LOw Power Heterogeneous Architecture
for NExt Generation of SmaRt Infrastructure and Platforms
in Industrial and Societal Applications

Exploitation Plan

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D8.2
DELIVERABLE TITLE	Exploitation Plan
RESPONSIBLE AUTHOR	Giulio URLINI (STMicroelectronics)

GRANT AGREEMENT N.	688386
PROJECT REF. NO	H2020- 688386
PROJECT ACRONYM	OPERA
PROJECT FULL NAME	LOw Power Heterogeneous Architecture for Next Generation of SmaRt Infrastructure and Platform in Industrial and Societal Applications
STARTING DATE (DUR.)	01/12/2015
ENDING DATE	30/11/2018
PROJECT WEBSITE	www.operaproject.eu

WORKPACKAGE N. | TITLE	WP8 | Exploitation, Impact, dissemination
WORKPACKAGE LEADER	TESEO
DELIVERABLE N. | TITLE	D8.2 | Exploitation Plan
RESPONSIBLE AUTHOR	Giulio URLINI (STM)
DATE OF DELIVERY (CONTRACTUAL)	30/11/2018 (M36)
DATE OF DELIVERY (SUBMITTED)	30/11/2018 (M36)
VERSION | STATUS	V1.0
NATURE	R(Report)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Giulio URLINI (STM);

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	ToC	07/08/2018	Giulio URLINI (ST)
0.2	Review of the ToC with partners and assignment of tasks	25/09/2018	Giulio URLINI (ST); Florin APOPEI (TESEO)
0.3	Integration of first contributions from partners	08/11/2018	Giulio URLINI (ST);
0.4	Integration of description of results related to the exploitation plans and other contributions	16/11/2018	Giulio URLINI (ST)
0.5	Integrated further contributions from the partners	19/11/2018	Giulio URLINI (ST)
0.6	Final version ready for review	28/11/2019	Giulio URLINI (ST)
1.0	Final version ready for submission	30/11/2018	Giulio URLINI (ST)

PARTICIPANTS		CONTACT
STMICROELECTRONICS SRL		Giulio Urlini Email: Giulio.urlini@st.com
IBM ISRAEL SCIENCE AND TECHNOLOGY LTD		Joel Nider Email: joeln@il.ibm.com
HEWLETT PACKARD CENTRE DE COMPETENCES (FRANCE)		Cristian Gruia Email: cristian.gruia@hpe.com
NALLATECH LTD		Craig Petrie Email: c.petrie@molex.com
ISTITUTO SUPERIORE MARIO BOELLA		Olivier Terzo Email: terzo@ismb.it
TECHNION ISRAEL INSTITUTE OF TECHNOLOGY		Dan Tsafrir Email: dan@cs.technion.ac.il
CSI PIEMONTE		Vittorio Vallero Email: vittorio.vallero@csi.it
NEAVIA TECHNOLOGIES		Stéphane Gervais Email: s.gervais@lacroix.fr
CERIOS GREEN BV		Frank Verhagen Email: frank.verhagen@certios.nl
TESEO SPA		Stefano Serra Email: stefano.serra@eiffage.com
DEPARTEMENT DE L'ISERE		Olivier Latouille Email: olivier.latouille@isere.fr

ACRONYMS LIST

Acronym	Description
ADAS	Advanced Driver Assistance Systems
BSP	Board Support Package
CAPI	Coherent Accelerator Processor Interface
CE	Chaos Engineering
CNN	Convolutional Neural Network
COCO	Common Objects in COntext
CPS	Cyber-Physical System
CPU	Central Processing Unit
DiaB	Data centre in a Box
DSP	Digital Signal Processing
ES	Evolutionary Strategy
FPGA	Field Programmable Gate Array
HPC	High Performance Computing
HPE	Hewlett Packard Enterprise
GPU	Graphical Processing Unit
IoT	Internet of Things
ISA	Instruction Set Architecture
LC	Linux Container
RDMA	Remote Direct Memory Access
RDS	Remote Desktop Services
SaaS	Software-as-a-Service
SDR	Software Defined Radio
SOC	System On Chip
SOTA	State Of The Art
ULP	Ultra-Low Power
VDI	Virtual Desktop Infrastructure
VMS	Variable Message Sign

LIST OF FIGURES

EXECUTIVE SUMMARY

This deliverable represents the last report on the exploitation activities. The previous deliverables on this topic tried to sketch the potential activities for the exploitation, at the light of the partial results obtained during the project.

At the end of the project, with the complete implementation of the use cases and the evaluation of the features implemented in the project against the requirements, we can define in which directions we can drive the research, on a side, and the production on the other.

The following document contain the evaluation of the project exploitation strategy, reporting the activities on IPR management.

The results of the activities focused on the use cases are listed with a special attention to the activities that can be extended in the future.

In the next section the individual exploitation plan of the partners is presented.

Finally, the evaluation of the impact of the project results is defined. A more detailed analysis of the innovation potential is presented in D2.8.

TABLE OF CONTENTS

1 PROJECT EXPLOITATION STRATEGY

The OPERA project has reached some of the results claimed at the beginning of the project, and modified the direction of the research due to the intermediate results and the unavailability of technologies not foreseen at the beginning of the project.

Several issues have been encountered with the prototypes based on non-mature technologies, like in the case of the ULP platform based on CNN accelerators.

The consortium has overcome most of these difficulties, but still some improvements are possible and new researches can continue starting from the project results. In this deliverable the actions that are foreseen by the partners regarding the continuation of research activities are presented. In this first chapter the main topics that can rule these progresses are presented, from the IPR management, that has been defined in collaboration with the European Commission, that provided a specific support for this topic, and with the definition of patents that will guarantee the intellectual property of some partners with the possibility to continue the joined collaborations.

1.1 IPR MANAGEMENT

A devoted webinar has been organized with the IPR helpdesk dedicated to the H2020 projects (www.iprhelpdesk.eu). The webinar has been followed by the project coordinator (STMicroelectronics), the technical coordinator (ISMB) and the WP8 responsible (TESEO).

The main outcome of the webinar was to highlight the possible barriers for the access to the market, like patents or similar, that could require licenses or royalties.

It has been underlined the potential in term of TRL of OPERA. One successful aspect of projects in these terms has been the application by IBM of the patent described in the following chapter.

1.1.1 Patents

IBM has filed for a patent entitled: *Method for remote page faults over RDMA*

BACKGROUND: When migrating a process (or container) using the post-copy method, most of the memory is left on the source machine and is not copied immediately to the target machine. When the process resumes on the target machine, it tries to access missing memory pages that have not yet been copied. This causes a page fault at the target machine, which can be serviced by copying a memory page over RDMA from the remote machine. The latency (speed) at which the page is copied is very important, because the process is stuck while waiting for the page.

PROBLEM: Using RDMA to copy the memory means the source machine doesn't know when the memory will be accessed, and in which order. That means basically it has to pin all the memory so when the NIC tries to read it, the read will succeed. Pinning slows down the operation of the machine for various reasons, and we would like to avoid pinning. Also, RDMA has a limited range of memory addresses, and can only pin up to 2GB in a memory region, which may also be a problem for large workloads.

SOLUTION: We can use technology such as CAPI or ODP which both allow the same thing - direct access by the NIC to the virtual memory of a process. Now the memory doesn't have to be pinned, because the NIC may cause a page fault on the source machine, which the OS will be able to handle. Also, both technologies allow access to the entire virtual address space, which means we are no longer limited to a 2GB memory region and can access the sparse address space of the process directly. Both of these reasons together improve the latency of moving the memory page to the target machine.

1.2 COLLABORATION WITH THE HETEROGENEITY ALLIANCE

The alliance is a formal association (non-profit & non-legal) formed of different organizations and financed project consortiums that are managed by a governance structure, and will pursue a common objective: influence the Heterogeneous Hardware and software markets. The alliance will focus on all phases of applications for H-HW&SW to enable a new wave of development and execution tools for next-generation applications:

- from design time,
- to enhanced execution,
- to parallel programming,
- and to an optimized runtime in a number of dimensions (energy, performance, data locality and security among others)

1.2.1 Heterogeneity Alliance Goal

Heterogeneity Alliance goal is to join efforts of organizations interested in the development of future technologies and tools to advance, and take full advantage, of computing and applications using heterogeneous hardware (H-HW&SW). Its main goal is to create an association in which anyone interested in the following technological areas can collaborate pursuing a common objective:

- founding an association with a common purpose of interaction, discussion and collaboration with other entities involved in any technological field related with the full life-cycle of H-HW&SW ,
- supporting the creation of a common and an extendable set of technologies and tools (assets) around development for heterogeneous hardware, shared as open-source, which can be improved for mass adoption utilizing technologies,
- in which all members will collaboratively take charge of promoting, supporting, and enhancing the assets shared by the Alliance,
- and benefiting from the synergies, opportunities, visibility and relevance that the community around H-HW&SW can generate for their own objectives”

OPERA: Share results from the OPERA project which main achievements are on Computing continuum, Low Power Scalable Form Factor Data Center, workload decomposition, Ultra Low Power Devices and FPGA acceleration for CNN.

Contribution will be on participating in the alliance to share ideas for technologies improvement on heterogeneous architecture, workload decomposition and resources allocation on Heterogeneous Low Power infrastructure, Computing Continuum architecture, Offloading.

2 RESULTS OF THE PROJECT DEMONSTRATED IN THE USE CASES

2.1 USE CASE 1 – TRAFFIC MONITORING

The results obtained in the demonstration produced some results that can be the starting point for an optimization and further research in the field, especially for the neural networks applications that increases the liability of the system.

STMicroelectronics is designing new SoCs based on the technology employed in OPERA thanks also to the potentiality of application of these technologies in real life contexts.

From the point of view of TESEO the solution of the Traffic Monitoring use case could be a real improvement because of the following advantages:

- Power Consumption: the box used in test site 1 and 2 is consuming a concrete amount of energy less than the other products on the market
- Infrastructure: the solution proposed in OPERA is smarter than the technology used today because of the less infrastructure used on the pole and the connectivity
- Maintenance costs: thanks to the technology used the OPERA solution is suitable to concretely reduce the maintenance costs of the entire solution

Neavia brought its know-how to OPERA project, about development of image processing and video analysis for road management applications. Thanks to the specificity of OPERA, Neavia could work on innovative and state-of-the-art hardware and architectures. On one side, regularly asked use cases like congestion and wrong way vehicle detection were developed using a “classical” image processing approach on a limited board, but having accelerated operations available; both partially close to some Neavia systems or concepts, and challenging since completely new. On the other side, the bicycle counting use case, leveraging on a shared and heterogeneous architecture in which the sensor on the field can directly embed AI on Edge, filtering specific events which can be pushed to a server for further CNN-based processing, without overloading it; proved the interest for both road management systems builders and road managers in integrating CNN in such systems, though still in beginning of development for now. This was challenging too as both the sensor and server are linked together and both use a CNN network as the detection layer in the application processes. CNN had not yet be used in Neavia till now but are of an evident interest for the future, and OPERA gave an insight of CNNs, on the conceptual perspective and the technical point of view. Each system was linked to the network, and so the Road Management Centre, through a new low-power connection mean enabling the deployment of video solutions even in isolated environment, whether regarding network or energy.

All of them offer new opportunities and open a new perspective about road management use cases, integrating HW-accelerated, energy efficient and performant architectures.

The point of view of the Department de L’Isere, the end user of this use case, is the following:

- The OPERA project showed the assets of the ULP autonomous platform, to implement innovative strategies of the road traffic management. The ULP autonomous platforms provide embedded smartness which will enable to collect more efficiently road information.
    - The video data can be permanently collected in many places thanks to the ultra-low power technologies : the experimentation described in D7.3 shows that OPERA technologies significantly reduce the energy consumption (shown on the test site 1) and can operate with a relatively light and compact energy autonomous system (shown on the test site 2): It makes possible to equip even more places, without requiring connection to wired energy and communication network, and without requiring heavy civil engineering infrastructure.
    - Traffic conditions and road events can be detected in many places thanks to the ultra-low power technologies: the experimentation described in D7.3 shows that OPERA

system provides the required smartness to detect road traffic conditions as start and end of road congestion, or to detect road events as the presence of wrong way vehicles.

- This ULP autonomous platform can contribute to really give better real time information to the road management center and the road user, as shown by the road detection use case.
    - The demonstration of the use case “detection of road congestion” showed how the OPERA technologies can be used to improve the traffic conditions monitoring and to produce finer and more accurate real time road traffic information. It will enable to deliver better information, both to the road users and the road operators.
    - The demonstration of the use case “detection of wrong way vehicle” showed how the OPERA technologies can be used to detect road events and to improve the safety of road users.

Based on these results, two innovation actions can be considered:

- The development of a mature product (not only a prototype) to detect in real time road traffic conditions and road events from a ULP autonomous platform that provides easy and cheap conditions of installation and operation.
- The investigation of new use cases supplied by the coupling between the ULP video autonomous platform and high efficiency servers through the off-loading process.

2.1.1 Development of a product to detect in real time road traffic conditions and road events

The results of the OPERA project show the proof of concept and the possible interest of OPERA technologies to collect some road events or some road traffic conditions.

These results can be used as a start point for the development of a more mature product, that will be able to operate more systemically and with more reliability under the complete panel of road environment constraints.

2.1.1.1 Required action for developing a pilot system

During this project, the implementation of the prototype required multiple tests and the development of interfaces between the technological bricks, sometimes done in the field, before the operation at real scale. The global system could be tested only in a few places (one test site per use case) and only during a relatively short period (from a few hours to a few days).

A next step will be expected to test several systems more systematically on a long period and under any road environment conditions. Starting from the OPERA prototype, we expect now a pilot system which can operate more completely under all the real conditions The required improvement is a more reliable system (including more reliable interfaces between the different technological bricks), an improvement hardware packaging to face all types of practical issues (positioning of the optical view, operation under snow conditions, etc…) and a easy-use human machine interface on road management centre.

The general idea is to extend the test at a larger scale to validate the operation of the product: longer test periods, more complete panel of environment conditions, more diversified use cases about road traffic conditions or road traffic events. The return of experiment will enable to optimize both from the hardware and the software devices of the product.

2.1.1.2 Autonomous energy system

According to the objectives of the OPERA project, the energy autonomous system developed in OPERA is only dedicated to demonstrate the ULP characteristics of OPERA technologies. The development and the optimization of a commercial energy autonomous system is expected to make possible a massive deployment of commercial devices across the road networks. Energy harvesting and energy storage technologies need to be optimized, specific mechanical packaging are required. The development of

mature commercial products has to eliminate no-useful energy losses at the interface board level. New opportunities which appear after the project in terms of energy harvesting and storage technologies and/or products can be investigated.

The results have to reduce even more cost, size and weight of the energy autonomous system to make possible a low cost and massive deployment across the road network. A commercial product has to be easily installable on any type of support without requiring expensive and bulky civil engineering infrastructure: easy installation on tipping pole, on pre-existing road signs or street fitting , on light dedicated pole , obviously without requiring any connection to energy or communication wired networks. Considering the successful tests of the OPERA project, such a target seems really realistic.

2.1.1.3 Advanced research for the neural network application

- A more advanced research is expected to increase the reliability of the detection of road events and traffic road conditions. Neural network applications have to ensure the reliable detection under the multiple constraints of the road environment: optical viewing considering the constraints to install the system in the field, diversified meteorological conditions including the most critical conditions in crisis situations (snow, fog, storm, etc…), diversified traffic conditions, etc.
- A more advanced research is expected to increase the panel of road events or road traffic conditions which can be detected: snow on the road, different typologies of car accidents, diversified topologies of natural risks (avalanche, rock fall, floods, etc…)
- The target is to use the results of OPERA project for an application to more use cases under more road environment conditions.

2.1.2 The experimentation of the off-loading process

The experimentation of the cycle detection use case investigated the concept of off-loading process coupling the ULP autonomous platform and high efficiency server. These tests show a new large field of possible opportunities for the road manager. This experimentation opened a new activities field. The ULP autonomous platform is no more only a detector of road event or road condition but the ULP autonomous platform can also enable to build a data base that can be analysed in postponed mode by a high efficiency centralized server. The treatment by a high efficiency server can enable to treat very complex images (as tested for the cycle detection use case), and we can imagine a high efficiency server treating the data collected in a large panel of different sites and/or treating the data collected in a large panel of different times.

Multiple use cases can be investigated considering the analysis of data video collected and transmitted by the ULP autonomous platform from an off-loading process: use cases relative to the better comprehension of the road traffic and the development of predictive tools, use cases relative to the analysis and monitoring of the road infrastructure, optimization of road operation by video data treatment, etc.

Large panel of use cases have to be considered, investigated, and the most relevant use cases have to be defined, tested and evaluated.

2.2 USE CASE 2 – TRUCK

The Protezione Civile organization and the truck operator have a very good opinion about OPERA Project, because now they can provide the same service consuming less energy, less time and less space. For this reason, they are working to set up a new vehicle but smaller than the current one.

EdgeLine 4000 chassis, which is replacing 2 older DL380 servers, has enhanced truck capabilities in multiple aspects:

- Performance: processing time for orthophoto reduced by a factor of 2 versus previous SOTA
- Efficiency: leveraging lower power CPU together with GPGPU enabled a saving of 5x while in processing mode and of 7x while only running RADIO services.
- Resiliency: radio and domino services are now hosted in virtual machines which can be started from any of the m510 located in the EL4000.
- Rack space: from 4U to 1U with NEBS compliancy (shock/vibrations and high temperature range)

These improvements, described in D7.6, can be achieved on other similar use-cases when hardware is integrated into any types of vehicles. Researches done on Autonomous Driving require similar architecture to process live data coming from all sensors (LIDAR, RADAR, cameras).

2.3 USE CASE 3 - VDI

During the project, we set up different configuration involving different hardware and software, many of them unknown to us, and for this reason we gained a great experience, specifically about containers and low power server. Thanks to them, it's possible to decrease a lot the energy consumption. But it's important to highlight that migrate business services form virtual machines to containers requires specific activities:

- to define policies for application architectures with microservices;
- to train people to manage the new technologies;
- to migrate the existing services toward the new configuration. If the current services are not containerised, it's necessary to consider a task for transforming the current services.

This effort depends on the enterprise and the type of applications, but considering that periodically it's necessary to review the applications (due to new business requirements), only the first two task could be considered as overhead. More details about that are available in D2.8.

3 INDIVIDUAL EXPLOITATION PLANS

3.1 STM

The activities of STMicroelectronics has been focused mainly in the development of the Ultra-Low Power device for the traffic monitoring applications considered in the OPERA project, and the integration of these devices in the OPERA framework, with the implementation of the offloading mechanism for the refinement of the detection and counting of elements in the scene (in the use case example implemented, the bicycle counting use case).

In this context STMicroelectronics has employed two embedded platforms for the edge computing, one for traditional video processing algorithms and one for the acceleration of Convolutional Neural Networks.

The first platform, named SecSoC, demonstrated the capabilities to execute video processing algorithms keeping a ULP consumption profile, with performances acceptable for the end user involved. The use of this platform could be the starting brick for the productization of low cost autonomous cameras for traffic monitoring and other applications. Several discussions on this topic have been started between NEAVIA and ST in order to evaluate the possibility to realize a product.

Regarding the more advanced platform for CNN execution, the SoC has demonstrated a high reliability level in terms of quality of detection. The use case adopted for its demonstration wasn't critical for the traffic managers, but can be applied in a variety of applications, that can't be supported by a classic video processing approach, at least with the same quality and reliability level.

The potentiality of this platform will be further exploited in new research projects, with the purpose of increasing the range of application evaluated and the support for third party to use this technology. Several project proposals have been submitted and another H2020 project that will use this technology has been started at the beginning of 2018. The results of OPERA will be considered in the development.

The platform will be industrialized by STMicroelectronics in several products that will be employed when available in large scale pilots' experimentations.

3.2 IBM

IBM has contributed all of the code developed within OPERA as open source. This includes changes to the Linux kernel, CRIU and RDMA libraries. We have initiated contact with organizations involved in the development of all of these projects through conferences such as Linux Plumbers, Open Source Summit and FOSDEM. This has given us exposure to a wide audience of potential adopters of the technology. Several companies have expressed interest, and started to use our contributions. For example, Google recently announced that they are starting to use CRIU, which includes our changes for post-copy migration of containers.

We plan to continue to work with these open source communities and their customers to increase adoption of the technologies that we and the other partners have created during the project. We have filed for a patent on one aspect of the remote page fault over RDMA work, which implies that IBM believes this technology has a business potential. In addition, we expect that our future work will be based in part on our developments in the OPERA project, and thus provides a base for future developments and experimentation.

3.3 HPE

HPE has joined OPERA project for its Moonshot portfolio. Moonshot, which is about providing power efficient servers for given use-cases, has been released 2 years prior OPERA beginning. A complete set of cartridges has been made available, from very low power ARM to higher end Xeon systems. Together with SRC, HPE has also released a FPGA cartridge based on Altera Stratix 4.

https://www.eetimes.com/document.asp?doc_id=1326707

Figure 1: Moonshot Saturn 1 cartridge

Saturn FPGA cartridge didn't meet its market due to multiple reasons: a high acquisition cost, a complex development interface and an old FPGA design.

With the rise of FPGA, HPE has started working more closely on this technology. Instead of developing products from the ground, a new strategy has been defined to leverage existing hardware from selected partners. The integration into existing EdgeLine chassis as well as upcoming product is now part of the priorities in term of qualification, same as what has been done for GPGPU.

HPE is currently working on the new IoT portfolio which is going to be an updated version of Moonshot and EdgeLine chassis. These new compute elements will be highly dependent on some results coming from OPERA, with extensive use of Redfish and no legacy protocol such as IPMI and SNMP.

Taking OPERA experience into consideration, EL8000 chassis will also be engineered in a way to be easily installed on remote site. This dense compute chassis is targeting heterogeneous workloads with both Arria 10 and Stratix 10 support, as PCIe accelerator.

This new platform will be targeting markets, such as 5G, which require deployment of extreme compute capabilities as close as possible from the antennas.

El4000 with m510 and Arria 10 FPGA is currently the platform of choice for high-quality video processing at the edge. eBrisk made a successful demonstration of 4k HDR transcoding with Nallatech FPGA board which initiated a large opportunity at an American broadcaster.

3.4 NALLATECH

3.4.1 CNN

Prior to the OPERA project Nallatech was actively research CNN acceleration on FPGA's. The additional CNN expertise acquired during the OPERA project, particularly the binary optimisation techniques applied, will be further developed for future commercial opportunities and European research projects. Nallatech is a member of the H2020 LEXIS project where FPGA acceleration of CNN type codes will also be applied.

Nallatech intend to start using machine learning IP as part of future FPGA based solutions. This has already included dissemination activities, including white papers and demonstrations as part of an effort to increase our customer base.

3.4.2 NALL: FPGA Device

The FPGA accelerator developed for the OPERA project has some unique capabilities that extend its functionally beyond what was implemented in the OPERA project. Within the OPERA project the FPGA device was used for offloading calculations. However, the device is also particularly well suited for edge computing given its IO and embedded ARM processor.

Target exploitation markets for the device include High Performance Computing, Image Processing and Network Analytics.

To enable the Network Analytics use case Nallatech have developed a host board to allow the FPGA device to be used outside the server environment in a portable small form factor package.

Figure 2: Nallatech FPGA MicroNode

The FPGA MicroNode platform is a low power miniature Linux Edge-computer using the Arria-10 SoC FPGA device developed for OPERA. This integrated platform makes use of the optical network I/O directly coupled to the FPGA fabric enabling ultra-low latency applications, as demonstrated in the OPERA device testing (D6.7). The embedded ARM processor on the ARM device runs a basic Linux stack and is the host in this micro system. This allows on the edge-of network applications to possible that combine both the FPGA and flexibility of the ARM.

3.5 ISMB

ISMB has been actively involved in the design and implementation of an (energy) power-aware Cloud orchestration system, the design and implementation of a directive antenna for efficient wireless communications and in designing the offloading mechanism. Regarding the implementation of the Cloud orchestration solution, ISMB plans to reuse such solution in other H2020 projects, as well as to continue its development. In this context, ISMB will study how to further improve the capability of the orchestrator to manage ever larger data centers and speeding up the execution of the static and dynamic orchestration algorithms. Concerning the design of the directive antenna, ISMB intends to continue improving its performance and adaptation capabilities, as well as to apply such solution in different contexts. As part of the offloading mechanism implementation work, ISMB contributed in designing the Convolutional Neural Network (CNN) that runs on the FPGA-accelerated server. To this end, the OpenCL framework has been used (High Level Synthesis –HLS). Thus, future activities will be two folds: on one hand, the OpenCL-HLS will be exploited for providing acceleration for other applications, ranging from Machine Learning algorithms to Big-Data applications. On the other hand, the developed CNN will be exploited in all the contexts (scientific and industrial applications) where it is required to quickly perform image analysis and object recognition/classification.

3.6 TECHNION

The Technion plans to exploit the OPERA findings and results in two directions.

First, the Technion plans to extend the models developed under WP5 to virtualized setups. The new models will be able to predict the runtime and energy of applications running in cloud setups (i.e.,

where applications are hosted in virtual machines) and not just in bare-metal setups (i.e., where applications are running natively on physical machines). Accurate prediction of cloud performance will hopefully allow cloud vendors to optimize their resource utilization via a smart placement of workloads in their heterogeneous data center. This exploitation activity will hopefully be carried out as a new EU project under ICT-11 topic B called "MAGIC: Massive data Acquisition and analytics for Growth of Innovation in key industrial sectors by Convergence of IoT, edge and cloud technologies" (the project proposal was submitted on November 14th, 2018).

Second, Technion plans to exploit the virtual memory research under WP4 in future research and innovation activities aiming at CPU design. The most important insight from our work is that the current, widely-available x86-64 platforms are inherently less scalable than hashed designs because they use more and more levels of translation as the memory size increases. In fact, in 2017 Intel announced its plans to extend the virtual memory address spaces supported in x86-64 CPUs [1]. The new Intel design will lengthen the page walk length from 4 to 5 memory references, further amplifying the overhead of virtual memory on modern, memory-intensive applications. Our proposed hashed page tables, in contrast, require only a single memory reference for address translation regardless of the address space size. The Technion already contacted two important CPU vendors -- Intel and Mellanox -- to disseminate the OPERA findings and drive future computer architects to use hashed page table in their designing of new computer processors. The Intel Haifa site is the home of the computer architecture group, which, among other things, is responsible for designing the next generations of Intel x86-64 processors. The Technion presented the virtual memory work to this group of engineers, which expressed interest in implementing it. The main concern of Intel was that the Technion proposal requires architectural changes, and Intel has strong legacy constraints of course. The Intel engineers therefore suggested to examine a virtual memory design that combines hashed page tables in tandem with the Intel radix page tables. The Technion didn't have the chance to carry out this research, but two other works have explored this approach and built on the Technion “Hash, Don’t Cache” paper (and cited it, of course). These two works were published in ISCA 2017, the premier forum for computer architecture research [2],[3]. The Technion also presented their work in Mellanox, which, after the acquisition of EZchip in 2016, is building high-performance network processors and multi-core processors. The Mellanox architects were therefore interested in the research results and the lessons learned, and the Technion hopes to collaborate with them on future projects.

Another way of exploiting the Technion’s research is optimizing the hashed page tables implemented in IBM Power platforms. Under OPERA, the Technion studied the Intel Itanium virtual memory design, proposed several ways to optimize it, and showed that our carefully optimized hashed page tables outperform virtual memory implementation with radix, hierarchical page tables in existing x86-64 processors. In future work, Technion hopes to also examine the IBM Power architecture for similar improvements and to exploit these results.

3.7 CSI

3.7.1 Truck Use Case

CSI worked on the set up and measurements of some different configurations in terms of hardware and software with the aim to test the heterogeneous architectures. Thanks to this experience, we appreciate the advantages provided by this type of solution in a very critical context where saving energy, space and time are critical element. In fact, the results were so good that we decided to introduce one of the tested configuration in real-life operations as reported in D7.6. In addition, thanks to low power servers and their small size (EL4000 is 1 rack unit), we are working on set up a small truck for Civil Protection.

3.7.2 VDI Use Case

Through Opera project we set up different way to provide virtual desktop infrastructure and we compared them taking into account the number of end users, power consumption and CPU workload.

Thanks to these results, we decided to investigate more about containers and microservices, specifically we planned for the next year two target:

1. To migrate some infrastructural services (as Web Servers) from virtual machines to containers
2. To define the rules and policies to develop business services using containers

In this way, we can progress in terms of technologies, reducing the effort to manage infrastructural services and at the same time saving energy.

3.8 NEAVIA

Neavia developed the software for the road use cases, both for local sensors and also on the remote server. All experience acquired during the project on new low-power and limited platforms will be used as a starting point, since it can clearly help solving existing or upcoming road management use cases and it matches the specific needs that Neavia focuses on: smart sensors for potentially isolated locations. Indeed, the systems used during the project are less energy consuming than available systems for which a large solar panel is required in order to operate autonomously. Thus, it could be feasible to decrease cost and impact on infrastructure and enable deployment in isolated locations. One could say the same about the communication link that OPERA relied on as isolated places like a railway in a mountain may not have a network access ready-ot-use. Also, CNN had been out for a long time now, its use in road applications is quite recent and even newer for AI on Edge. As lots of people and companies, Neavia think it could help rising robustness and accuracy of road management systems and lead to a smarter and safer road usage. As demonstrated during the OPERA project, embedding CNN on the field is already feasible, though the network needs to be shrunk, preserving energy and speed but providing less accuracy. This is where comes the use of the offloading, completing the process thanks to a more performant network. In the same respects, Neavia identified the shared architecture as a short-term solution to bring AI into its products and upcoming projects. For instance, sensitive existing cameras with “classical” image processing can be used as potential event filters, events which can then be confirmed by a CNN-based application running on a server. Work was already carried out for congestion detection and is ongoing for stopped vehicles aside the road. New use cases and augmented services can also be anticipated such as advanced statistics of speed profiles, counting, road occupation, each of them possibly classified by vehicle type; securing crossroads, detect pedestrians in perilous contexts, or protect animal along forest roads.

3.9 CERTIOS

The exploitation plan of CERTIOS, should be proportionate to the scale of the project, and should contain measures to be implemented both during and after the end of the project. This means that for CERTIOS a consulting company, exploitation is carried out by giving advice to third parties. The most likely to profit from the knowledge CERTIOS has gained during the project, and who are already in the customer base of CERTIOS are data centres. The paid-for consultancy is normally based on advice given, and projects that lead to implementation of measures to improve energy efficiency in the data centre.

Along the lines of the current ways CERTIOS takes up business from its clientele, the innovative OPERA products will be introduced as well. The introduction of the OPERA products will be done in presentations and (pre-)sales meetings with the aim of increasing consultancy value.

CERTIOS is involved in the Data Centre Alliance[1], with an active role in the Board of Commissioners. The DCA helps the UK government institutes and the JRC to set standards, making policies and to endorse the European Code of Conduct for energy efficient Data centres[2]. The DCA dissemination channel is a very effective one for the introduction of innovative OPERA products for the data centre industry.

3.10 TESEO

Teseo is focused mainly in Eiffage Group to promote the results obtained in OPERA project and especially in the Traffic Monitoring use case. Teseo is planning to have another meeting in APRR, an Eiffage enterprise, for showing the result and the functionality of the solution after the 2nd implementation phase in Grenoble.

Another effort planned for Teseo is to promote OPERA project in IoT and Industry 4.0 sectors and also in sectorial Enterprise group as Unione Industriale.

3.11 ISERE

After investigating the results of OPERA project, ISERE considers the deployment of ULP autonomous platform as a major innovation axis for the improvement of road management.

Even if OPERA project shows the interest of OPERA technologies, ISERE cannot deploy OPERA system for the moment because OPERA system is not yet a mature product.

ISERE will investigate how to deploy ULP autonomous platform able to detect and to diagnose in real time road events and road traffic conditions. The use cases are identified, the proof of concept is done, and first results show the interest. A mature commercial product is still expected.

Three opportunities can be considered:

- To buy equivalent commercial system if the advanced market becomes able to offer it.
- To test and to deploy the same type of technologies but implemented in a mature product dedicated to an immediate use at high real scale by end-users. Required technical cooperation and financial support can be provided by innovation projects as European project “public procurement of innovation”. The result will be both an ULP autonomous platform coupled with an ULP radio link.
- To integrate OPERA technologies into the road data collection system SYNCRO (results of the European project SYNCRO), which is currently deployed across the ISERE road network. One target can be to couple data collection from video system and data collection from road sensors and from connected vehicle (done by SYNCRO system). Innovation project could support both the required financial support and the framework for a partnership with industrial partners (as ST and NEAVIA).

ISERE will investigate how to deploy ULP autonomous platforms to collect data which will be analysed in real time or in postponed mode by an off-loading process. The proof of concept was investigated (applied for the use case “cycle detection”). But we still require investigating possible use cases, to test more completely the technical capabilities of such systems. A large diversity of use cases can be imagined about monitoring of road infrastructure (analysis of the state of road infrastructure, etc.) about road exploitation (in particular in case of winter conditions, prevention of natural events, etc.) ISERE will be interested to participate to R&D projects on these topics.

[1] https://www.data-central.org/

[2] https://ec.europa.eu/jrc/en/energy-efficiency/code-conduct/datacentres

4 EVALUATION OF IMPACT

4.1 OFFLOADING

Heterogeneous systems (in terms of hardware components, i.e., CPUs and accelerators) became predominant in the last years, since they offer more performance and reduced power (energy) consumption. Offloading is the capability of a (heterogeneous) computing system to reduce the pressure on the main CPU by moving the execution of tasks on a more specialised hardware device. From a pure programming perspective offloading is easily enabled by modern programming models, such as OpenCL and CUDA; however, offloading generally concerns closely tight systems (e.g., accessing to a GPU within the same server node, or within a cluster). Extending such mechanism to the broader domain of remote connected devices open the door to new applications. While, the mechanism has been demonstrated in the context of the traffic monitoring use case using FPGA accelerators, the OPERA approach can be extended to the use of any other accelerator type.

In this context, the impact of the OPERA solution is of worth. Indeed, small connected devices are generally equipped with low-power processors providing limited computing resources, although they are demanded for ever more complex tasks. The offloading mechanism, as devised and implemented in OPERA, provides an effective solution for leveraging on powerful accelerators (GPUs, FPGAs, etc.) available on modern Cloud data centers, and to save energy. On large network of IoT sensors it provides a way to effectively tackle the problem of processing massive amount of data, without negatively impacting on the energy consumption of the IoT device. On the other hand, the (remote) offloading mechanism enables small connected devices to exploit dedicated computational resources thus virtually enlarging their computing capabilities, by distributing computational heavy tasks on the accelerators, thus allowing more accurate data analysis. The reader can recognize such situation represented by scenarios such as large networks of sensors for the environmental monitoring, as well as video surveillance applications.

4.2 ULP AUTONOMOUS PLATFORM

The OPERA project showed to the road user the interest of OPERA ULP autonomous platform to detect in real time road traffic conditions and events. The experimentation of the OPERA prototype convinced ISERE to investigate how to test pilot system and how to acquire commercial products based on the OPERA results. Such a system should improve the capacity to characterize traffic conditions aiming at delivering finer and better information to road users and to the road exploitation team. But ISERE still expect a mature commercial product.

If a commercial product based on these technologies is offered, road managers can characterize the traffic conditions on a large number of sites: the road manager can easily and cheaply deploy the ULP autonomous platform in many places (compact and cheap energy autonomous system) and the road manager can easily diagnose traffic conditions (by using the embedded smartness of platform that delivers information on traffic conditions to the road management centre). The road manager can deliver very finer and more accurate real time traffic information to road users, and to the road operation team. One key impact is to contribute to reduce traffic congestions and their impact as the reduction of accident risks, the reduction of travel time for road users, and the reduction of air pollution.

If a commercial product based on these technologies is offered, road managers can automatically detect road events on a large number of sites: we can imagine that the example of wrong way vehicle can be opened to other types of events: car accidents at critical places as in high issues crossroads, natural events in critical places as flood stream, avalanche, rock fall, malicious actions against road infrastructure, etc. Road managers can operate even more efficiently the road network (road cut, road

diversion, road securing works etc.) and can give better information to road users. The main possible impact is obviously to improve the safety of road users.

The OPERA project opens a new investigation field about the treatment of video data by off-loading process. It enables the road manager to become aware of new opportunities provided by postponed treatment of video data by high efficiency servers (and probably by new artificial intelligence tools).

Monitoring of road infrastructure (and not only road traffic), better comprehension and anticipation of road traffic, better operation of the road under winter conditions are possible examples. Combined with new artificial intelligence tools and big data process, the innovation can deeply modify the road management by creating new opportunities that are to be further investigated.

REFERENCES

[1] Intel Corporation. (2017). 5-Level Paging and 5-Level EPT. https://software.intel.com/sites/default/files/managed/2b/80/5-level_paging_white_paper.pdf
[2] "*Do-It-Yourself Virtual Memory Translation*", http://dl.acm.org/citation.cfm?id=3080209
[3] "*Rethinking TLB Designs in Virtualized Environments: A Very Large Part-of-Memory TLB*", http://dl.acm.org/citation.cfm?id=3080210