OPERA: a Low Power Approach to the Next Generation Cloud Infrastructures

Alberto Scionti*, Pietro Ruiu*, Olivier Terzo*, Joel Nider†, Craig Petrie‡, Niccolò Baldoni*
* Istituto Superiore Mario Boella (ISMB), Torino, Italy
† IBM Research and Development, Haifa Research Lab
‡ Nallatech Ltd., Glasgow, UK
* Teseo S.p.A, Torino, Italy
E-mails: *{scionti,ruiu,terzo}@ismb.it, †joeln@il.ibm.com, ‡c.petrie@nallatech.com, *n.baldoni@teseo.clemessy.com

***Abstract*— The continuous evolution of information and communication technology has led to a change in the adopted computing paradigms over time. Cloud computing is an emerging paradigm in which users, depending on their specific requirements, access to a shared pool of computing resources dynamically allocated. Cloud computing represents, with respect to Grid computing, the evolutionary step towards the implementation of a ubiquitous computing service. Such paradigm leverages on the infrastructural capabilities (compute, storage, and network) of modern data centers to provide an adequate level of computational power able to satisfy users' requests. However, trying to continuously increase such capabilities comes at the cost of an increased energy consumption. Energy efficiency is, therefore, one of the major challenges that cloud providers must address.**

The OPERA project aims at bringing innovative solutions to increase the energy efficiency of cloud infrastructures, by leveraging on modular, high-density, heterogeneous and low power computing systems, which are able to cover the whole computing continuum. To this end, the project will design a high-density server solution in which low power processors and FPGA devices will be used to accelerate cloud workloads. High-speed optical interconnections will be used to connect the proposed server with high-performance nodes, such as OpenPOWER-based machines. Cyber-Physical Systems (CPS) represents a natural extension of cloud infrastructures since they can collect and process data locally, more specifically where they were generated. OPERA aims at researching energy efficiency of such cloud end-nodes by designing an ultra-low power computing system with reconfigurable radio frequency capabilities. The effectiveness of the whole platform will be demonstrated with key scenarios, specifically a road traffic monitoring application, the deployment of a virtual desktop infrastructure, and the deployment of a small data center on a truck.

I. Introduction

The continuous advancements in silicon technology allow the implementation of always more capable computing systems. Nowadays, such systems known as Cyber-Physical Systems (CPS), are enough powerful to enable a direct real-time interaction with the surrounding environment. CPS are at the basis of the implementation of new services, as well as the generation of an enormous amount of data to analyze, thanks to their capability of sensing/acting on the environment where they are deployed. To process the enormous amount of data generated by networks of these smart systems, large computing infrastructures are required in the back-end. Cloud computing is the set of technologies that allow to process and analyze such data, as well as to provide advanced services responding to the societal and industrial needs. These technologies also provide the substrate that enables CPS to transparently transfer data and get back useful information to react to the changes in the environment where they operate. However, such welcome capabilities are today counterbalanced by the high power consumption and energy inefficiencies of the silicon-based technology. New ways of implementing these systems are thus required to provide real-time responsiveness, scalability, and energy efficiency. To properly respond to these challenges, the OPERA project aims at researching the integration of high-performance, low power processing elements within highly modular and scalable architectures. Furthermore, OPERA realizes that heterogeneity is a key factor for improving energy efficiency and computational capabilities at the same time. To this end, solutions delivered in the project will largely leverage on reconfigurable devices (FPGA) and specialized hardware modules to maximize performance and reduce power consumption. The project started in December 2015, will span over 3 years. The coordination of the project is carried out by STMicroelectronics, along with the technical support of the Istituto Superiore Mario Boella – ISMB (technical coordinator). STMicroelectronics is also responsible for designing an ultra-low power CPS equipped with a reconfigurable radio communication module. HPE, along with the strong contribution of Nallatech, is responsible for designing a highly-scalable high-density server. In particular, Nallatech will develop a processing board sporting a high-performance FPGA device, capable of managing optical link interconnections. Nalltech also provides the tools needed to easily programming such kind of boards. Partners IBM, ISMB, and Technion will contribute to the definition of an automated method to decompose workloads into independent tasks which are executed as microservices on the cloud infrastructure. Partners CSI Piemonte, Le Département de Isère, Neavia technologies, and Teseo–Clemessy will provide support in setting up the three use cases. To ensure that energy efficiency improvements will be achieved over current cloud infrastructures, partner Certios will provide expertise for the definition of metrics used to assess the energy efficiency of the OPERA cloud platform.

II. Project objectives

The OPERA consortium recognizes the ambitious challenge of amply improving the energy efficiency of modern cloud infrastructures. To overcome current limitations, we identified the following specific objectives regarding the design of next generation scalable servers:

- Exploiting heterogeneous multicore processors (i.e., ARM, X86_64, and OpenPOWER), along with spe-

978-1-5090-2817-7/16 $31.00 © 2016 IEEE
DOI 10.1109/DSD.2016.63

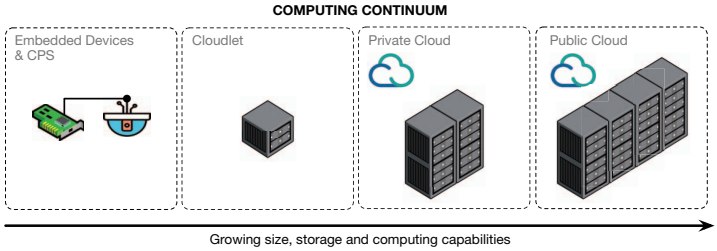

Figure 1. A visual representation of the computing continuum: from left (smart embedded devices) to right (public cloud data centers), computing and storage capabilities grow. Along with compute and storage, energy consumption grows as well.

cialized accelerators based on reconfigurable devices (FPGAs);
- Automatically splitting workloads into a set of independent modules which can be executed as microservices on the most appropriate device;
- Leveraging on cutting-edge technologies such as direct optical links and the Coherent Accelerator Processor Interface (CAPI), in order to maximize the performance of the cloud infrastructure;
- Designing a scalable and standardized data center-in-a-box, i.e., a small form factor enclosure that is capable of supporting up to hundreds of processing boards (also known as microservers);
- Defining and integrating appropriate metrics with a more holistic approach to continuously monitor the efficiency of the cloud computing platform;

Being part of the cloud infrastructure as edge nodes, Cyber-Physical Systems used in video surveillance applications will be made more energy-aware by achieving the following objectives:

- Exploiting ultra-low power manycore processors with a dedicated energy-aware instruction set architecture, in order to speedup computer vision algorithms;
- Leveraging on cutting-edge technologies such as processing acceleration through specialized functions;
- Exploiting reconfiguration and adaptability of the radio communication subsystem to enhance the CPS efficiency;

Succeeding in all these objectives requires a common working ground among different research areas. OPERA tackles this challenge by bridging extremely specialized skills such as designing System-on-Chip, programming highly performing FPGA accelerators, optimizing the workloads among multicore chips, integrating different technologies under stringent low power constraints. OPERA research and innovation results will significantly contribute to the creation of more advanced cloud services, spanning from the CPS support to the HPC-cloud convergence. Furthermore, OPERA results will greatly help to reinforce the Europe position in leading low power computing technologies, as well as to increase its role in growing markets (e.g., HPC-oriented infrastructures, IoT, etc.).

III. The OPERA platform

As stated in the previous sections, OPERA aims at deploying a scalable computing platform, explicitly looking at the integration of two main elements: (*i*) remote high- and energy-efficient computational capabilities; (*ii*) local acquisition and preprocessing capabilities by means of ultra-low power CPS which incorporate sensors, computing resources, and radio-communication interfaces. These two elements are at the basis of the evolution of the whole computing continuum, as depicted in figure 1. At the edge, there are embedded devices and CPS, equipped with small low power processors which provide the necessary computing power and storage space to preprocess data acquired through attached sensors. However, the capabilities of such systems do not allow to maintain a large amount of data for a long time. To this end, multiple flows of data are aggregated by means of a cloudlet [1], i.e., a nearby computational resource represented by a data center-in-a-box with wired/wireless connectivity towards Internet, multicore processors, and a stable power supply source (i.e., it is plugged into a power outlet). For instance, a cloudlet can be a small group of servers that act as the primary gateway for mobile devices to access remote services. Whenever larger capabilities are required (e.g., to perform a complex analysis of a large set of streamed data), private cloud computing infrastructure are exploited, albeit resources are upper bounded. Conversely, public data centers offer an almost-infinite amount of resources that can serve complex and heterogeneous tasks, with different requirements.

The primary objective of the OPERA project is to make more energy efficient all the stages of this chain. To this end, we envision a more scalable cloud platform. The whole platform organization is depicted in figure 2. The platform comprises two main blocks. The first regards the implementation of a scalable low power data center resorting to a mix of advanced hardware technologies and management software. Hardware heterogeneity provides the proper computational power required by complex tasks running on top of diverse frameworks (e.g., Apache Hadoop, Apache Spark, MPI, etc.). To this end, OPERA integrates low power and high-performance processing architectures (ARM, X86_64, and OpenPOWER) into highly dense and interconnected server modules. Enclosures such as the HPE Moonshot [3] and the HPE The Machine [2] make possible integrating hundreds of different processing elements by exploiting a microserver design: a single cartridge contains the specific processing element (i.e., CPU, DSP, GPU, or FPGA), the main memory and the storage. Unlike in single embedded systems, data center workloads offer more opportunity to improve efficiency by adopting specialized hardware structures. In fact, several tasks are intrinsically parallel, thus, they can benefit from being executed on a more parallel hardware, as offered by high-end FPGAs. Examples of such tasks include, but are not limited to, Memcached key-value store systems [4], search engine ranking

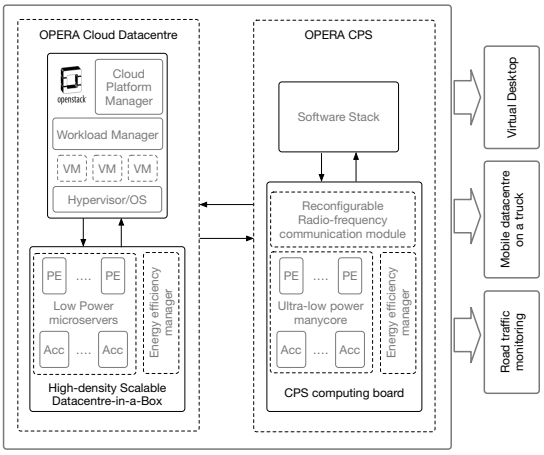

Figure 2. Envisioned OPERA cloud platform with interaction between ultra-low power CPS and scalable, high-density, low power data center nodes.

algorithms [5], and database operations [7]. Besides cloud-specific applications, other algorithms may largely benefit from FPGA acceleration. Among the others, in OPERA we are particularly interested in accelerating image processing algorithms, and network communication functions. In our case, we use the CAPI technology to make the FPGA accelerator easily accessible by the application: the accelerator logic can transparently access the memory hierarchy of the processor that intends to use it. To this purpose, commands and responses for the accelerator are encapsulated within the protocol messages belonging to the interconnection link. Unlike other acceleration platforms, making the external accelerator coherent with the rest of a multicore processor avoids the use of complex performance-limited techniques to transfer data (e.g., GP-GPUs force to transfer programs and data from the main memory of the host processor to the private memory of the accelerator). To support task parallelization, OPERA leverages on two key elements: (*i*) a software components named Workload Manager (W-MGR) that is responsible for decomposing cloud applications into small independent services and mapping them on specific processing elements; (*ii*) Linux containerization (e.g., Docker [10], Linux Containers [9], etc.) to encapsulate these software services into lightweight virtual machines (VMs). The workload manager uses dedicated algorithms for resource provisioning, as well as to configure and selectively switching on-off of the lower building blocks of the infrastructure, such as VMs, I/O devices, memory, multicore processors, HW accelerators, and transmission components, in order to increase the system efficiency. On the other hand, it relies on profiling information to properly split the application into a set of independent tasks and to map them to specific processing elements.

The second key part of the envisioned platform consists of an ultra-low power (ULP) Cyber-Physical System, targeting video surveillance and monitoring applications. In this context, we are planning to manufacture this system around a highly parallel manycore-based System-on-Chip, equipped with hardware acceleration functions specifically designed for computer vision applications. Furthermore, by leveraging on an advanced manufacturing process, the CMOS camera sensor will be placed close to the computing back-end (i.e., the ultra-low power SoC). In addition, to enhance the system capabilities of operating in environments which lack for a stable communication link (e.g., in mountain roads), a reconfigurable radio-frequency communication module will be integrated into the system. At its basis, there is a reconfigurable antenna (i.e., a physical antenna connected to a signal processor) that in general is capable of reconfiguring its sensing frequency, polarization, and radiation pattern. These capabilities enable the radio system to accommodate different communication protocols and handle variable channel conditions.

Both the scalable high-density server and the ultra-low power CPS will integrate HW/SW features intended to monitor the efficiency of the system in terms of used resources and consumed power. This information will be integrated into the algorithms used by the workload manager to map computations with the appropriate processing elements either on the cloud side or on the CPS.

A. Low power heterogeneous data center architecture

Heterogeneity has been recognized as a viable solution to improve performance and reduce power consumption at the same time [12]. At its basis, there are hardware accelerators with their specific programming models, that make them very difficult to be exploited in virtualized environments. Although the full exploitation of such hardware accelerators in cloud computing environments is still very low, OPERA intends to reverse the situation by making FPGAs transparently accessible by cloud applications. To this end, our accelerating boards are wrapped by an optimized Board Support Package (BSP) that deal with the low-level details of the FPGA architecture and peripherals (e.g., PCIe, CAPI, SerDes I/O, SDRAM memory, etc.), and that fits into the high-level OpenCL toolflow [14]. This allows the application (or a portion) to be represented as highly portable C code kernels, while the W-MGR can decide at the last minute if these kernels can be off-loaded from software running on the host processor, to silicon gates which execute quickly, hence efficiently yielding improved performance. Furthermore, our accelerators furnish the high-density server (we refer to this solution as "Data center-in-a-Box") and OpenPOWER nodes with PCIe and CAPI attached programmable logic. With the Coherent Accelerator Processor Interface (CAPI) [18], the attached accelerator appears as a coherent CPU peer over the I/O physical interface. Specifically, it can access to a homogeneous virtual address space spanning the CPU and the accelerator, as well as a hardware-managed caching system. The advantages are clear: a shorter software path length is required compared to the traditional I/O model (i.e., there is less overhead associated with the OS and drivers). On the FPGA side, two hardware sub-blocks are placed: the Power Service Layer (PSL) which is an IBM proprietary module, and the Accelerator Functional Unit (AFU) itself, i.e., the silicon block implementing the acceleration logic. The PSL contains hardware blocks that maintain cache coherency with the outside world, through a cache and memory management unit. In addition, an Interrupt Source Layer (ISL) is available in order to create an access point to the AFU for the software layer. On the OpenPOWER side, the Coherent Attached Processor Proxy (CAPP) block acts as a gateway for the requests coming from and directed to the external AFU. The accelerator also acts as a high bandwidth, low latency bridge connecting processors hosted on the Data center-in-a-Box (towards a more energy-efficient architecture, we will

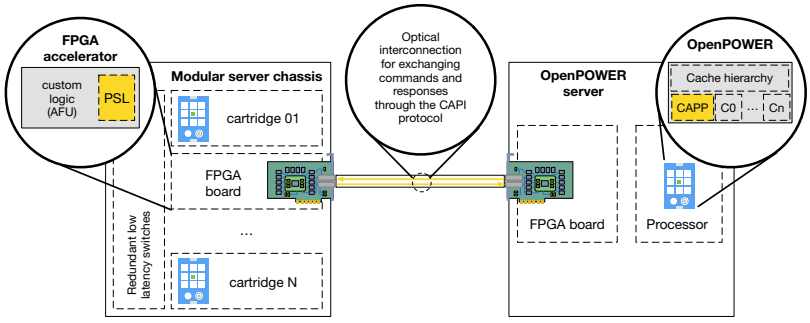

Figure 3. Scalable high-density low power server chassis are accelerated by FPGA hosted cards, also acting as a bridge towards high-performance OpenPOWER-based nodes thanks to the CAPI interface.

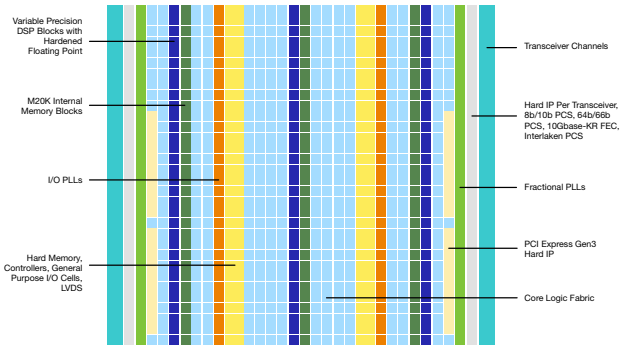

Figure 4. The internal organization of the Altera Arria 10 FPGA device [13].

explore more the ARM landscape) with the OpenPOWER platform. Specifically, this link will be implemented using optical interconnects operating a Serial Lite protocol that minimizes the overhead of transporting vast amounts of data between platforms. In particular, we resort to standard Quad Small Form-factor Pluggable – QSFP – modules that permit up to 40 Gb/s of bandwidth physically configured either as a single 40 Gb/s link or split into four independent 10 Gb/s links using an optical splitter cable. This unprecedented flexibility allows for a wide range of topologies to be supported without the high cost and delay of re-designing new PCBs. Thus, all these features allow dramatic scalability and tight coupling of very different processor technologies within a single computing platform. In fact, interconnecting X86_64 and ARM processor with OpenPOWER systems using conventional Ethernet or Infiniband cards would adversely impact application performance and overall system efficiency. Figure 3 shows the whole data center organization, as proposed by OPERA: scalable high-density and low power nodes take advantage of the acceleration capabilities offered by FPGA devices. Such devices connect, through fast optical links, OpenPOWER nodes, which are made coherent with reconfigurable logic through the CAPI interface.

At the basis of the OPERA accelerator, there is the Altera Arria 10 SoC, which is a high-performance reconfigurable device directly supporting the IEEE-754 Floating Point arithmetic through newly designed DSP blocks. Figure 4 shows the internal organization of the Altera Arria 10 FPGA device [13], where different reconfigurable logic blocks (e.g., PLLs, DSPs, Core-logic fabric, etc.) are highlighted. Such SoC also features a second-generation dual-core ARM Cortex-A9 MPCore processor, which is integrated into the hard processor system (HPS) in order to obtain more performance and security with respect to the previous generation or equivalent soft-cores. The HPS partially substitutes on-die reconfigurable logic but provides more flexibility to the platform. Each Cortex-A9 core is fully compatible with the ARMv7 instruction set architecture and supports vector operations through the NEON extensions. Cores can be clocked at 1.2 GHz, and present a superscalar, variable length, out-of-order pipeline with dynamic branch prediction. The cache memory totals almost 1 MiB, and it is distributed in 32 KiB I-cache + 32 KiB D-cache (L1), 512 KiB L2 cache (shared, 8-way, set associative), 256 KiB on-chip scratchpad, and 64 KiB for the on-chip ROM. The internal bridge allows transparent communication between the reconfigurable logic and the HPS.

As previously mentioned, making easy to program such complex devices is a challenge. To address this issue, in OPERA we aim at providing accelerator boards with easy-to-use OpenCL BSP. Such software package allows leveraging on the FPGA device, as well as all the high-performance on-board peripherals. The hardware design issues and constraints are abstracted away and automatically handled by the Altera OpenCL compiler, leaving the software programmer to deal only with specific algorithms of interest. The compiler allows optimizing the C-based code enabling OpenCL channels (an OpenCL language construct) to be used for kernel-to-kernel or IO-to-kernel high bandwidth, low latency data transfers. Channels are also used to implement an application program interface (API) intended for the host to communicate with the hardware accelerator generally mapping the PCI Express interconnect, or kernels to communicate one another without host intervention. To speedup the development of kernel applications, the compilation flow is based on a software-based debug and optimization cycle, which limits the need for FPGA HDL compilation only when most of the code has been optimized.

B. *Workload decomposition*

The deployment of a set of interconnected tasks in a heterogeneous context is critical since different processing architectures are touched. To take full advantage of hardware specialization, a mechanism to automatically assign tasks must be put in place. To this purpose, we propose to adopt the *microservice* model. It has recently emerged in the cloud computing community as a development style which allows

building applications composed by several small independent but interconnected modules [15], [16], each running its own processes and communicating with others by means of a lightweight mechanism (typically consisting of an HTTP-based REST API [17]). The result is an asynchronous, shared-nothing, highly scalable software architecture that avoids monstrous and difficult to maintain monolithic code.

Given this premise, OPERA aims at exploiting such architectural pattern in the following way: (*i*) the application is partitioned into independent tasks; (*ii*) each task is characterized (through profiling tools) in order to ease the mapping with the available hardware; (*iii*) tasks are wrapped by thin modules that export the HTTP-based REST communication interface; (*iv*) communication among the microservices takes place by means of the exported API. To make this process effective a mechanism to describe the microservices in terms of their relationships and interaction with other microservices and cloud infrastructure layers is needed. On the other hand, to allow energy-efficiency algorithms to properly work, feedback information coming from the servers executing the microservice instances should be collected.

The first issue can be solved by describing different modules composing the microservices through an application descriptor, which allows the abstraction of different components of cloud applications, as well as describes their relationships and interconnections. This artifact enables the portability of cloud applications and services across different platforms. Among the various standards, OASIS TOSCA [19] is one of the most complete, and it has recently been extended to support Linux containers. TOSCA is an XML-based language and metamodel which can be used to describe composite cloud-based applications in a modular and portable fashion. A TOSCA file consists of two parts: (*i*) a topology template – a graph in which typed nodes represent service's components and typed relationships connect nodes following a specific topology; and (*ii*) plans – workflows used to describe managing concerns. Figure 5 shows an example of a service template. Nodes, as well as relationships, are described by appropriate type descriptors. Each type defines the list of capabilities that the particular module can serve, and the list of requirements necessary to deliver its service, along with the set of properties and the interface associated with it. In OPERA, we will enrich the capabilities in order to use them to describe possible processing architectures that are suited for the execution of the specific microservice, as well as to express the expected performance and power consumption. Similarly, plans will specify an execution order for the set of microservices, along with their possible instantiation on the specific processing elements.

Mapping microservices with the available hardware requires the ability of scheduling microservices depending their specific requirements (e.g., minimum performance level) and policies for maximizing energy efficiency of the whole infrastructure. The workload manager (W-MGR) is the software component devoted to this purpose. It receives the TOSCA descriptor and, after its compilation, it provides a schedule. It performs this operation satisfying the following criteria:

- Must be able to monitor energy/power efficiency of each node;
- Must be able to quickly react to the changes in workload demand;

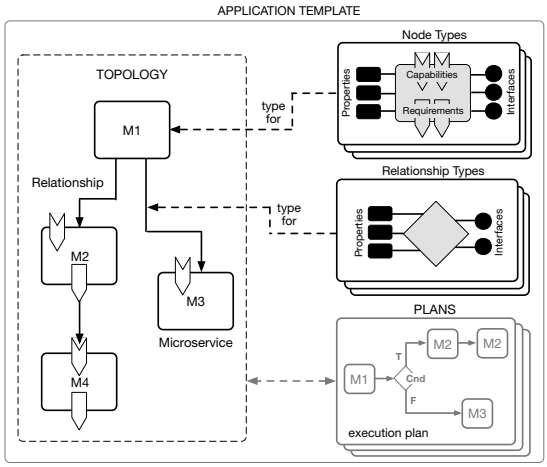

Figure 5. An example of TOSCA template to describe a cloud application as a set of interconnected microservices.

- Must be aware of each node's capabilities and specific features.

The first criteria can be easily met since modern hardware contains embedded sensors to keep track of operational parameters (e.g., processor temperature, fans speed, etc.), as well as several tools are available to monitor the load of the nodes. The second criteria can be satisfied by leveraging the capability of the underlying infrastructure to scale up and down the allocated resources. Finally, the third criteria is simply satisfied by adopting a standardized way of describing application requirements and capabilities of each node (TOSCA standard).

C. Ultra-low power embedded system

Among the various applications of CPS, video surveillance is one of the most interesting. In OPERA, intelligent cameras will be used in the smart city context as end-nodes of a cloud infrastructure, to monitor urban traffic aiming at recognizing potential situations of risk. Such kind of application covers multidisciplinary fields related to computer vision, pattern recognition, signal processing, and communication. When the task has to be performed in a time-constrained manner, the situation is exacerbated. Satisfying such computational demand requires implementing advanced hardware systems, albeit generally with a low energy efficiency. From this viewpoint, integration of functionalities in the form of hardware modules is considered as a technological key feature to increase CPS efficiency. An image sensor placed closely to a computing layer will form such a device. Our state of the art device will be composed of an image sensor (capturing static images or videos) and a computing layer in charge of pre-processing frames, applying filters, and executing other manipulation algorithms.

Another source of potential efficiency improvement is given by the high degree of parallelism that image processing exposes. In fact, processing functions are applied to (groups of) pixels in parallel. For this reason, our design adopts a highly-parallel architecture based on a ULP manycore solution designed to operate with very low supply voltage and currents. For instance, we can use a manycore solution equipped with energy efficient cores (EE-cores) operating near 1.0V, requiring

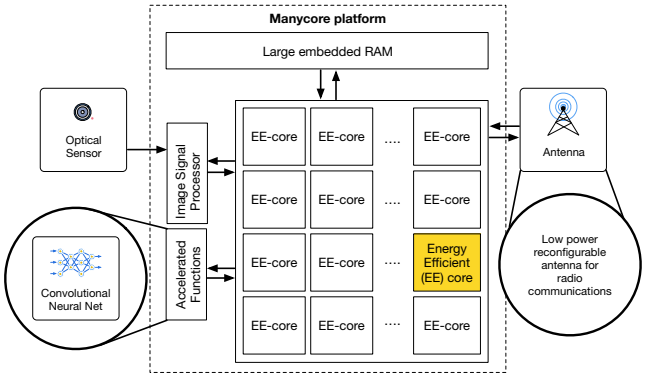

Figure 6. OPERA Cyber-Physical System architecture: an ultra-low power manycore processor with acceleration function for CNNs is directly attached to the camera sensor and to the reconfigurable communication antenna.

only a few tens of μA and sporting several hardware acceleration functions. With such features, the envisioned computing layer can perform operations only requiring few pJ of energy. A dedicated image processing unit allows performing complex operations, such as moving object detection and image/video compression, at a low energy consumption when compared with standard embedded platforms. In addition, OPERA aims at further improving the performance/power ratio by integrating HW/SW components to accelerate convolutional neural networks (CNNs). The approach of CNNs will allow, with a limited increase in the used resources, to improve the identification of classes of objects on the scene (this task is the basis for any video monitoring application). Aiming at optimizing this platform, we will leverage on the integration of imaging sensors and processing cores, which are typically realized with different manufacturing technologies. In addition, we improve the overall efficiency of such CPS by introducing two advancements in the design of the computing layer: (*i*) the use of an energy-optimized instruction set architecture, and (*ii*) the adoption of an advanced manufacturing technology (e.g., the FD-SOI technology) that will allow the extreme miniaturization of the chip with a significant reduction of the power consumption with respect to the traditional bulk silicon devices. The complete CPS design envisages a low power reconfigurable radio-frequency (RF) communication interface. This kind of communication interface is of particular interest for video surveillance and monitoring applications where the environmental context is particularly critical, such as in the case of mountain roads where connectivity is not reliable. To deal with this issue, our reconfigurable RF module will be capable of adapting the transmission to the best channel/protocol features. Since RF transmission is generally power-hungry, we will design the RF interface with very low power components.

IV. Use cases

OPERA intends to provide technology demonstration on energy efficiency, scalability, and computational performance resorting to three use cases based on the following real-world applications.

A. Virtual desktop infrastructure

The purpose of this use case is to demonstrate the capability of the OPERA platform to scale, while keeping energy efficiency under control. The idea is to leverage on the cloud concept of as-a-service to provide a virtual desktop infrastructure (VDI), that is remotely accessible by means of low power thin-clients. Users are increasingly demanding access to their applications and data anywhere and from any device. The rapid growth of knowledge workers who roam from one computer to another within the office had led companies to provide access to the users desktop experience at any computer in the workplace, effectively detaching the user from the physical machine. Through virtualization, employees can access their applications and data very safely over a network and the risk of data loss is minimized, while IT departments can save costs by consolidating services on physical resources.

Due to the high number of systems concurrently connected to the nodes of the data center, we expect a high CPU/storage load. In fact, each connected client may run any typical desktop application (possibly more than one at the same time) expecting a very low latency in the response time. To address this challenge, OPERA will implement a solution based on the open-source framework OpenStack. As far as storage is concerned, OpenStack Cinder, along with the Ceph file system, could cover block storage needs for virtual machines. Ceph exploits commodity hardware and some of its components (e.g., Mon – host monitor) are themselves CPU intensive. Given the low-latency requirements of this use case, network management must be concerned. To this end, we want to exploit the flexibility furnished by the OpenStack Neutron module, to completely adhere to a "software-defined" paradigm. In addition, network latency will be kept low by leveraging on a more powerful remote desktop protocol w.r.t. the traditional protocols (e.g., VNC, RDP, etc.). Finally, to keep as low as possible the overhead of the software virtualization layer, a mechanism based on the KVM hypervisor and containerization will be put in place to run lightweight virtual machines on low power servers.

B. Mobile data center on a truck

OPERA intends to deliver mobile IT services for the Italian agency called Protezione Civile. IT services, such as forecasting and risk prevention, contrast and overcome emergencies, and mitigation of risks, will be delivered through a truck (operated by partner CSI Piemonte) equipped with electronic instruments which allow: (*i*) creating a satellite communication link; (*ii*) acquiring images and videos of the truck surrounding areas; and (*iii*) processing, temporary archiving and transferring of acquired data (videos and images). Images and videos will be acquired by means of a drone equipped with an ultra-low power computing module, which is in charge of controlling the flight of the drone itself, as well as to stream the data to the base station on the truck. The use of a drone is helpful also in the case of dangerous or difficult access to the target site. Once acquired, images and videos are then processed on the truck, which will be equipped for this purpose with the scalable high-density low power server described in section III-A. Data processing on truck consists of these two steps: the arrangement of the videos by deleting not useful parts, adding comments, etc; and the creation of orthoimages for their subsequent comparison with others archived. For instance, 20 minutes of flight yield 300 photos that require approximately 15 hours to be processed with a standard X86_64 machine for a resolution of 1.0 cm. Moving from a standard computer to a high-density but low

power server equipped with FPGA accelerators will enable OPERA to greatly speedup the processing task while keeping low the impact on the power source of the truck (currently, a gasoline power generator).

C. Road traffic monitoring

One of the contexts where OPERA foresees a growing interest is the application of CPS for traffic monitoring in urban and rural contexts. Deploying ultra-low power CPS equipped with video processing capabilities and a wireless communication interface makes possible to monitor large geographic areas. For instance, it becomes possible to quickly detect accidents or any situation of risk and communicate alerts to vehicles (future VANET standard). To this end, collected data are transferred and further processed into low power servers located in remote data centers for proactive actions intended to reduce such risk situations. OPERA will also explore emerging cooperative environments (e.g., Car2Infrastructure technologies).

V. Related work

The OPERA project will exploit many different technologies with the ambition of integrating them into a more efficient platform serving as the basis for creating the next generation cloud infrastructures. In this section, we discuss some works that have been of interest for the development of this project and provided the necessary inspiration and vision to develop ideas behind the design of our scalable data center and ultra-low power CPS.

Heterogeneous architectures have been popularized in HPC to accelerate computations, however, they still remain not common in the cloud computing domain [12]. Although cloud services (e.g., Amazon AWS, Microsoft Azure) offer now cloud instances running on powerful GPUs, the approach with FPGAs is less easy since the difficulties in programming such devices. Some works tried to reverse the situation, explicitly targeting cloud workloads [4], [5], [6], [7], [8]. Both industry and academic community look at the adoption of reconfigurable devices as a way to reduce the impact of computing equipment on the power consumption of the data center. Although from an architectural standpoint accelerators are able to process more data compared to general purpose CPUs, from a technology viewpoint they still make use of power reduction techniques such as dynamic voltage and frequency scaling (DVFS), adaptive voltage and frequency scaling (AVFS), and partial reconfiguration for FPGAs [22] to keep power consumption under control. Such techniques must be correctly integrated into the whole HW/SW stack to really getting advantages on the power consumption. OPERA intends to introduce such kind of heterogeneity in scalable, modular, high-density server enclosures, while keeping away the burden of programming the FPGAs at low-level [14].

CPS are today largely deployed in urban contexts as cloud-connected smart sensors. Often, they are demanded for computationally intensive tasks to avoid large data transfer to the cloud back-end. Recently PULPino [21] has been proposed as an ultra-low power architecture targeting IoT domain. Similarly, ReISC core [20] architecture offers high-performance within an ultra-low power envelope, thanks to a 32 bit RISC-based 5-stage pipeline supporting SIMD operations, a set of high-performance DSPs, and an energy-aware ISA (it allows to improve the efficiency of the system during idle modes). Since communication capabilities (both wired and wireless) are largely exploited in CPS deployed in unmanned contexts, having a reconfiguration antenna capable of adapting to the physical channel characteristics [11] greatly help the deployment of such systems in rural areas.

Modularization of cloud applications ([15], [16], [17], [19]) is emerging as the next architectural pattern to develop highly scalable cloud services. Similarly, CNNs are considered an efficient and yet powerful model to classify objects within an image, as well as to analyze even more complex patterns such as human action detection [23], [24]. However, despite their effectiveness, integration within an efficient CPS architecture is still lacking. Here, is where the OPERA aims at providing innovation.

VI. Conclusions and future work

In this paper, we have presented hardware and software layers that we are developing in the OPERA H2020 European Project. With this project we aim at improving the energy-efficiency of current cloud computing infrastructures by two order of magnitude when compared with current state-of-the-art systems. To accomplish with this challenging objective, we have envisioned to introduce several advancements on the data center side as well as on the end nodes of the cloud, with particular attention to the integration of (ultra) low power high performance technologies.

OPERA foresees to gain efficiency on the data center side, by proposing a modular, high-density server enclosure equipped with small low power server boards, and accelerator cards. To maximize the efficiency, FPGA devices will be used in the accelerator boards to provide acceleration for specific kernels as well as low latency connectivity towards OpenPOWER nodes. In addition, applications leverage on a modular design (microservices) that allows to better scale on low power heterogeneous components. On the other hand, cloud end-nodes (i.e., CPS) are made less power-hungry by integrating manycore processing elements with camera sensors and reconfigurable wireless communication interfaces. To assess the feasibility of the envisioned platform, OPERA will test its designs on three real-world application, albeit the results carried out in the project are of interest for a broader community.

Acknowledgment

This work is part of the OPERA project, which has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 688386.

References

[1] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," in *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14-23, Oct.-Dec. 2009.

[2] R. Courtland, "Can HPE's "The Machine" deliver?," in *IEEE Spectrum*, vol. 53, no. 1, pp. 34-35, January 2016.

[3] HPE Moonshot. https://www.hpe.com/us/en/servers/moonshot.html

[4] M. Lavasani, H. Angepat and D. Chiou, "An FPGA-based In-Line Accelerator for Memcached," in *IEEE Computer Architecture Letters*, vol. 13, no. 2, July-Dec. 15 2014.

[5] A. Putnam et al., "A reconfigurable fabric for accelerating large-scale datacenter services," *2014 ACM/IEEE 41st International Symposium on Computer Architecture* (ISCA), Minneapolis, MN, 2014.

[6] Shan Yi et al., "FPMR: MapReduce Framework on FPGA", *Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays* (FPGA'10), 2010, Monterey, California, USA, pp.93–102, ACM.

[7] A. Becher, F. Bauer, D. Ziener and J. Teich, "Energy-aware SQL query acceleration through FPGA-based dynamic partial reconfiguration," *2014 24th International Conference on Field Programmable Logic and Applications* (FPL), Munich, 2014.

[8] Naoki Tanida, Mary Inaba, Kei Hiraki, Takeshi Yoshino, "Hardware Accelerator for Full-Text Search (HAFTS) with Succinct Data Structure", *International Conference on Reconfigurable Computing and FPGAs* (RECONFIG), 2009.

[9] LXC-Linux Containers. http://linuxcontainers.org

[10] Merkel D., "Docker: Lightweight Linux Containers for Consistent Development and Deployment, *Linux Journal*, March, 2014.

[11] S. Ciccia et al., "Reconfigurable antenna system for wireless applications," *Research and Technologies for Society and Industry Leveraging a Better Tomorrow* (RTSI), IEEE 1st International Forum on, Turin, 2015, pp. 111-116.

[12] Crago, S.P. and Walters, J.P., "Heterogeneous Cloud Computing: The Way Forward", *IEEE Computer*, vol. 48, no. 1, pp.59–61, 2015.

[13] Altera Arria 10 FPGAs. https://www.altera.com/products/fpga/arria-series/arria-10/overview.html

[14] Khronos Group, "The open standard for parallel programming of heterogeneous systems". https://www.khronos.org/opencl/

[15] Newman S., "Building microservices: designing fine-grained systems", 2015.

[16] Thones J., "Microservices", *IEEE Software*, vol. 32, no. 1, 2015.

[17] R. T. Fielding, "Architectural Styles and the Design of Network-based Software Architectures", *Ph.D Dissertation*, University of California, Irvine, 2000.

[18] J. Stuecheli, B. Blaner, C. R. Johns and M. S. Siegel, "CAPI: A Coherent Accelerator Processor Interface," in *IBM Journal of Research and Development*, vol. 59, no. 1, 2015.

[19] Organization for the Advancement of Structured Information Standards, "OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA)", 2015.

[20] N. Ickes, et al., "A 10 pJ/cycle ultra-low-voltage 32-bit microprocessor system-on-chip", *Proceedings of the ESSCIRC*, Helsinki, 2011.

[21] A. Traber, et al., "PULPino: A small single-core RISC-V SoC", *RISC-V Workshop*, 2016.

[22] Nunez-Yanez, J., "Energy Efficient Reconfigurable Computing with Adaptive Voltage and Logic Scaling", *SIGARCH Comput. Archit. News*, ACM, vol. 42, no. 4, September 2014.

[23] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, Jan. 2013.

[24] J. Jin, et al., "An efficient implementation of deep convolutional neural networks on a mobile coprocessor", *International Midwest Symposium on Circuits and Systems* (MWSCAS), 2014.