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

Exploitation plan - preliminary report 1

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

DELIVERABLE NUMBER	D8.6
DELIVERABLE TITLE	Exploitation plan - preliminary report 1
RESPONSIBLE AUTHOR	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	STMicroelectronics
DELIVERABLE N. | TITLE	D8.6 | Exploitation plan - preliminary report 1
RESPONSIBLE AUTHOR	Giulio URLINI, STM
DATE OF DELIVERY (CONTRACTUAL)	31/03/2017
DATE OF DELIVERY (SUBMITTED)	31/03/2017 (M16)
VERSION | STATUS	V5 | Final
NATURE	R(Report)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Giulio URLINI (STM), ; Jean Christophe MAISONOBE (ISERE); Simone CICCIA, Alberto SCIONTI (ISMB); Stéphane BARBIER (NEAVIA); Idan Yaniv (TECHNION); Frank Verhagen, Dirk Harryvan (CERTIOS); Luca SCANAVINO (CSI); Joel NIDER (IBM); Richard CHAMBERLAIN (NALLATECH); Gallig RENAUD (HPE)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Table of Contents	30/12/2016	Giulio URLINI, STM
0.2	First review	21/03/2017	Giulio URLINI, STM
0.3	Integrated the contributions from several partners	27/03/2017	Giulio URLINI, STM; Jean Christophe MAISONOBE, ISERE; Simone CICCIA, Alberto SCIONTI, ISMB; Stéphane BARBIER, NEAVIA; Idan Yaniv, TECHNION; Frank Verhagen, Dirk Harryvan, CERTIOS; Luca SCANAVINO, CSI
0.4	Integrated other contributions and harmonized the content	28/03/2017	Giulio URLINI, STM; Florin APOPEI, TESEO; Joel NIDER, IBM
0.5	Final version ready for internal review	30/03/2017	Giulio URLINI, STM; Richard CHAMBERLAIN, NALLATECH; Gallig RENAUD HPE
1.0	Final release ready for the submission	31/03/2017	Giulio URLINI, STM

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)		Gallig Renaud Email: gallig.renaud@hp.com
NALLATECH LTD		Craig Petrie Email: c.petrie@nallatech.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: s.serra@teseo.clemessy.com
DEPARTEMENT DE L'ISERE		Olivier Latouille Email: olivier.latouille@isere.fr

EXECUTIVE SUMMARY

This deliverable contains the first evaluation of the potential exploitation impact of the OPERA project on three main pillars:

- New market segments: the introduction of new paradigms can enable new markets. In the context of video surveillance devices autonomous in terms of power consumption and intelligence of scene interpretation will be the base for a new approach, in the road management field, as deeply analysed in the project, as well as in other fields, like monitoring scenarios, critical contexts and environments where the lack of power and communication infrastructure is an important issue. The autonomy of the device in terms of scene and event interpretation could also enable application where the human supervision should be reduced or allow the deployment of a bigger number of devices, that can't be controlled by enough personnel. In the truck demonstrator a new application of small form factor servers can enable new types of applications with local mobile servers.
- Existing market segments. In this area the innovation introduced by OPERA can impact on the optimization of the actual servers, with a significant reduction of power consumption with the related effects, like the reduction of heat dissipation.
- New research directions: the virtual desktop infrastructure (VDI) use-case of OPERA initiated an innovative research in the fields of virtualization and workload placement. The OPERA project wish to utilize virtualization and containerization technologies to achieve its energy efficiency goals under this use-case. Consolidating many virtual machines on a small number of physical servers is a well-known technique to reduce the power costs of large-scale databases. The OPERA project adds another level of sophistication to this virtualization model by introducing multiple heterogeneous platforms in the same cluster, which requires: (1) new technologies to migrate applications between different instruction set architectures (ISA), (2) accurate models to predict the performance and energy consumption of workloads running on different machines, and (3) new optimization techniques that are able to find efficient workload placements across the available machines.
  The OPERA work under WP4 also explore alternative energy efficiency models and methods, e.g., energy harvesting techniques that may exploited in the future for many other use-cases.

The exploitation plan of the OPERA project is based on the continuous evaluation of the potential impact that will be described and updated in the intermediate deliverable D8.10 and finalized with the indication of the actions to be executed after the end of the project in D8.2.

TABLE OF CONTENTS

1 GLOBAL EXPLOITATION STRATEGY

The main goal of the exploitation of a research project is the use of the results during and after the project's implementation. It can be for commercial purposes but also for improving policies, and for tackling economic and societal problems.[1]

The exploitation plan must consider the elements in the research that will impact specific topics and needs to drive the research and the development to increase such impact in interesting fields.

This document will describe the analysis of the main research topics of OPERA and the foreseen impact that they could have on the market and on new research activities.

In the second part of the document the strategies that can be applied during the execution of the project for the maximization of such impact will be analysed and described. Finally, the last section will be the report on the exploitation actions foreseen to be executed after the end of the project.

These two sections will be very short in this version, and they will be populated in the next versions of the exploitation document, with the results of the project and their influence to the exploitation.

The exploitation activity will be connected to the development of a business plan, described in the deliverables produced by task 8.3, "Business planning". The Business Plan will be related to the exploitation strategy and it will identify also the general strategy of the project, the market, the market segmentation, the business model, the financial analysis and the economic impacts.

[1] https://ec.europa.eu/research/participants/portal/desktop/en/support/faqs/faq-933.html

2 ANALYSIS OF THE POTENTIAL IMPACT

2.1 REVIEW OF THE MAIN GOALS OF OPERA

The OPERA project has many different goals for the achievement of the computing continuum and its demonstration in three different scenarios. The project itself has also defined a mechanism for the measure of the improvement in terms of efficiency and power consumption, globally considered at system level. The second method implemented in the project for the evaluation of the results is the cyclic lesson learned process in order to evaluate the temporary results and provide the corrections for the next cycle. The first round of evaluation and production of the first lesson learned is planned for M21, in August 2017, when the deployment of the first prototypes on the real life scenarios will be completed and the evaluation of the results conducted.

The activities of the consortium has identified several aspect among others that can have a strong weight on the exploitation.

2.1.1 Small form factor data center – impact

The OPERA research is focused on reducing the power consumption inside a small form factor data center. Several approaches to achieve this goal has been chosen, each of which has a potential impact on the way data centers are to be built and managed in the future.

By improving the workload management, we can better match the compute power available to the work that needs to be done. This is an important improvement, because we can better estimate the hardware that is needed for a given workload. By choosing heterogeneous hardware, we can better control power consumption of the solution by migrating the load to the best available architecture and match an optimal power to performance ratio.

We see seamless container migration as a major hurdle to overcome on the way to adopting heterogeneous cloud architectures. By developing the technology required to make container migration easy and efficient, cloud architects will gain more freedom when designing new data centers, or retrofitting older ones. By selecting compute nodes from among a larger pool of available machines gives the architects a wider range of options in CPU, memory and I/O capabilities. In addition, the freedom to choose from multiple chip vendors can also have an impact on the capex (capital expenses) when building the data center, by providing multiple options for providing compute power. Increased competition between vendors usually benefits the end-user by increasing quality and lowering costs.

Our developments have already had an impact on the open-source project CRIU, which is the basis for migration in container solutions such as Docker, LXC and OpenVZ. Our patches provide an option for post-copy migration, which reduces down-time, and reduces network utilization during the migration. Post-copy migration is made possible by using a feature called “userfault fd” in the Linux kernel, which was developed as part of the EU funded FP7 project ORBIT. Our future developments related to the low-latency interconnect will have a further impact on post-copy container migration by reducing the time taken to serve a remote page fault. This will bring heterogeneous container migration one step closer to reality, and acceptance in the industry.

The architecture design and the modifications applied to match requirements for truck use-case opened new concrete possibilities. By leveraging FPGA board together with Moonshot cartridges in a small chassis, we are able to provide new disruptive solutions in multiple areas which require dense compute at low energy.

For instance, eBrisk has ported their 4K HDR encoding library to this solution, providing to their customers a 4x improvement in term of density versus the densest solution available today:

http://www.prnewswire.com/news-releases/ebrisk-enables-4-hevc-4k60fps10bithdr-channels-in-1-rack-unit-610269645.html

2.1.2 Ultra-low power computing impact

The ULP camera designed in OPERA gives a considerable and previously unseen capability to collect data across interurban road networks for supporting advanced road traffic management.

The ULP camera designed in OPERA enables to monitor all areas (including areas far from electrical grids and communication wired network) thanks to energy autonomous and ULP wireless communication technologies. Such a technology enables to systematically monitor high road traffic issues areas, in particular if no bulky energy autonomous installations, minimizing equipment's and civil engineering infrastructure are developed. OPERA enables to monitor under realistic economical and operational conditions any area (even if isolated and far from conurbation), like critical interurban cross-road and roundabout, road structures like bridge and tunnel, access to ski resorts or high mountain passes, etc…

The ULP OPERA micro servers provide the required embedded smartness to cover a large scope of use cases: detection of congestion, detection of wrong way vehicles, and detection of cycles as tested in OPERA. However, other use cases might be explored: detection of stopped vehicles, detection of snow covering roads, detection of natural risks, etc… More and more advanced algorithms will enable to increasingly use the capability of this embedded smartness, which can be operated thanks to ULP technologies.

The use cases cover all road traffic issues: road traffic density, road traffic meteorological conditions (in particular during winter periods), natural events, human events

The local information treatment enables to automatically transmit an alarm to the operators of the traffic management center which makes possible the monitoring of a very large number of areas across the road network, without exceeding the viewing capability of human operators.

ULP OPERA camera combines a large covering of both geographical areas and road traffic use cases and a new capability of real time treatment of a large number of events: It enables to give real time information to road users who are thus better informed and to operate the road network more and more efficiently and in real time. Considerable impact is the improvement of the safety and quality of life of road users, the reduction of the negative effect of congestion on people life, environment and economic activity.

The ULP OPERA camera also enables to collect and to store video data that can be treated in postponed mode by centralized smartness. Big data processing enables to analyse video data collection in a large number of areas and/or during a long period. Statistics can be elaborated to enlighten public policies and to improve the road infrastructure management. As an example a new process to monitor the road infrastructure (road surface, bridge, and tunnel) could be investigated.

ULP camera combined with ultra-low power server will be decisive to collect and to treat road big data, in particular in postponed mode. If they full meet the ULP objectives, ULP technologies could be applied to an increasing part of the European road network (more than 5 000 000 km of roads, any type of road included)

From business perspective, benefits from ULP design will enable introduction on the market of a brand new generation of video solution for road applications. Firstly, it will reduce drastically energy needs and thus will lead to much more compact solution. Time to market will be shorten as development phase during the project will give a competitive benefit over competition. The benefits will also allow embedding video solution in product without any video solution today. One can think of traffic lights or public lighting equipment. Today those equipment cannot embed such video solution due to electrical needs and/or energy storage needs. Those hurdles will be overcome by the use of ULP platform as develop in Opera project. This breakthrough on the market should allow massive use of video solution in the scope of road monitoring and should open the door to new use cases. One can think of V2X business where video solutions are not present today. Having road infrastructure with relevant and compact video solution will allow pushing in the car some video scenes useful for the driver: dangerous situation, pedestrian crossing detection among others.

2.1.3 Investigation of low power optimized communication

Nowadays, the trend of autonomous wireless sensors is oriented to high data rates and minimum energy consumption. Unfortunately, the two parameters are in the same direction and switch on/off the transmitter is no longer sufficient. The ability to reconfigure an antenna plays a crucial role in this scenario. This aspect is not yet fully exploited. OPERA wireless links want to demonstrate superior performance by means of wireless systems able to control the radiated energy of antennas. In particular, all radiated energy is concentrated towards the desired direction. This mechanism requires less power to transmit the same information. Therefore, innovation can be identified in the fact that the communication link adopts reconfigurable antennas able to establish high rate communication at minimum energy consumptions.

Several applications can benefit from this energy-aware transmission mechanism, especially those applications that require to be completely autonomous in terms of power sources. For instance, deployable autonomous computing nodes can be built, by integrating low power computing elements, capabilities of being powered by renewable sources, and reconfigurable antennas. Other potential applications can be seen in a more effective and efficient management of unmanned vehicles (e.g., drones).

2.1.4 Heterogeneous architecture and optimal workload allocation

Data centres (DCs) are continuously demanding performance and low power technologies to properly addressing the challenges posed by providing new services. Nowadays, heterogeneity has been recognized as the best way to address such challenges, by considering computing architectures and processing elements, spanning from different architectural implementations of traditional CPUs, to GP-GPUs, to more specialised devices such as FPGAs and custom ASICs. Such heterogeneity is necessary to better run (i.e., improving performance – the amount of elaborated information, and reducing as much as possible power/energy consumption) applications and services, that nowadays, cover different domains: SaaS applications, machine learning, big data, etc. Correctly managing the workloads (that are also heterogeneous) becomes of primary importance for Cloud providers. From this viewpoint, the OPERA action aims at contributing in designing a set of mechanisms and policies to correctly instantiating and managing cloud applications. These mechanisms and policies will converge into a software module that will be tested on a small-scale cloud system. By testing such mechanisms and policies within this small scale testbed environment, OPERA will demonstrate the benefit of such "smart" way of instantiating applications also in large-scale cloud infrastructures. Further interest can be arisen in the developers' communities regarding the development of cloud orchestration tools. At the moment of writing this report, there is no a consolidated solution capable of addressing "cloud orchestration" taking care of the optimization of the energy efficiency of the underlying infrastructure.

The development language selected for the application used in the truck use-case is key to the success of optimizing the resource allocation. OpenCL being a standard compatible with a wide variety of accelerators, we ensure that the work done within OPERA will fit well on nowadays architectures but also on future ones.

The porting of MICMAC application will require to understand the way it is working internally through a profiling exercise. From this result, we will understand which part of the application will be providing the best performance / energy ratio on each of the 3 available architectures (x86, ARM and FPGA). This internal workload allocation is very important to avoid, for instance, over-consuming FPGA logical elements where the ARM could perform just as good in a lower power envelop.

In the OPERA project the MICMAC application has been chosen for several reasons, but it is not the only application supported, for this specific use. Other applications can benefit from the acceleration provided by the OPERA platform, enlarging the scope of the exploitation.

2.1.5 Cross-evaluation on power metrics and new paradigms of the evaluation and optimization of power balance

The application of the power metrics defined in OPERA will provide us a new way of comparing traditional approaches and innovative aspects introduced by OPERA. This comparison will influence the exploitation of the OPERA research activities.

The Energy Efficiency Models chosen for use in the OPERA project products are detailed in D4.1. The following choices for the energy efficiency modelling are made:

- efficiency will be calculated using use phase energy only;
- The final use phase energy will be measured in kWh and not converted to primary energy;
- Dependant on the products, project demarcations will either be system or equipment level.

Energy Efficiency Metrics (EEM): In line with available generic EEM descriptors, specific EEMs have been described for use within the OPERA project, these EEMs all have the form:

$$Energy\ Efficiency = \frac{functional\ unit}{energy\ used\ to\ produce\ this\ output}$$

Where energy is measured in the “final use phase” on a system or equipment level as dictated by the energy efficiency model used for each specific product and use case.

The models and EEM’s thus defined and described for each specific use case have been evaluated and approved using RACER criteria. RACER stands for Relevant, Acceptable, Credible, Easy and Robust. The EEM’s can and will be used to substantiate the OPERA energy efficiency claims and during the project to evaluate the effectiveness and progress of the methods developed for improving energy efficiency.

Central to the methods that will be used in the OPERA project is the concept of heterogeneity. As indicated in this document, the use of heterogeneous compute architectures can have a significant positive impact on computational efficiency in general and in the use cases that are part of the project specifically.

A management framework and tool, that manages workload on a heterogeneous compute architecture in such a way as to maximise energy efficiency is currently unavailable and will be created as part of the OPERA project. Although developments in both IT hardware and software as well as developments in data center facility management have resulted in improved energy efficiency, none of these developments have put the application energy efficiency in the centre of the framework.

OPERA uses all the latest available technologies to create the currently unavailable management tooling, specifically tailored to support heterogeneous architectures. This heterogeneity is a key element in achieving the energy efficiency claims within the OPERA project.

To exploit heterogeneity without incurring efficiency penalties, it has been concluded that workload decomposition into so called micro-services running in virtualisation containers is most likely to result in the highest energy efficiency. These containers will be made portable across the heterogeneous architecture to be able to shut down or power up computational elements as load conditions dictate. The concept of application portability over different Instruction Set Architectures (ISA's) is another one of the innovations of OPERA. The project will deliver extensions to open source compilers to enable this functionality.

For the novel system management tooling to identify the most efficient placement of a specific micro-service, extensions to an application description framework are created. The TOSCA framework has been selected as the basis of this work. TOSCA is again an open source development, enabling the OPERA project to create modifications while at the same time giving an avenue for sharing the developments with the community.

Since the virtualisation layers also add complexity and overhead in the memory management of the servers, a novel memory mapping methodology is tested to assess its impact in overall energy efficiency.

To improve the efficiency of the workload migration, different protocols will be investigated for the OPERA system node interconnect. A low latency interconnect will allow memory sharing between different system nodes, in turn allowing post migration memory copying that results in very small application delays during an application migration.

The expected impact of these developments mentioned above, is described in the OPERA Innovation Potential (D2.5) and this deliverable will be reviewed another 3 times until the very end of the project.

3 ACTIONS FOR THE MAXIMIZATION OF IMPACT

The objective of the partners involved is to progress in the research activities while considering the outcome of the aspects more connected to the exploitation at the end of the project trying to identify the elements that can improve the exploitation itself.

A first step towards the dissemination of the results both in terms of academic presentations and relation to industrial actors has been the creation of a resident set of demonstrators, designed in a portable suite, in order to be able to present the results of the project, for the moment only partial, and to demonstrate practically these results.

To present these results two brochures have been designed, one for an academic target, and another for a more industrial audience.

This activity has impacts on the dissemination activities of the project as well as on the exploitation of the results. It will have also an impact on the business development activities once more industrial actors will be involved.

For the moment the main exploitation activities already conducted are related to the traffic monitoring use case that involves anyhow all the elements of the project. The main reason is that one of the partner of the project is a potential end user of the results that can evaluate directly the impact on its activities and on the market segment in general, the *Departement de L'isere*.

3.1 ULTRA-LOW POWER COMPUTING PLATFORM

The evaluation of the actions for the improvement of the impact of the research around the autonomous ULP device for local computing in the context of traffic monitoring applications is started with a detailed analysis of the issues related to the installation, maintenance and working condition of such a device in a real scenario. The requirements defined by the end user (ISERE) and the industrial partner working in this sector (NEAVIA) have been compared to the development activity in order to answer to all these requirements, possibility with a solution is some cases, and a clarification of the actual limits of the solutions in others. In addition to this activity an evaluation of the possible solution of all the issues in a real products has been conducted, increasing the visibility of the end user and the potential producer of a final device about the potentiality of the research.

The details of this study will be presented in the lesson learned deliverables that will describe the choices made to fulfil these requirements and the description of the possible solutions for issues not covered by the OPERA project.

In summary for the exploitation the main elements considered were the cost and the power budget of a system that should be autonomous in terms of power enabling the video processing functionalities keeping low the data rate transmitted, processing locally the major part of the data and leaving to a remote processing only the features. In parallel other systems, like the heating of the lens and the possible harvesting techniques has been evaluated in order to provide a power level required for these type of devices in the real final use.

These elements are considered in the global picture, considering the total power budget of a real product and the reduction of personnel required for the surveillance, increasing the level of detection.

This first step should be extended in the coming months with the lessons learned that will bring the consortium to a new prototype, more in line with the project and application requirements.

3.2 SMALL FORM FACTOR DATA CENTER

The hardware family used to answer the small form datacentre constraints in OPERA project has been existing inside HPE's portfolio since 2013. It's been evolving over the years following technology trends and customers' feedback, mostly from the enterprise IT market.

With OPERA project and more especially due to some requirements formulated by CSI, modifications such as the integration of Redfish to enable monitoring and management of the entire heterogeneous system, open the path to new or better integrations capabilities into open platforms such as OpenStack. This kind of integration of standard API is key to get people's interest and insure:

- The development made for OPERA will be reused in other projects
- The standard grows and establish itself as de-facto solution

The integration on the heterogeneous low-power architecture done through the "Datacentre in a truck" use-case provides a perfect technology validator and demonstrator. Due to the space saving which is coming from the use of small form factor systems, we are able to expand the feature set by bringing high-availability, on top of providing better performance and lower power usage of the overall solution. Bringing industrial resiliency into low form factor datacentre is a typical example of feature which could benefit to new research domains, such as autonomous vehicles.

3.3 INVESTIGATION ON LOW POWER COMMUNICATION

The technologies related to wireless communication are changing very rapidly in these years for the support of a variety of small, possibly autonomous devices in the IoT domain. A long range communication that reduces the power needs and keep enough bandwidth is the Holy Grail of the IoT domain explosion.

The research conducted in OPERA is somehow transversal to the various emerging protocols. It is focused on the optimization of the communication channel through long distances, trying to reach the two challenging goals described before, less power and enough bandwidth. The exploitation of this technology so will be transversal to the new wireless protocols emerging in these days (Lora, Narrow Band-IoT among others), trying to support the devices implementing these protocols with the optimization of the channel. The activities in this area for the enhancement of the exploitation will be the evaluation of the use of the antenna in new contexts, with industrial partners as STMicroelectronics planning new products integrating the antenna.

3.4 HETEROGENEOUS ARCHITECTURES

The combination of technologies provides difficulties in terms of algorithm porting. Combining x86, ARM and FPGA technologies has the potential to significantly improve compute performance and reduce energy consumption. Part of the Opera project will involve the porting of the image processing software MICMAC as part of the "Datacentre in a truck" use case. This will involve exploiting the x86, ARM and FPGA technologies where appropriate. The lessons learned from this task will help the consortium to understand how different processing elements can be optimally utilised in terms of performance per watt.

To understand the power benefits gained from heterogeneous systems it is necessary to carefully monitor the power consumption. This requires the extension of industry standard power monitoring software to support FPGA accelerators. This software (RedFish) will used to measure the effects of algorithm partitioning to best understand which technology is most energy efficient during the upcoming MICMAC porting.

The development of a software tool flow to support the FPGA prototype will allow the consortium, in the coming months, to quickly evaluate the different technologies and is key to exploiting the system to its full potential.

4 EXPLOITATION ACTIVITIES AFTER THE END OF THE PROJECT

This chapter is left empty for the present version. It will be populated during the analysis of the results of the field trials and enhanced by the actions taken by the partners and the consortium as a whole for the exploitation of the results that will be available in the coming months.