Towards an Open Heterogeneous Smart Cloud

Josep Ll. Berral, Aaron Call, Ramon Canal
Universitat Politècnica de Catalunya, Barcelona Supercomputing Center
Barcelona, Spain

Introduction and Motivation

Recent advances on Pervasive Computing integration (IoT, wearables and Edge), Big Data technologies (analytics and Cloud) and Artificial Intelligence capabilities (machine learning for recommendation and prediction) allowed the expansion of ICT technologies in the day-by-day lives of citizens. With an potential exponential growth on data, resources and usage (users and involved devices), systems and infrastructures require adaptability, scalability, resilience, trustworthiness and sustaniability. And those properties are not always fully available or even aligned, and decisions must be made to keep a fair trade-off in used resources (including energy) and quality of service (including privacy and responsiveness).

To achieve this, different aspects must be targeted: (1) Open access to the underlying technologies; (2) Awareness of the impact on energy, availability and trust; and (3) Automation of the management procedures.

Initiatives such as RISC-V are investing on the creation of Open Hardware Specifications to grant access of efficient computing technologies to all levels of industry and society. Through an open specification, manufacturers and developers are allowed to freely create a new technological fabric of services and applications, on top of a new mesh of an heterogeneous mesh of devices unified by a common standard. Also at this time, Cloud architectures have become commodity for services and applications. Being capable to move processing closer to the Edge and IoT devices is allowing new applications to provide immediate and autonomous response, keep data closer and more "in control" of users, and amortize consumed energy by low-power Edge devices. For this, open architecture initiatives also focus on adoption in and out of the Cloud towards embedded on IoT and Edge.

Proper management of resources, from Cloud to Edge and IoT, requires "smart" approaches. **We have reached a limit on current methodologies for management, where the exponential growth of users and resources make traditional orchestration inefficient (not scalable, not reliable, not adaptive).** For this, leveraging Artificial Intelligence methods towards knowledge acquisition on top of monitoring is an advance that will allow better decision making on resource provisioning. Being able to treat distributed resources as "whole" resources, and virtualize centralized resources into smaller available resources, allows efficient access for applications to tailored resources, but makes orchestration more complex.

Therefore, while current research topics move in different and specialised directions on efficiency in Cloud/Edge infrastructures, coordination of distributed networks of computing devices, design of new and open standards, and feasible applications of Artificial Intelligence into ICT, there is the need to converge once again into a common ground where all advances help to push ICT society in a focused direction (before diverging again on future needs and discoveries).

State of the Art

The European Commission, following the strategic interest in EU industry and academia, is pushing towards a mature RISC-V development around the European Processor Initiative (EPI) [1], seeking technological sovereignty over the reduced group of non-European designers and manufacturers. While the software industry is investing into RISC-V portability, e.g. Linux Fedora distributions [2, 3] or Google Android support for such [4], the hardware industry is advancing on RISC-V standards along with designers and manufacturers [5].

Smart usage of Edge-Cloud resources implies the use of AI towards resource management, and provisioning towards AI on such resources. Critical AI applications in different domains such as medical [6][7] and space [8], involving data analytics and scientific computing, are moving towards Distributed Learning (i.e., Federated and Swarm Learning). Edge-Cloud Continuum architectures play an important role on the feasible deployment, and current research focuses on leveraging the same advances on AI for autonomous and private resource management [9][10].

Current efforts on heterogeneous resource optimization Cloud-Edge focus on disaggregation, tailoring components in co-design (HW and SW levels). As an example, NVIDIA is releasing GPU virtualization through rCUDA [11], allowing distributed GPU sharing along hyper-distribtued architectures. Privacy and security primitives are on the spotlight such as authentication methods and access control, to enforce secure access policies to private data and systems, through attestation techniques and cryptographic APIs. Hardware manufacturers such as Intel, AMD or ARM, are making first-steps on secure processors (e.g ARM Trustzone [12]), waiting for higher-level software and platform components to become fully effective on the Cloud-Edge. Same with hardware advances on fault tolerance on virtualization, e.g. [13], moving towards system availability and dependability.

Research Challenges

The principal challenge on expanding the Cloud-Edge into heterogeneous and novel architectures (i.e., RISC-V) focuses on its transparent management and scheduling to meet Quality of Service (QoS) guarantees. This includes monitoring and controlling the underlying HW from the OS, then identifying the proper operating points of such HW to guarantee Quality of Service (QoS). Performance, energy, reliability and security define the QoS, not only for regular workloads, but even more for Data Analytics and AI-based tasks, in which privacy, accuracy and trustworthiness must be ensured. Nonetheless, energy efficiency must be proportionated when leveraging resources to enhance AI methods. So, architecture orchestration must consider energy consumption for minization, while ensuring availability and dependability.

The second challenge is the seamless integration of heterogeneous resources (including accelerators) to ensure QoS. Heterogeneity also includes the choice of different resources for computing, transmission and storage. The Cloud-Edge is composed by a wide and distributed set of resources, each designed for different purposes and properties, that must be orchestrated aware of applications requirements and resource properties. As there is no "one fits all" policy, the management must consider application and resources performance, power, security and availability as indicators, capable of adapting to dynamic circumstances for load, traffic, variability and faults, and aware of data generation, location and consumption.

Finally, the third challenge is the integration of AI and Machine Learning (ML) as a fundamental part for knowledge-driven management in hyper-distributed and federated infrastructures. Data collection, model training and inference for decision making must be deployed across the infrastructure, and not only specific resources or applications, maximizing automation of resource and data provisioning. With the increasing amount of actors in the Edge/IoT ecosystem, handling users, data sources and consumers, devices and resources, ad-hoc or static policies need to be substituted or enhanced by more dynamic ones. And Knowledge Systems such as AI/ML methods have demonstrated the potential to drive the Continuum in a trustworthy manner.

References

[1] M. Kovač, "European processor initiative: The industrial cornerstone of eurohpc for exascale era," in *Proceedings of the 16th ACM International Conference on Computing Frontiers*, ser. CF '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 319. [Online]. Available: https://doi.org/10.1145/3310273.3323432

[2] "Fedora/RISC-V project homepage," https://fedoraproject.org/wiki/Architectures/RISC-V, accessed: 2023-03-23.

[3] M. N. Ince, J. Ledet, and M. Gunay, "Building an open source linux computing system on risc-v," pp. 1–4, 2019.

[4] "Android Open Source Project to RISC-V," https://www.eenewseurope.com/en/android-open-source-project-ports-to-risc-v/, accessed: 2023-03-06.

[5] "RISC-V Shines at Embedded World With New Specs and Processors," https://www.allaboutcircuits.com/news/risc-v-shines-at-embedded-world-with-new-specs-and-processors/, accessed: 2023-03-23.

[6] "INCISIVE: Improving cancer diagnosis and prediction with AI and big data," https://incisive-project.eu/, accessed: 2023-04-14.

[7] "SECURED: Scaling Up Secure Processing, Anonymization And Generation Of Health Data," https://secured-project.eu/, accessed: 2023-04-14.

[8] "CALLISTO: Copernicus Artificial Intelligence Services and data fusion," https://callisto-h2020.eu/, accessed: 2023-04-14.

[9] "NEARDATA: Extreme Near-Data Processing Platform," https://neardata.eu/, accessed: 2023-04-14.

[10] "CLOUDSKIN: Adaptive virtualization for AI-enabled Cloud-edge Continuum," https://cloudskin.eu/, accessed: 2023-04-14.

[11] C. Reaño, F. Silla, G. Shainer, and S. Schultz, "Local and remote gpus perform similar with edr 100g infiniband," in *Proceedings of the Industrial Track of the 16th International Middleware Conference*, ser. Middleware Industry '15. New York, NY, USA: Association for Computing Machinery, 2015. [Online]. Available: https://doi.org/10.1145/2830013.2830015

[12] "TrustZone for Cortex-A: SoC and CPU System-Wide Approach to Security," https://www.arm.com/technologies/trustzone-for-cortex-a, accessed: 2023-04-14.

[13] X. Jin, S. Park, T. Sheng, R. Chen, Z. Shan, and Y. Zhou, "Ftxen: Making hypervisor resilient to hardware faults on relaxed cores," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 2015, pp. 451–462.