OPERA

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

Innovation Potential of OPERA Platform 3

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

DELIVERABLE NUMBER	D2.7
DELIVERABLE TITLE	Innovation Potential of OPERA Platform 3
RESPONSIBLE AUTHOR	Dirk Harryvan (Certios)

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	WP2 | Low Power Computing Requirements and Innovation Engineering
WORKPACKAGE LEADER	ISMB
DELIVERABLE N. | TITLE	D2.7 | Innovation Potential of OPERA Platform 3
RESPONSIBLE AUTHOR	Dirk Harryvan (Certios)
DATE OF DELIVERY (CONTRACTUAL)	31/03/2018 (M28)
DATE OF DELIVERY (SUBMITTED)	31/03/2018 (M28)
VERSION | STATUS	V1.0 | Final
NATURE	R (Report)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Dirk Harryvan (Certios); Frank Verhagen (Certios), Joel Nider (IBM); Richard Chamberlain (Nallatech); Gallig Renaud (Nallatech); Idan Yaniv, (Technion); Nathalie Viollet (HPE); Albert Scionti (ISMB); Simone Ciccia (ISMB); Giulio Urlini (ST).

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
v 0.1	Table of Content (TOC)	07/02/2018	Frank Verhagen, Dirk Harryvan (Certios)
v 0.2	Preparation of adding content	09/03/2018	Frank Verhagen (Certios)
V 0.7	Contributions from partners	28/03/2018	Joel Nider (IBM) Idan Yaniv (Technion) Richard Chamberlain, Gallig Renaud (Nallatech) Nathalie Viollet (HPE) Alberto Scionti (ISMB).
V0.8	Final draft version after review by Nallatech	28/03/2018	Richard Chamberlain (Nallatech)
V0.9	Final draft version after review by TESEO	29/03/2018	Roberto Peveri (TESEO)
V 0.10	Final review by the editor	30/03/2018	Dirk Harryvan (Certios)
V1.0	Final review by the coordinator	31/03/2018	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)		Cristian Gruia Email: cristian.gruia@hpe.com
NALLATECH LTD		Craig Petrie Email: Richard.Chamberlain@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: s.serra@teseo.clemessy.com
DEPARTEMENT DE L'ISERE		Olivier Latouille Email: olivier.latouille@isere.fr

ACRONYMS LIST

Acronym	Description
ADAS	Advanced Driver Assistance Systems
API	Application Programming Interface
ASID	Address Space Identifier
BSP	Board Support Package
CAPI	Coherent Accelerator Processor Interface
CDNN	CEVA Deep Neural Network
CNN	Convolutional Neural Network
CPS	Cyber-Physical System
CPU	Central Processing Unit
CRIU	Checkpoint/Restore In Userspace
D2.5	Deliverable 2.5
DiaB	Datacenter in a Box, a hardware product of the OPERA project
DSP	Digital signal processing
EC	European Commission
FD-SOI	Fully Depleted Silicon On Insulator
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
GPGPU	General-Purpose Graphics Processing Unit
HPC	High Performance Computing
HW/SW	Hardware/Software
I/O	Input/Output
IoT	Internet of Things
ISA	Instruction Set Architecture
ISP	Image Signal Processing
LLVM	Low Level Virtual Machine
PCIe	Peripheral Component Interconnect Express
PUE	Power Usage Effectiveness
PWC	Page Walk Cache
RDMA	Remote Direct Memory Access
RF	Radio-frequency

RF MEMS	Radio Frequency Microelectromechanical system
SoC	System on Chip
SOTA	State of the Art
SWaP	Size Weight and Power
T2.3.1	Task 2.3.1
TCO	Total Cost of Ownership
TMC	Traffic Management Center
ToC	Table of Contents
TOSCA	Topology and Orchestration Specification for Cloud Applications
TPU	Tensor Processing Unit
ULP	Ultra-Low Power
WP	Work Package
YOLO	You Only Look Once

Table 1 Acronyms List

LIST OF FIGURES

Figure 1 Linked WP's 6
Figure 2 Timeline iterations D2.5, 2.6, 2.7 and 2.8 6
Figure 3 The OPERA main objectives mapped on the reference overall architecture 11
Figure 4 Nvidia Volta specification 15
Figure 5 271 MW in power means 2.4 TWh/year 20
Figure 6 PUE formula 22

LIST OF TABLES

Table 1 Acronyms List 5
Table 2 List of actions and roles 7
Table 3 Impact of saving energy in data centres 22

EXECUTIVE SUMMARY

1.1 POSITION OF THE DELIVERABLE IN THE WHOLE PROJECT CONTEXT

This project's deliverable D2.7 researches the latest innovation potential of the OPERA project and the products realized in the context of this project and the deliverables D2.5 and D2.6. There will be 4 iterations; this document is the result of the third iteration. The documentation of the innovation potential is part of the WP2. The relationship of this work package to the other WP's of the OPERA project, can be visualized in the following graph:

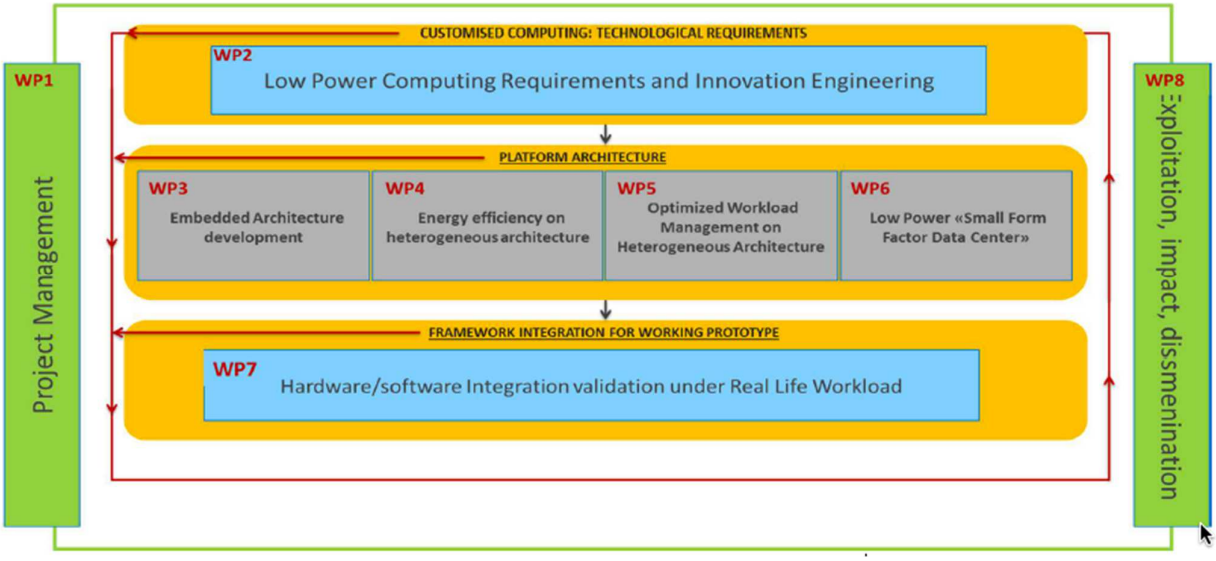

Figure 1 Linked WP's

1.2 DESCRIPTION OF THE DELIVERABLE

Like deliverable D2.6, this deliverable D2.7 is delivered as an addendum to the D2.5 Innovation Potential deliverable. Presenting D2.7 in this way avoids repeating text of D2.5 and D2.6. The four iterations have been spread in time, over the duration of the OPERA project. Where the D2.5 deliverable had the original idea of delivering the OPERA project Potential in M10 of the project, the following iterations have been added after M10, in order to enable addendums (D2.6 and D2.7) to D2.5, to be integrated as a final document on the OPERA Innovation Potential, D2.8 in M36. The decision to deliver 4 (intermediate) reports is consistent with the idea of OPERA as a research and development project; in the progress of the project, more and more is known about the innovative potential of OPERA products.

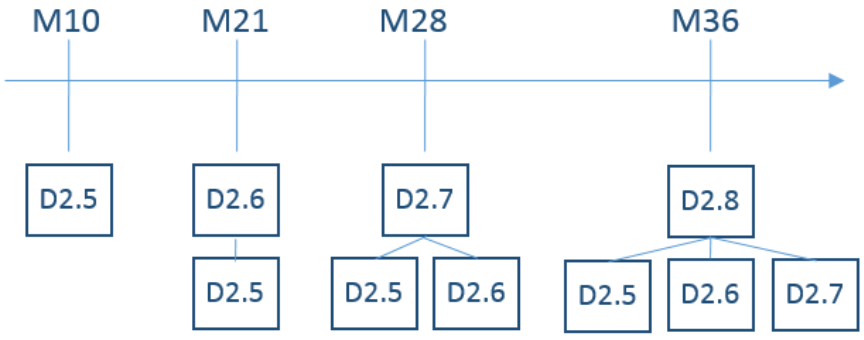

Figure 2 Timeline iterations D2.5, 2.6, 2.7 and 2.8

1.2.1 Positioning the iterations

1.2.1.1 D2.7 (M28)

In D2.6 there was already a strong reference to the D2.5 deliverable that was submitted in M10 and updated and resubmitted in M18. In the D2.7 addendum, thanks to the progress of the OPERA products, new product market combinations are likely to be identified. In D2.7 we will focus on the way of translating the innovations, the new technology, to the market and creating opportunities for OPERA partners, for each of the innovations and involved partners. In addition to this, the section 4, identifying of Potential Markets and section 5, Market Analysis have been enriched; identifying extra new entrants in the competitive arena for OPERA products.

1.2.1.2 D2.8 (M36)

The final document D2.8 will be an integration of the latest insights, the first deliverable D2.5 and the 2 intermediate deliverables, the addendums D2.6 and D2.7. Deliverable D2.8 will be an independent report that can be read independently from the earlier iterations.

1.3 LIST OF ACTIONS AND ROLES

LIST OF ACTIONS											
ACTIVITIES LIST AND PARTNERS ROLES	**CERTIOS**	CSI	HPE	IBM	**ISMB**	LD38	NALLATECH	NEAVIA	**ST**	TECHNION	TESEO
Summary of the initial requirement	P										
Components to be integrated	P		P	P	P		P		R		
Integration output: first iteration (D2.5)	P		P	P	P		P	R	R		R
Integration output: second iteration (D2.6)	P	R	P	P	P		P		P	I	I
Integration output: third iteration (D2.7)	P	I	P	P	P		P		P	I	P
Integration output: fourth iteration (D2.8)	P	I	P	P	P	R	P	I	P	I	P

Table 2 List of actions and roles

- P = Participating (includes I & R)
- I = Input delivery (Includes R)
- R = review
- **Bold** → assigned in project proposal to contribute to this the task

As shown in Table 2, especially the privately held organizations are expected to participate, give this deliverable input and review the document D2.6 itself. The further the project will progress; the more organizations will be able to contribute in detail of what to expect from OPERA's innovations.

TABLE OF CONTENTS

1.1 POSITION OF THE DELIVERABLE IN THE WHOLE PROJECT CONTEXT 6
1.2 DESCRIPTION OF THE DELIVERABLE 6
1.2.1 Positioning the iterations 7
1.3 LIST OF ACTIONS AND ROLES 7
2 INTRODUCTION 10
2.1 OBJECTIVE OF THIS DELIVERABLE 10
3 INNOVATION OF OPERA SOLUTIONS 11
3.1 PRODUCTS AND SERVICES 11
3.2 TECHNOLOGY INVOLVED IN OPERA 11
3.3 WHAT TECHNOLOGY ARE OUR PARTNERS WORKING ON? 12
3.3.1 HPE 12
3.3.2 IBM 12
3.3.3 ISMB 12
3.3.4 Nallatech 12
3.3.5 ST 13
3.3.6 Technion 13
3.4 WHAT IS THE CURRENT STATE OF THE ART? 13
3.4.1 HPE 13
3.4.2 IBM 13
3.4.3 ISMB 14
3.4.4 Nallatech 14
3.4.5 ST 15
3.4.6 Technion 15
3.5 HOW ARE YOU GOING TO MOVE PAST THIS STATE OF THE ART? 16
3.5.1 HPE 16
3.5.2 IBM 16
3.5.3 ISMB 16
3.5.4 ST 17

3.5.5 Technion 17
4 POTENTIAL MARKETS 18
4.1 MIGRATING CONTAINERS 18
4.2 ARTIFICIAL INTELLIGENCE / DEEP LEARNING 18
4.3 VIDEO SURVEILLANCE 18
4.4 INDUSTRY 4.0 19
4.5 DATA CENTRES 19
4.5.1 Example in the Dutch market 19
4.5.2 Attribution of energy use to servers 20
5 MARKET ANALYSIS 21
5.1 MIGRATING CONTAINER 21
5.2 ARTIFICIAL INTELLIGENCE / DEEP LEARNING 21
5.3 DATA CENTRES 21
6 IMPACT ENVISAGED 22
6.1 VIDEO SURVEILLANCE 22
6.2 DATA CENTER 22
7 REFERENCES 24

2 INTRODUCTION

2.1 OBJECTIVE OF THIS DELIVERABLE

The objective of this report is to present the innovation potential of the OPERA products and services. This report is an update of D2.5 Innovation potential of OPERA – Platform 1 which was submitted in M10 and resubmitted in M18. This report, deliverable D2.7, is a third iteration in a sequence of 4 iterations in total:

- D2.5 Innovation potential of OPERA – Platform 1 (M10)
- D2.6 Innovation potential of OPERA – Platform 2 (M21)
- D2.7 Innovation potential of OPERA – Platform 3 (M28)
- D2.8 Innovation potential of OPERA – Platform 4 (M36).

In order to prevent producing a document that is growing 'bigger' with each iteration, and in which it is increasingly difficult for the reader to find the latest input to the document, we have chosen to leave D2.5 as the base document and only to present new insights (new since producing the former iteration) in the newer iterations. The Table of Contents (ToC) in the later iterations will only focus on the 'dynamic' parts of the content, i.e. the contents that are progressing over the duration of the project. The more 'static' parts of the project, like the

- Methodology (D2.5, section 3)
- Innovation Envisaged (D2.5, section 4)
- How to achieve the 9 objectives (D2.5, section 5)
- Full description of the OPERA products and services (D2.5, section 6)
- Description of Other H2020 Projects (D2.5, section 9)

will therefore, like in D2.6, not be repeated in the iteration D2.7.

The objective of this iteration, D2.7, is to give some update of the dynamic parts of the D2.5 and D2.6 and see what the innovation for the different partners is going to be like (Section 3):

- Section 3 Innovation of OPERA Solutions (reference: D2.5, section 6 and D2.6 section 3)
- Section 4 Potential Markets (reference: D2.5, section 7 and D2.6 section 4)
- Section 5 Market Analysis (reference: D2.5, section 8 and D2.6 section 5)
- Section 6 Impact envisaged (reference: D2.5, section 10 and D2.6 section 3).

3 INNOVATION OF OPERA SOLUTIONS

3.1 PRODUCTS AND SERVICES

When we look at the objectives of OPERA and what objectives need to be reached in order to make the products of OPERA work (see Figure 3), we expect to develop the following technical solutions as described in detail in Section 6 of D2.5:

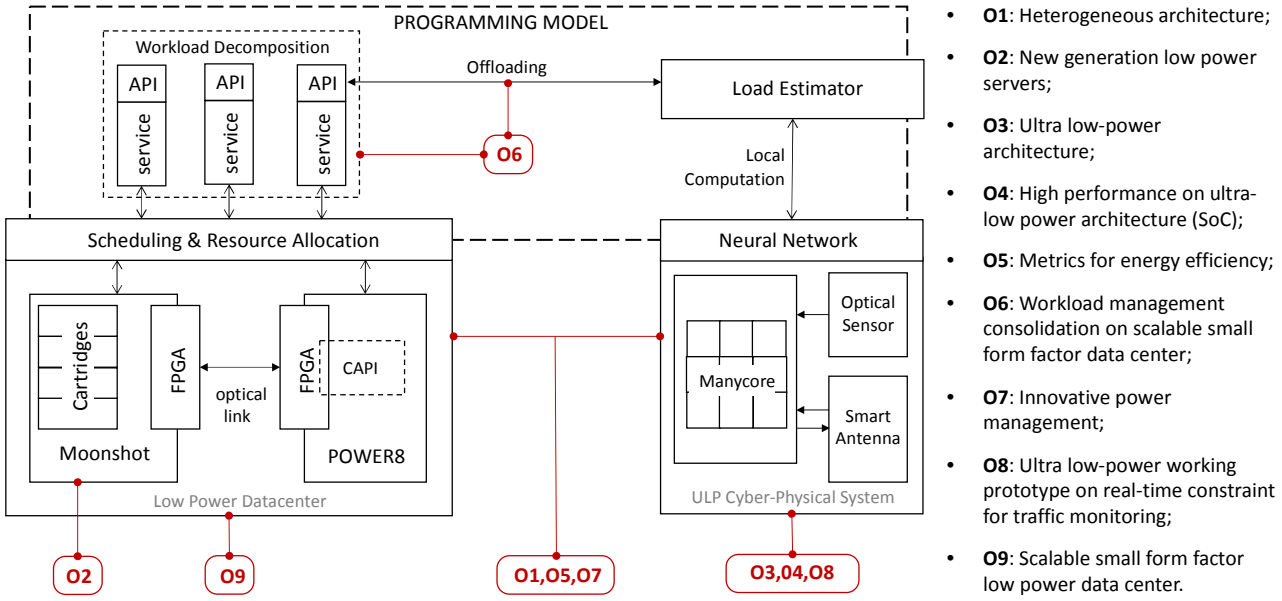

Figure 3 The OPERA main objectives mapped on the reference overall architecture

1. An FPGA card as application accelerator and high-speed interconnect, developed for OPERA also useable outside of the OPERA platform
2. Low-power heterogeneous data center architecture incorporating both HPE Moonshot and IBM POWER8
3. An ultra-low power autonomous traffic monitoring video sensor.
4. Extension of open source Cross ISA compiler to work with POWER8, enabling application migration over different CPU architectures
5. An energy aware workload management to maximize power efficiency of the heterogeneous architecture under different application loading
6. Smaller mobile datacenter solutions offering superior compute power per Watt in a small form factor for the mobile application
7. Extension of the TOSCA application description TOSCA is an open standard developed with the aim of ease the description and deployment of an application in a cloud environment.
8. Using CAPI (Coherent Accelerator Processor Interface) to attach a low latency interconnect for communication between servers
9. Virtual Memory Model, Developed for OPERA platform and applicable also outside of the OPERA platform
10. Incorporation of the reconfigurable antenna for low power wireless connectivity
11. Implementation of Redfish on the FPGA (PCIe) card
12. Coupling of multiple disparate FPGA boards.

3.2 TECHNOLOGY INVOLVED IN OPERA

To illustrate the innovation potential, it is necessary to illustrate the difference between the state of the art at the time of the start of the OPERA project. For the partners involved, we here go into the technology developments, the innovations, in specific areas of OPERA in the next sections of this paragraph. After

indications that the state of the art (SOTA) has been surpassed, we will go into the market opportunities that these innovations may unfold in the next section, in chapter 4.

3.3 WHAT TECHNOLOGY ARE OUR PARTNERS WORKING ON?

The different partners are working both individually as well as in combinations to achieve the products and services listed in paragraph 3.1. The link between the technology and these products and services is given, in square brackets [ ], for each of the mentioned technologies using the numbering schema from paragraph 3.1. This linkage is given in order to better understand the role of the technologies that these partners are working on within the OPERA project.

3.3.1 HPE

HPE is developing Moonshot, which is a Low Power and Heterogeneous workload optimized servers, as detailed in deliverables linked to WP6.
HPE is adding the integration with additional hardware (such as GPU, FPGA) to deliver best in class efficiency for given workloads.

3.3.2 IBM

Post-copy migration using CRIU (checkpoint restore in user space). Today, CRIU is able to work in two modes:

1. static checkpoint/restore to a disk
2. live migration by using the pre-copy method.

We are adding a third mode of operation which enables accessing memory on-demand from the remote machine during migration. This reduces the initial amount of data that needs to be copied, and therefore reduces downtime incurred by the migration.

3.3.3 ISMB

ISMB is working on three main aspects concerning both the data center side and the remote ultra-low power (ULP) cloud edge nodes. On the data center side, ISMB is actively working on implementing an efficient workload management system, as well as on machine learning algorithms acceleration using FPGA devices. Concerning the workload management, ISMB is responsible to implement innovative data center resources allocation policies that are able to improve global energy efficiency. To this end, different strategies have been studied and are currently under implementation. The result of such activity will be a software module that will be also integrated into well know and widely used cloud management stacks, such as the OpenStack framework. Also, ISMB is active in supporting the effective porting of Deep Learning algorithms on acceleration hardware. Along with Nallatech, a state-of-the-art convolutional neural network (CNN) has been successfully ported on a reconfigurable energy-efficient device (FPGA). Beside the OPERA project, further refinement of the current solution will be of worth also in other contexts. On the remote ULP nodes, ISMB is developing an advanced RF-communication system, based on a reconfigurable antenna. Specifically, we are developing steerable and directive antennas able to reduce the energy consumption in wireless devices, by exploiting the higher gain of the antenna, which reflects in a reduction of the transmitted power and/or the possibility to enlarge the communication distance without increasing the former by having a higher Signal to Noise Ratio (SNR).

3.3.4 Nallatech

The development of CNN networks running on FPGA to support an offload compute capability for ULP device. Particularly focusing on reduced bit precision networks where FPGA performance is particularly high.

3.3.5 ST

The activities of STMicroelectronics related to the development and utilization of the power efficient accelerator for Deep Learning Convolutional Neural Networks had a significant progress in the last six months. In particular the platform has been shared with the partner NEAVIA for the development of specific applications for the traffic monitoring field of usage, and the CNN approach has been shared with Nallatech and ISMB in order to provide different scales of performance among the OPERA continuum.
The development of a custom YOLO Network on the Orlando embedded platform, in parallel with the use of the standard full featured YOLO on the FPGA based server enabled a new class of applications. In terms of innovation, this development is creating a new set of applications, and enabling new markets for video surveillance, demonstrating the capabilities of autonomy in terms of power and processing capabilities of the Orlando solution.

3.3.6 Technion

Under WP5, the Technion deals with performance modelling of compute-intensive and memory-intensive applications. Today, most of these applications are running in cloud environments ("About 75% of x86 server workloads are virtualized", according to this (Bittman, Dawson, & Warrilow, 2015). Cloud providers, like Microsoft and Amazon, thus wish to optimize their data center utilization by improving the resource allocation among the applications they host. In other words, cloud providers want to provide better performance to their customers while reducing their operational costs (which are mostly related to energy). To this end, the Technion develops performance models that are able to predict the runtime and energy of applications as a function of their compute and memory resources. Besides guiding resource allocation decisions among different applications that run on the same physical machine, these models will be able to guide placement algorithms that have to choose the optimal scheduling of different applications on a heterogeneous cluster of machines, like those managed by cloud vendors. The cloud orchestration software developed by ISMB under WP5 takes the latter approach and will be based on the Technion findings.

3.4 WHAT IS THE CURRENT STATE OF THE ART?

3.4.1 HPE

Moonshot is recognized as one of the best Low Power platform on the market[1]. There is consensus[2] that offloading applications to special purpose cores is the way to minimize the power consumption of a given workload. Getting orchestration software to understand both hardware and application is key for the overall solution efficiency.

3.4.2 IBM

Container migration between two host servers is currently done with a pre-copy method. The pre-copy method creates a copy of the memory pages associated with the container on the target host. When the target memory is populated, a checkpoint is created, and the container is started on the target host.
The pre-copy method will send multiple copies of the same memory page in the case that the contents of that page changed before the migration process completes. This process causes higher than necessary network bandwidth, which wastes energy as well as puts pressure on the network. On the positive side, this is a relatively safe method which means the chances of losing the container during migration is very low.

[1] https://itbrandpulse.com/it-pros-vote-2017-server-and-database-brand-leaders/ (Dense Low Power Microserver HPE's Moonshot earned its third straight Market Leader achievement, along with Performance, Reliability, Service & Support, and Innovation).

[2] http://ieeexplore.ieee.org/document/1310259/ (General-purpose versus application-specific processors) https://web.eecs.umich.edu/~taustin/papers/CASES02-asp.pdf (Application Specific Architectures: A Recipe for Fast, Flexible and Power Efficient Designs).

In the case of a catastrophic failure (network connectivity is lost between the hosts, or the target host goes down) the migration process may be cancelled, and the container can be resumed on the source machine without any damage.

3.4.3 ISMB

Cloud computing paradigm is based on the capabilities of the service provider to manage the infrastructural resources in such way that computing and storage demand for incoming jobs can be satisfied. Nowadays, such capabilities rely on software stacks that ease the control and monitoring of used resources. The main objective for such tools is the selection of the most suitable set of nodes to satisfy the application requests. However, current tools use allocation policies that are unaware of the underlying infrastructural (energy) power consumption. Moreover, the used strategies are not designed to consolidate workloads, thus providing solution for the allocation of resources that are not efficient from the energy perspective.

Recently, cloud computing has seen an explosion of applications demanding for deep learning capabilities. In order to support such demand, various algorithms have been ported on different platforms (CPUs, GPUs). However, most of such platforms, although providing good performance, are not very efficient in terms of (energy) power consumption. A big improvement is thus represented by flexible (reconfigurable), as well as energy efficient hardware supporting the acceleration of such algorithms. Furthermore, the most of the past research works presented designs and implementation for non-state-of-the-art algorithms.
ISMB has developed a directive reconfigurable antenna with a steering capability that can cover a sector of 90° (±45° around the broadside direction). This antenna allows, according to the use case requirements specified in D3.1 (i.e. 700 meters) to reach this communication distances. This already represents an improvement over the conventional antenna designs.

3.4.4 Nallatech

As part of the offload capability being developed to support the ULP device and to demonstrate interoperability between the ULP system and the datacenter, the Moonshot server and FPGA are being used to accelerate convolutional neural networks (CNN). In this scenario a more complex CNN is implemented to address the limited capabilities of the Tiny YOLO network used by the ULP system.

The ULP has limited CPU and memory capabilities and is thus restricted to the Tiny YOLO implementation. Offloading more difficult operations to the more powerful Moonshot with FPGA allows more complex analyses to take place while maintaining a low energy profile and consequently long autonomy time in the ULP device.

CNN is one of the most significant developments in computational techniques in recent years, used for many machine learning applications. As a result, there is significant industry interest with many vendors developing hardware specifically to address this area. The current state of art compute platform for the CNN networks is the Volta range of GPU's from Nvidia, featuring their Tensor Core architecture[3]. This device is capable of 125 TFlops of TensorFlow operations (Reduced floating precision operations optimised for CNN processing). GPU's are used for both training CNNs and inference. It is the later that will be implemented on the FPGA as part of the OPERA project. Here, there GPU is not as efficient as it is for training, running approximately at 60% efficiency. Significantly the power consumption of a Volta GPU with NVLINK is 300 Watts.

[3] https://www.nvidia.co.uk/data-center/tensorcore/

	Tesla V100 PCIe	Tesla V100 SXM2
GPU Architecture	NVIDIA Volta	
NVIDIA Tensor Cores	640	
NVIDIA CUDA® Cores	5,120	
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS
GPU Memory	16 GB HBM2	
Memory Bandwidth	900 GB/sec	
ECC	Yes	
Interconnect Bandwidth	32 GB/sec	300 GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink
Form Factor	PCIe Full Height/Length	SXM2
Max Power Comsumption	250 W	300 W
Thermal Solution	Passive	
Compute APIs	CUDA, DirectCompute, OpenCL™, OpenACC	

Figure 4 Nvidia Volta specification

Figure 4 lists the specification of the two Volta variants available.

3.4.5 ST

The actual state of the art, in line with the research conducted by STM in OPERA and in other parallel activities, is the Orlando R&D System on Chip. It is the first version of a CDNN processor, still in progress and with foreseen improvements. The use of this SoC in the OPERA context is planned for the final prototype, and if possible also in the next version if its integration in the global system can be completed in the needed timeframe.

3.4.6 Technion

Related work by Lizy K. John's group from UT Austin developed a learning-based prediction of performance and power consumption on in-order ARM processors (represented via the gem5 cycle-accurate simulator) from performance statistics obtained through hardware counter measurements on a host processor (e.g., Intel Core i7-920 with 24GB of RAM). Their first work, (Zheng, Learning-based Analytical Cross-Platform Performance Prediction, 2015), aggregated the performance counters of the whole program runtime and got error of more than 40% for some benchmarks. Later work, "Accurate Phase-Level Cross-Platform Power and Performance Estimation" (Zheng, John, & Gerstlauer, Accurate Phase-Level Cross-Platform Power and Performance Estimation, 2016) extracted the counters at the granularity of program phases (i.e., 500-20,000 basic blocks of LLVM intermediate representation) and showed prediction accuracy of over 97% (worst case) at speeds over 500 MIPS for both performance and power. They also showed that the average accuracy of per-phase prediction may be low; however, accuracy is considerably higher when aggregating all the phases to obtain the overall performance due to averaging effects.

3.5 HOW ARE YOU GOING TO MOVE PAST THIS STATE OF THE ART?

3.5.1 HPE

With the work delivered on OPERA, HPE gathered experience and valuable feedback from other consortium members. HPE is developing a new, more flexible way to integrate heterogeneous compute to provide even better power efficiency. HPE is also working with application providers to leverage offloading capabilities.

3.5.2 IBM

We are implementing post-copy migration which means the container resumes operation on the target machine very early in the migration. Most of the memory belonging to the container is left on the source machine and copied on-demand. This makes migrations very fast compared to pre-copy and especially naive migration methods. The network utilization is relatively low as compared with pre-copy since each page is copied only once. There are two drawbacks to this method. The first, is that the container runs slower after resuming on the target machine. This is due to the fact that the memory pages that are not yet present must be copied over the network, essentially stalling the thread that needs the memory page. As copying of the working set nears completion, page faults become rarer which serves to amortize the cost of a page fault over a longer period of execution which means it has less of an effect on the running time. We are attempting to address this problem by using special hardware to accelerate the page fault handling. The second drawback is the exposure to catastrophic errors. If the target host goes down while the migration is in progress, it is likely not possible to cancel the migration and resume execution on the source machine. This is similar to the case if the network between the hosts goes down (or the source host itself) during the migration. In that case, the page faults would not be able to be resolved, effectively crashing the application since it would be able to continue execution without its missing memory.
The runtime drawback is thus actively targeted by the OPERA project by utilising a high bandwidth low latency interconnect. The mentioned drawback concerning fault tolerance will remain, but the use of high quality hardware will in a extremely low probability of such a failure occurring.

3.5.3 ISMB

To progress beyond the current state-of-the-art, we are working on three main aspects. First, we are refining the proposed workload allocation policies to get an effective energy-aware tools for the management of data centers. This is based on the integration of a two-steps approach: we use an energy-aware greedy strategy for the initial placement of application tasks on the data center, by selecting the most power-efficient nodes, and then by consolidating the overall workload in the data center. Second, we are working with Nallatech to accelerate a state-of-the-art CNN, specifically the YOLO9000 which has been demonstrated to provide high levels of accuracy for image classification. We consider to this end a very efficient hardware accelerator, based on a state-of-the-art FPGA device.

Third, we also progress over the current state-of-the-art designs for the RF antennas. To this end, we proposed a new antenna prototype with a geometry that allows to selectively cover a 360° in azimuthal plane by steps of about 60°. This feature should overcome the steering limitation posed by the state-of-the-art antenna (Sec. 3.4.3), by paying in a shorter short communication distance (according to the requirements defined in D7.2). This requirement allowed also to reduce the dimensions of the antenna.

3.5.4 ST

The results obtained with the Orlando SoC, even if it is not yet a product available on the market, is that this solution is capable to compete with far more complex solutions based on servers and GPU accelerators with an outstanding power efficiency profile.
The acceleration of the FPGA on the server side is also demonstrating that the heterogeneous approach, at different scales of magnitude, is a winning choice.

3.5.5 Technion

The main weakness of the previous works by Lizy K. John's is that it concerns only (relatively) simple in-order ARM processors. Our main challenge will be to show that similar ideas are possible to implement HW-counters-based performance prediction of x86-64 platforms.

To this end, the Technion develop a framework that is able to (1) control the resource allocation, and (2) measure runtime and energy to plot performance curves on x86-64 platforms. Each curve describes the behaviour of a given application on a given system (CPU + OS) as a function of the allocated resources. As previously noted, the generated curves can guide the placement algorithm developed by ISMB.

4 POTENTIAL MARKETS

Next to the industries mentioned already in D2.5 and D2.6, thinking about industries, extra markets have been identified by the OPERA consortium:

4.1 MIGRATING CONTAINERS

IBM explicitly mentioned applications for this technology anywhere that containers are being migrated. It may be the case as we are showing in the VDI use case of OPERA, that migration is needed for load balancing. However, there are other valid uses for migration such as machine evacuation (moving all running applications in order to perform maintenance on the machine).

Post copy migration might also offer a way to migrate highly active applications (VM's/containers). The issue that these applications have with pre-copied memory is that the memory pages change so often that a stable identical copy on the target host is never achieved. The pre-copy process in this case will keep copying changed memory pages and in most of these cases will time-out without migration.

4.2 ARTIFICIAL INTELLIGENCE / DEEP LEARNING

Artificial Intelligence is at the center of most of the technological conferences and is seen as a way to improve many aspects of human condition in a certain extent. Companies but also countries are investing a lot of resources on AI to become an undisputed leader in their area.

It will take time for humanity to develop the knowledge and technology to build a "full AI" system, able to understand and react to complex situations in any domain. We are already doing gigantic progresses on specific domains such as image recognition with new Artificial Neuronal Network able to classify images better than any human.

Deep Learning, which can be defined as a sub-category of Machine Learning, is leveraging Convolutional Neuronal Network (CNN) to mimic the human brain behaviour. OPERA is leveraging CNN for the traffic monitoring use case to understand the road conditions and increase the security of drivers by taking actions such as generating alerts.

This use case itself could already be sold as a solution to cities and roads operators. It allows to automatically process the output of 1000s of cameras and only display "problematic scenes" to the operator, potentially with pre-programmed actions.

In the future, merging these capacities with autonomous driving vehicles, which embed 10s of sensors, would extend the detection capabilities and dramatically reduce the number of accidents. To stay in the image recognition domain, many other domains are investigating or already implementing Deep Learning to greatly improve accuracy and performance.

4.3 VIDEO SURVEILLANCE

Safety: Leveraging video surveillance cameras input coupled with Deep Learning system enables authorities to identify or tag and follow people. A Box or luggage left behind in a public place can potentially be a threat and needs to be detected. By using input from multiple cameras, this box can be identified, and a volume can be calculated.

Road safety: The use of autonomous low power systems with embedded intelligence will enable road management to vastly increase the amount of road that is monitored without swamping road managers with useless video. In current operations, all cameras are monitored by human eyes, severely limiting the amount of video streams that can be monitored. The use of intelligent cameras that evaluate the image first and only transmit when necessary will greatly increase the number of cameras that can be handled by a single road manager. This new of working coupled with the autonomous character of the OPERA solution will enhance the ability to equip many more road segments with video surveillance.

Medical image: it has been demonstrated that a well-trained and designed Neural Network has a better accuracy to detect tumours than the best specialists on the planet. You still need the expert to provide a good reading and explanation, but the detection phase can easily be automated.

Going from a traditional CNN to a Binarized Neural Network approach increases inference speed and also reduces the size of the model, allowing its use in low-power devices. Using binarized weights rather than floating point has negligible influence on the accuracy but reduces the size of the model by 32 and increases speed by 16x, when running on the FPGA.

4.4 INDUSTRY 4.0

Industry 4.0: processing video feeds installed in a factory to increase security, product quality and optimized workers efficiency. On this last topic, one company working on truck maintenance has been using this method to improve by 50% efficiency. To achieve this, they fed video feeds to a CNN to track every move in the factory.
The output highlighted frequently used routes and provided guidance on how to reduce unnecessary movements and result in reorganizing the working place.

4.5 DATA CENTRES

With the potential for energy savings in data centres, the data centre industry can be seen as an interesting prospect for OPERA innovative and energy saving products.

4.5.1 Example in the Dutch market

This section discusses the question: What is the addressable energy use for the OPERA products?
In order to answer this question, Certios conducted literature research as well as a survey of a number of datacenters varying in size from very small to large.
A report published by CE Delft created on the request of the Dutch government organization RVO (Afman & Scholten, 2016) describes the estimation of the total energy use by ICT in the Netherlands.

Aside from home use, the report attributes a substantial energy use to commercial and private datacenters, also referred to as the supply and demand side, respectively. The distinction between the two is less based on any technical reasoning but rather on the availability of data, the commercial datacenters being easier and therefore more accurately characterized. Combining trend data with the reported energy use of 1.36 TWh for the year 2013, the energy use of servers in private datacentres is estimated at 0.9 TWh (including an average PUE of 1.7[4]) for the year 2017.

The energy use in the same period for commercial datacenters can be estimated from 2 different reports: The first is from the company CE Delft in 2014 (CE Delft, 2014). In this report, the 2017 energy use by commercial datacenters is estimated at 2.7 TWH (1.9-3.5 (error margin)), including a slightly better than private datacentres PUE of 1.4). It is important to note that this estimate also contains the energy for storage and network devices, which has to be excluded for the OPERA addressable market.
The 2014 CE Delft estimated is corroborated by a more recent publication of the Dutch Datacenter Association (DDA) which published commercial Datacenter power draw in MW (Vermeulen, 2016).

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

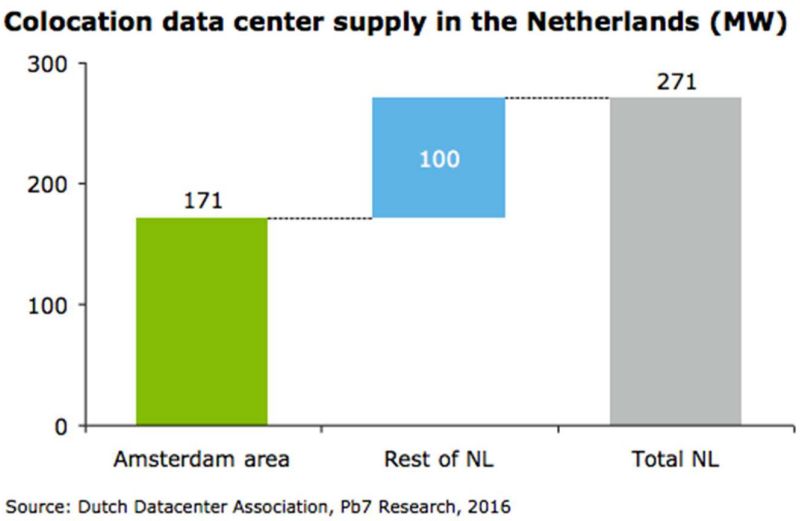

Figure 5 271 MW in power means 2.4 TWh/year

Simple arithmetic yields that this power draw corresponds to 2.4 TWh. The number is also deemed valid for 2017 (Peters & Dijk, 2016).
Given the source, namely "self-reporting" by datacenters and the recent nature of the numbers, 2.4 TWh for 2016 will be used, this again is including facilities, storage and network equipment, that will have to be excluded.

4.5.2 Attribution of energy use to servers

This section answers the question, what portion of the energy estimate is associated with servers? This question is important since servers are a target for replacement with OPERA technology.
The afore mentioned report (Afman & Scholten, 2016)) suggests that 67% of data center energy use is directly related to servers. More recent research in the USA, based on sales figures from hardware vendors, see "United States Data Center Energy Usage Report (Shehabi, 2016)", give an estimate of 77%.
Given the quality of the source data, OPERA will use the 77% estimation, resulting in the estimated commercial data center energy use attributed to servers in 2017 of 1.9 TWh. The 1.9 TWh includes facility power (cooling) which is logical given the near linear relation between ICT and facility energy use.
Given the earlier estimate of 0.9 TWh for servers in private datacenters, the total addressable energy use in the Netherlands was 2.8 TWh for 2017. Using a KWh price of 0.15 EUR, the annual server energy costs in the Netherlands are 420 M EUR.

In the coming period, these estimations will be refined to include the adoption of server (host) power management and cluster workload management to finally arrive at an estimated energy savings potential for the OPERA solution, and the subsequent savings in energy costs, which will drive the market.
These will then be expanded to the entire European market using what numbers are available about the total size of this market in terms of energy used and costs (Avgerinou, Bertoldi, & Castellazzi, 2017).

5 MARKET ANALYSIS

In this section we will have a look at some market analyses that we could add to the earlier analyses of D2.5 and D2.6.

5.1 MIGRATING CONTAINER

With reference to the application mentioned in section 4.1, as far as we know, this is the first implementation of post-copy migration for containers. Post-copy migration for virtual machines already exists (implemented as part of QEMU) and is used in production systems. We believe that post-copy migration for containers has the same potential and can be used in new or existing container cloud installations for the purposes mentioned.

5.2 ARTIFICIAL INTELLIGENCE / DEEP LEARNING

The potential market of Artificial Intelligence enablers is growing faster in these months, due to the investment in this research area of the main actors in the field. Companies like ARM, Movidius, (an Intel company), Ceva, STMicroelectronics, like many others, are investing in the research of small and efficient processors for the implementation of Neural Networks, demonstrating the high level of interest in these applications. Other companies, like Facebook, Google, Nvidia, are investing in architectures for more powerful and performant processor for server-based applications. Also, in this case the research conducted in OPERA for the acceleration of CNN in the server applications domain, with the heterogeneous accelerator based on FPGA, is targeting this new field.

Deep Learning market growth will be double digits for the coming 10 years. Analysts are talking about reaching 110Bn $ by 2027 just for North America. One of the company taking the most advantage of this enormous growth is Nvidia with their GPGPUs. Today, they are seen as the reference hardware platform for training Neural Networks and doing inference.

Google, with their TPU (Tensor Processing Unit) has been investing into hardware design and demonstrated during AlphaGo challenge that 176 GPU (44KW) could provide better results than 48 TPU (1.9 KW). Google also demonstrated that the work on the algorithm is not completed and is the one with potentially the biggest impact. Their last NN version, using 4TPU (160W), beat 100:0 the previous NN running on 48 TPU. There is a good chance that new types of hardware targeting high precision, complex and fast inference emerge in the coming years when some of the network designs are stabilized.
Another hardware implementation, which will be going after the volume/edge market is the dedicated CNN ASIC, such as the one located in the ULP camera from ST we are using in the OPERA project.

We now have a “AI” processor into each last-generation high end smartphone, and there is a good chance this will extend to more and more devices in the coming years.

FPGA is still in a pretty good position for this market, especially with some of the researches conducted on binarizing the Networks, which dramatically reduce the memory requirements. It also provides another level of flexibility which greatly extends the life of the hardware platform. When a new idea of optimizing Neural Network is discovered, it is very easy to put it in practice.

FPGA is the technology selected by Microsoft for some of their AI use cases and we see a massive interest of universities in Europe which are starting to build large FPGA clusters for applying CNN to many applications (industrial design, aerodynamics, earthquake prediction).

5.3 DATA CENTRES

The market for data centres is developing quickly, and as announced in 4.5.2, in D2.8 the analysis for the potential market will be extended. After seeing the market potential in one country (the Netherlands), the analysis needs to be extended to the rest of Europe as well, which will be done in D2.8.

6 IMPACT ENVISAGED

In this section the impact envisaged will be described, for the several potential markets the consortium partners will target after the OPERA innovative products and services will have become marketable. The descriptions in the next sections are an addition to the impact envisaged sections of D2.5 and D2.6.

6.1 VIDEO SURVEILLANCE

Regarding the impact envisaged related to the introduction of the CNN Orlando SoC in the traffic monitoring use case, it has been presented to several potential investors and customers in the sector of video surveillance. The feedback has been very positive, enabling new opportunities for applications not considered in the past with the traditional approach for video processing devices.
At the moment the impact of these experiments is already perceived, and it will become stronger in the next period, with the presentation of complete results in terms of quality of detection and power efficiency.

6.2 DATA CENTER

The impact thanks to energy savings will be quite impressive and will have an impact in the data centre industry (see section 4.5).

To give an idea of the energy cost structure in data centres with regards to today's server equipment, there is a way to calculate the annual energy costs of IT equipment ((Bashroush, 2018). This is to illustrate the impact of energy costs. Let's work out an illustration for a small imaginary commercial data centre offering hosting services:

To estimate the IT equipment annual energy costs, one needs the following data:

Item for calculation	Assumption
Number of servers [S] and server details (make and model)	400 Servers, 3-5 years old
Idle [Ei] and 100% utilisation [Eu] power consumption (from SpecPower DB) in Watts	[Ei] 150W [Eu] 250W
Average annual utilisation of servers [u] (this can be read from the hypervisor or monitored over a sample period) (as a unit, i.e., 0 - 1)	10%
Energy cost per kWh [c]	0,15 EUR/KWh

Table 3 Impact of saving energy in data centres

Then, you can calculate the annual energy consumption cost for every server to be:

$$S * (E_u * \mathbf{u} + E_i * (1 - \mathbf{u})) * 8.76 * \mathbf{c}$$

$$= 63{,}115 \text{ EUR}$$

The average PUE in the data centres that are participant of the EU Code of Conduct for Energy Efficient Data Centres[5] is 1.7.

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

Figure 6 PUE formula

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

This means that for every 1 Watt a server draws (IT Equipment Power), 1.7 Watts is drawn (Total Facility Power) in order to arrange for its facilities like lighting, cooling, power distribution in the data centre. The total savings in this simplified data centre example is 1.7 * 63,115 EUR = 107,296 EUR.

This example indicates the energy saving impact when the ratio *Server Performance per KWh* will increase. This can be achieved by OPERA, through:

- reduction of idle power
- increase of utilization
- reduction of utilization energy usage.

This will be elaborated more in the final iteration D2.8.

7 REFERENCES

Afman, M. R., & Scholten, T. (2016). *Energiegebruik ICT in Nederland 2013, trendontwikkeling 2020 en 2030.* CE Delft. Delft: CE Delft.

AMD Inc. (2013). AMD64 Architecture Programmer's Manual, Volume 2.

artemis-emc2. (2016, 8). *artemis-emc2*. Retrieved from http://www.artemis-emc2.eu

Avgerinou, M., Bertoldi, P., & Castellazzi, L. (2017, sept 22). Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency. *MDPI*. Retrieved from www.mdpi.com: www.mdpi.com/1996-1073/10/10/1470/pdf

axiom. (2016, 8). *axiom*. Retrieved from http://www.axiom-project.eu

Barr, T. W. (2010). "translation caching: Skip, don't walk (the page table)". *In ACM/IEEE International symposium on Computer architecture (ISCA), pages 48-59, 2010.*

Bashroush, R. (2018). A Comprehensive Reasoning Framework for Hardware Refresh in Data Centres. *IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING*, 1-14.

Bittman, T. J., Dawson, P., & Warrilow, M. (2015, 07 14). *Magic Quadrant for x86 Server Virtualization Infrastructure.* Retrieved from Gartner: https://www.gartner.com/doc/3093222/magic-quadrant-x-server-virtualization

CE Delft. (2014). *Nederlandse commerciële datacenters 2014-2017 Nieuwbouwplannen en ontwikkeling energiegebruik.* Delft: CE Delft.

Consortium, O. (2015). *OPERA_PartB_section1-3_2016_08_05.* OPERA. ST.

Distributed Management Task Force inc. (2017, 03 30). *redfish developers hub homepage*. Retrieved from redfish developers hub: http://redfish.dmtf.org

DRedBox. (2016, 8). *DRedBox*. Retrieved from http://www.dredbox.eu

Ena-HPC.org. (n.d.). Retrieved 2017, from http://www.ena-hpc.org/index.html

exanest. (2016, 8). *exanest*. Retrieved from http://www.exanest.eu

Gartner. (2016, april 27). *G00301050*. Retrieved from Cool Vendors in Cloud Infrastructure.

Gartner. (2016, april 26). *G00301053*. Retrieved from Cool vendors for compute platforms.

Gartner. (2016, 5 13). *G00301565*. Retrieved from Cool vendors in sustainability.

Gartner. (2016, 5 11). *G00302247*. Retrieved from Cool vendors in internet of things analytics.

Gartner. (2016, July). *Gartner Hype Cycle for Data Science.*

Gartner. (2016). *Internet of Things Primer for 2016 doc.nr.G00303518.* Gartner.

Gitlab. (2016, 8). *aparapi-ucores*. Retrieved from https://gitlab.com/mora/aparapi-ucores

HPE. (2017, 03 30). *HPE Moonshot system family guide*. Retrieved from HPE Moonshot system: https://www.hpe.com/h20195/v2/getpdf.aspx/4AA4-6076ENW.pdf

Intel Corporation. (2015). Intel 64 and IA-32 Architectures Software Developer's Manual, Volume 3B.

Intel Corporation. (2017). *5-Level Paging and 5-Level EPT.*

Intel Corporation. (2017). *Stratix 10 Product Table.* Retrieved 2017, from www.altera.com: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/pt/stratix-10-product-table.pdf

Intel Corporation. (n.d.). *Intel Arria 10 Product Table.* Retrieved from https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/pt/arria-10-product-table.pdf

Intel. (n.d.). *Intel Arria Product Table.* Retrieved from www.altera.com: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/pt/arria-10-product-table.pdf

Kim, W. C. (2005). Blue Ocean Strategy. Harvard Business review Press.

Kolluri, S. (2015, 10 15). *UltraScale Architecture Low Power Technology Overview*. (Xilinx, Producer) Retrieved 2017, from www.xilinx.com: https://www.xilinx.com/support/documentation/white_papers/wp451-ultrascale-pwr-reduction.pdf

Nallatech. (2017). *Intel Stratix 10 FPGA – Nallatech 520 FPGA Acceleration Card.* Retrieved 2017, from Nallatech.com: http://www.nallatech.com/store/fpga-accelerated-computing/pcie-accelerator-cards/nallatech-520-compute-acceleration-card-stratix-10-fpga/

NVIDEA. (2016, august 16). *correcting some mistakes.* Retrieved from https://blogs.nvidia.com/blog/2016/08/16/correcting-some-mistakes/

Nvidia. (2017). *WELCOME TO THE ERA OF AI.* Retrieved 2017, from www.nvidia.co.uk: https://www.nvidia.co.uk/data-center/tesla-v100/

Parker, M. (2017). *Understanding Peak Floating-Point Performance Claims.* Retrieved from www.altera.com: https://www.altera.com/en_US/pdfs/literature/wp/wp-01222-understanding-peak-floating-point-performance-claims.pdf

Peters, M., & Dijk, A. v. (2016). *Dutch Digital Infrastructure 2016.* Deloitte.

Shehabi, A. e. (2016). *United States Data Center Energy Usage Report LBNL-1005775.* Berkeley: Ernest Orlando Lawrence Berkeley National Laboratory.

Spear, K. (2017, 1 3). *Cisco Supports Redfish Standard: API Enhances UCS Programmability*. Retrieved 2017, from Cisco Blogs: https://blogs.cisco.com/datacenter/cisco-supports-redfish-standard-api-enhances-ucs-programmability

Standard Performance Evaluation Corporation. (n.d.). *www.spec.org*. Retrieved 2017, from https://www.spec.org/cpu2006/Docs/

Techtarget.com. (n.d.). Retrieved 2017, from http://searchstorage.techtarget.com/definition/cold-storage

Vermeulen, P. (2016). *Dutch Data Center Report 2016.* PB7. DDA.

Villa, N. (2016, august 30). *ICT and sustainability: unleashing innovation.* Retrieved from ICT4S keynotes: http://2016.ict4s.org/program/keynote-speakers/nicola-villa/

Vineyard. (2016, 8). *Vineyard*. Retrieved from http://www.vineyard-h2020.eu

Xilinx Inc. (2016, Feb 19). *Xilinx Corporate and ortfolio Transformation.* Retrieved from youtube.com: https://www.youtube.com/watch?v=9jtvNPySPkc

Zheng, X. (2015). Learning-based Analytical Cross-Platform Performance Prediction.

Zheng, X., John, L. K., & Gerstlauer, A. (2016). Accurate Phase-Level Cross-Platform Power and Performance Estimation. *Accurate phase-level cross-platform power and performance estimation.* Austin, TX, USA.