objectives and ensure optimal use of the infrastructure. SERRANO orchestration mechanisms are also responsible for coordinating the application deployment and the efficient movement of required data across the selected resources. Finally, the service assurance mechanisms based on real-time telemetry of services and appropriate machine reasoning techniques safeguard that applications perform as intended (step d), while they proactively execute any required re-optimization.

VI. SERRANO Platform

The SERRANO platform integrates the project's developed technologies and mechanisms, coupling them with orchestration and telemetry/monitoring functionalities. SERRANO targets a hierarchical architecture (Fig. 6), consisting of a central Resource Orchestrator and resource-hosted Local Orchestrators, for end-to-end cognitive orchestration together with closed-loop control, based on the principles of observe, decide and act [8].

The Resource Orchestrator receives individual resource and performance requirements from the AI-enhanced Service Orchestrator based on applications the latter serves and decides the placement of the service mesh. Multi-Objective AI/ML algorithms are used for allocating the edge, cloud and HPC resources so as to satisfy the applications' requirements. Next, the Resource Orchestrator, based on the Resource Optimization Toolkit (RTO)'s outcomes, assigns the workloads to the selected resources along with the desired performance state (intent) and coordinates the required data movement (utilizing the Distributed Secure Storage API). SERRANO Resource Orchestrator follows a declarative approach, instead of an imperative one, for describing the workload requirements to the Local Orchestrator. This provides several degrees of freedom to the Local Orchestrator for serving in an optimal manner the "request", satisfying both the central orchestrator and the resource's objectives (step c). Then, the control is passed to the Local Orchestrators that are responsible for the actual deployment based on the desired performance requirements. Service assurance mechanisms also take place in the Resource Orchestrator and in the Local Orchestrator level for safeguarding that the overall application and the locally executed workloads have the desired performance (step d). Service assurance will be met through continuous machine reasoning analysis over the collected telemetry information (Enhanced Telemetry Agents and Central Telemetry Handler). The service assurance procedures may trigger self-optimization procedures in central and local level and pro-active actions to mitigate any issues. The goal can be achieved both reactively (when the state of the resources degrades to an extent that it violates a goal, this is followed by a reaction to overcome the disturbance and reach the goal again) and proactively (using predictions based on past experience we can foresee a likely change in the state of the resources and act in advance to avoid the violation of the goal).

SERRANO's hierarchical solution enables the Resource Orchestrator to manage the underlying heterogeneous infrastructure at a more abstract and disaggregated manner compared to the resource-aware Local Orchestrators. To further enhance the intelligence of the overall platform, SERRANO telemetry mechanisms will feed the developed AI/ML methods with key performance metrics and post-execution validation information from the individual edge, cloud and HPC platforms. The communication between the Resource Orchestrator, the Local Orchestrators, the service assurance and monitoring mechanisms will be performed through the Unified Resource API and Telemetry API, part of the SERRANO SDK.

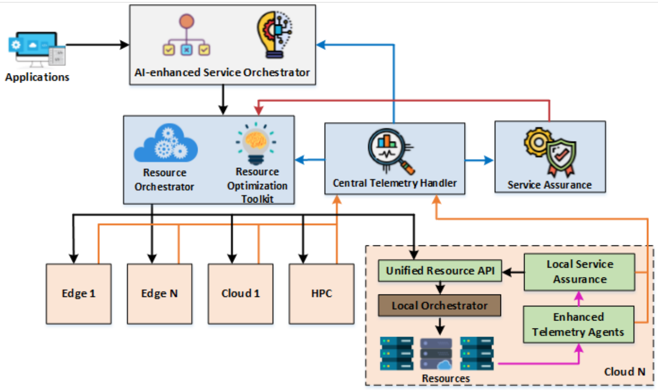

Fig. 6 SERRANO distributed orchestration and service deployment

The success and wide acceptance of any platform largely depends on the functionality and services offered to end users and application developers. Moreover, it is widely recognized that a company's competitiveness depends directly on its capacity to realize new ideas with the minimum time to market. SERRANO aims to deliver a holistic Service Development Kit (SDK) based on seamlessly heterogeneous architecture paradigm to improve the productivity of developers in building, deploying and managing novel applications, while having greater control over computing, storage, and network infrastructures. SERRANO SDK will include a set of well-defined APIs, adopting a transparent approach where there are no hidden internal APIs. More specifically, the SDK will include the ARDIA API, Unified Resource API, Telemetry API, Distributed Secure Storage API, Plug&Chip API, REMAP – VVUQ API.

VII. Use Cases

To highlight the proposed ecosystem's scientific and technological significance, SERRANO will demonstrate three high impact use cases (i) Secure Distributed Storage, (ii) High-performance Fintech Analysis and (iii) Machine Anomaly Detection in Manufacturing for Industry 4.0

- **Secure Distributed Storage:** This use case focuses on providing secure and high-performance storage and sharing of various data types at the edge of the network. In particular, SERRANO aims to break the typical trade-off between security and performance, by utilizing a combination of multiple edge locations and even multiple cloud computing and storage services/providers. The targeted architecture consists of multi-cloud and multi-edge subsystems that will (i) deliver a more robust and secure platform, and (ii) use novel security and resource allocation techniques to manage data privacy for the stored and processed data.
- **High-performance Fintech Analysis:** This use case focuses on supporting fintech analysis procedures with improved security and performance. In particular, the SERRANO platform will enable the transparent deployment of fintech related workloads across multiple resources (edge, cloud, HPC) and respective providers. This will lower the associated costs and enable on demand scalability, selecting the most appropriate cloud and HPC infrastructures for the compute intensive operations required for portfolio and market analysis.
- **Machine Anomaly Detection in Manufacturing for Industry 4.0:** Some of the state-of-the-art techniques for predictive maintenance require for a machine to stop, before performing the respective analysis. Another approach is to perform these analyses continuously, while machines continue running at 100%. In this way, the manufacturer can detect critical events in real time. SERRANO platform will be used in the latter approach, to orchestrate optimally data and computational movement in the edge, cloud and HPC continuum, handling the high volumes of data generated in real-time by the high-frequency and high-accuracy sensors.

VIII. Conclusions

In this paper, we describe the concept and the innovations of the SERRANO project [1]. SERRANO is creating an abstraction layer that automates the applications' deployment process in edge, cloud, and HPC resources. SERRANO is also developing hardware and software-based mechanisms that provide security, privacy and multi-tenancy by design, along with acceleration. Finally, SERRANO targets a hierarchical orchestration architecture for end-to-end cognitive resource allocation, together with closed-loop control for automatic adaptation.

Acknowledgment

This work is supported by the EU research project SERRANO (101017168).

References

[1] SERRANO project web site, www.ict-serrano.eu

[2] Dan Williams, "Solo5: Building a Unikernel Base From Scratch," CIF16, Jan 2016.

[3] A. Madhavapeddy et al, "Unikernels: The Rise of the Virtual Library Operating System," 57, pp. 61-69. 10.1145/2541883.2541895, 2014.

[4] A. Dimakis et al, "Network coding for distributed storage systems," IEEE Int. Conf. on Computer Comm. (INFOCOM), pp. 2000–2008, 2007.

[5] NVMEoTCP, Mellanox, https://www.mellanox.com/news/press_release/mellanox-propels-nvmetcp-and-roce-fabrics-new-heights, last accessed April 2021.

[6] G. Zervakis et al., "VADER: Voltage-Driven Netlist Pruning for Cross-Layer Approximate Arithmetic Circuits", IEEE Trans. on VLSI Systems, Vol. 27, no. 6, pp. 1460-1464, June 2019.

[7] D. Diamantopoulos et al., "Plug&Chip: A Framework for Supporting Rapid Prototyping of 3D Hybrid Virtual SoCs", ACM Trans. Embed. Comput. Syst. 13, 5s, Article 168, Dec. 2014.

[8] K. Christodoulopoulos, et al., "ORCHESTRA-Optical performance monitoring enabling flexible networking." IEEE International Conference on Transparent Optical Networks (ICTON), 2015.