The Potential Influence of Workload Management Across Heterogeneous Server Systems on Datacenter Energy Use and Power Draw.

D.H.Harryvan
Certios BV
Amsterdam, The Netherlands
dirk.harryvan@certios.nl

R. Chamberlane
Nallatech
Molex
Cumbernauld, UK
richard.chamberlain@molex.com

Alberto SCionti
Istituto Superiore Mario Boella
(ISMB)
Torino TO, Italy
scionti@ismb.it

Giulio Urlini
Advanced Systems Technology
STMicroelectronics
Agrate Brianza (MB), Italy
giulio.urlini@st.com

O. Terzo
Advanced Computing and
Electromagnetics (ACE)
Istituto Superiore Mario Boella
(ISMB)
Torino TO, Italy
terzo@ismb.it

***Abstract*—Energy efficiency is a key part of the European energy policies and 2020 climate targets. Project OPERA is working to create an energy aware workload manager for heterogeneous systems that will allow microservices to migrate between systems with differing instruction set architectures. The Energy savings potential of such technologies is enormous and is estimated at 47 TWh per year in Europe, 95% of the energy consumed by servers in Europe. The impact of such technologies on datacenter operations is profound. Significant and fast variations in power draw over time are expected, a fact that operators need to consider when retrofitting or designing new facilities.**

Keywords--workload management; heterogeneous server systems; project OPERA; energy savings

I. INTRODUCTION

Energy efficiency is a key part of the European energy policies and the 2020 and 2030 energy and climate targets include a specific part for energy efficiency [1]. Throughout the EU, policies and measures have been introduced to improve energy efficiency in every sector of the economy, including ICT and data centers [2]. For a number of years now, the datacenter industry has put tremendous effort in optimizing the Power Usage Effectiveness (PUE) as the main path to energy savings. From the definition of the PUE [3] it follows that:

$$Total\ datacenter\ energy\ use = PUE\ X\ IT\ energy\ use$$

The effect of best practices, such as the European Code of Conduct on Data Centre Energy Efficiency [2], has been published and a gradual improvement in the PUE, ~~use~~ has been reported [4]. This improvement has however, not led to a decrease in total energy use (see figure 1) and, given the reporting period, it is not expected that major improvements in the average PUE and total energy use are imminent.

Other than decreasing the PUE, another option for lowering overall energy use is decreasing the energy use of the IT equipment. Since the peak in CPU clock frequencies in 2005, manufacturers have improved system performance in other directions such as developing multicores, dedicated accelerators and massive memory expansions. As a result, computational power has increased by orders of magnitude, while the overall power draw of these systems has dropped or stayed level. The total effect of these improvements is summed up in the so called "Koomey's Law" [5] that states that it takes about 2.7 years for peak-output efficiency to double. Koomey does mention in this article that these 2.7 years is a marked slowdown of the improvements over the rate governing the years prior, which implies that new directions for improving efficiency need to be developed.

An issue with the current computer systems is that in order to truly exploit their energy efficiency, these systems should be running near maximum workload all the time. Even in the most modern systems efficiency drops at low utilization. This was recognized by Jonathan Koomey in 2015 and his article mentioned earlier, introduced "typical use efficiency" as a way to describe these advances; not only in chip efficiency but also in power and workload management.

II. PROJECT OPERA.

The This paper reports on the advances in "energy aware workload management" across heterogeneous systems. This work is conducted under the flag of the "OPERA" project; a research project under funding by the European Union through the Horizon 2020 Research and Innovation Programme under the grant agreement No. 688386 [6]

Workload management in itself is not a new concept, VMware Vmotion technology is a well-known example but as to date has been limited to homogeneous systems; the workload can be moved between nearly identical computers. The focus of OPERA is to broaden this to heterogeneous systems. Heterogeneity literally means "composed from dissimilar parts" and refers to the way that system manufacturers are currently pushing the boundary of performance. The combination of specialized processors such as GPU's and crypto accelerators with more general-purpose

processors with differing instruction set architectures (ISA) like ARM, X86, and Power, has the potential of bringing additional performance at lower energy cost.

The currently most common way of exploiting heterogeneity is through adapting an application so that specific instructions are performed by dedicated hardware components. Examples of this can be found in almost all graphics-oriented application where most of the computations needed for displaying images are handled by the Graphics Processing Unit (GPU). OPERA is advancing this research by developing a system comprised of heterogeneous nodes, in the OPERA case an IBM system equipped with Power8 processors and a HP Moonshot equipped with X86 processors, coupled through a high speed, low latency interconnect based on Mellanox RDMA-enabled Ethernet cards. When this work is completed, the live workload, comprised of microservices, on the OPERA system can be moved without interruption between various nodes with differing ISA's (ARM, Power8, X86) or similar architectures but different manufacturers such as AMD and Intel, driven by the application need and the drive for power efficiency [7]**.** To accomplish this OPERA implemented a remote page fault handling, that will allow applications running on one node to access memory pages on another node regardless of the underlying ISA [8].

III. Energy use in European datacenters

A number of terms that are interrelated are common in the discussions that focus on datacenters and their energy use. Power, PUE, Energy and Energy efficiency and greenhouse gas emissions are interrelated topics, but their relations are important in light of the climate goals of European Union [9].

- 20% cut in greenhouse gas emissions (from 1990 levels)
- 20% improvement in energy efficiency

Greenhouse gas emissions, CO_2 and other chemicals are often expressed in their Global Warming Potential (GWP) relative to that of CO_2 over a 100-yearlong period (GWP100 (EPA, 2018)) Several studies on the full Life Cycle of electronics have been published that include the GWP. A selection of these have been analysed to highlight the parameters specific to the OPERA project and to provide scientific support for choices made in the context of OPERA. As described in the Vestlandsforsking-rapport Report nr. 6/2010 Life Cycle Assessment of Electronics [10], the ISO 14040/44 standards have ensured a rather consistent set of Global Warming Potential (GWP) for the studied products. However, consumer electronics are more difficult than many other product types due to a lack of transparency about sourcing of raw materials and detailed composition of the electronic products by manufacturers. The report nonetheless homogenised several older studies and concluded that based on the survey of published work, recycling and other end–of–life processes constitute a tiny share of the total GWP100 score for consumer electronics.

One of the most recent, high-quality investigations into the total life cycle impact of enterprise servers, is published as work conducted for the Joint Research Committee of the EU (JRC) titled the "Analysis of the material efficiency requirements of enterprise servers" [11]. The report contains a very detailed study of available literature and states that IT manufacturing companies like Fujitsu, Dell and Apple have published environmental reports on enterprise servers. The conclusions drawn in these reports are comparable in the fact that all indicate that the use phase generates about 90% of the total environmental impact, manufacturing contributes from 5 to 7% to the footprint and the remaining impact is due to transport and recycling [12]. Thus, these reports not only support the conclusions in the Vestlandsforsking-rapport nr. 6/2010 that the end-of-life impact is small, but also that the manufacturing stage and therefore the embodied energy in electronics is minor compared with the use phase energy consumption. This information caused the OPERA project to focus only on the use phase final energy use measured in kWh.

From the article "Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency" [4] table 1, we can show the total datacenter use phase final energy use of datacenters in the EU. Combination with the reported and estimated PUE in the same periods we can extract the IT energy use shown in figure 1. The energy used by servers in the datacenter is estimated by Shehabi et al, United *States Data Center Energy Usage Report* [13] to be 77% of the IT energy.

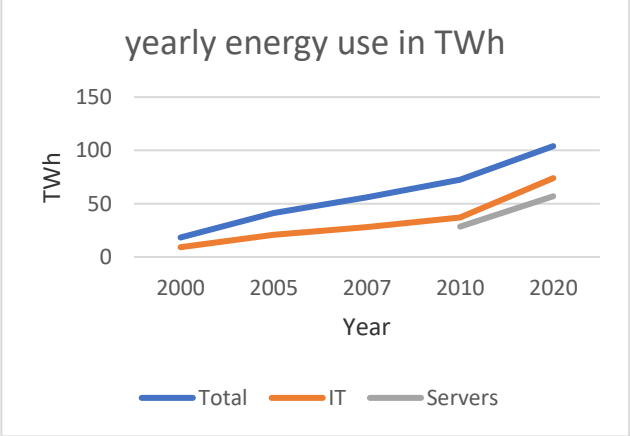

figure 1: energy use in EU datacenters.

From these figures we can conclude that the potential target for OPERA technologies there for represents approximately 50TWh in the year 2018

IV. Expressing overall energy efficiency

Since it is not only unlikely, but also undesirable to reduce the amount of work that is done in the datacenters, the logical step towards reducing energy use is to increase the operational efficiency. The ITU, in its recommendation "Energy efficiency metrics and measurement methods for telecommunication equipment" [14], discusses energy efficiency metrics and states:

- *"The energy efficiency metric is typically defined as the ratio between the functional unit of work and the energy necessary to deliver the functional unit; the higher the value of the metric, the greater the efficiency of the equipment."*

This definition clearly differentiates efficiency from effectiveness. The PUE discussed earlier is a measure of effectiveness as it does not define nor use any unit of work. Although very useful, the PUE is there for unsuited to register progress against the 2020 goal for energy efficiency. The difficult element in defining IT efficiency has always been to define the unit of work. Since the type of workloads being

handled in the datacentre are so diverse, within the OPERA project we have taken a very practical approach that is directly related to the use cases that are part of the project. Since the use case being discussed in this paper is the redesign of a virtual desktop infrastructure, we use the user of the VDI service in defining the unit of work. The functional unit for this metric will therefore be one week of VDI services provided per single user as fraction of the maximum number of supported users.

$$EE_{vdi} = \frac{1\ week\ fully\ operational\ VDI\ (AU)}{Energy\ used\ during\ 1\ week\ per\ user(kWh)}$$

Analysis of the existing configuration yielded the observation that a single HP Proliant DL380 G9 (2X Intel E5-2660 v3 2,66Ghz) production server currently provides VDI services for up to 240 users at a particular installation.

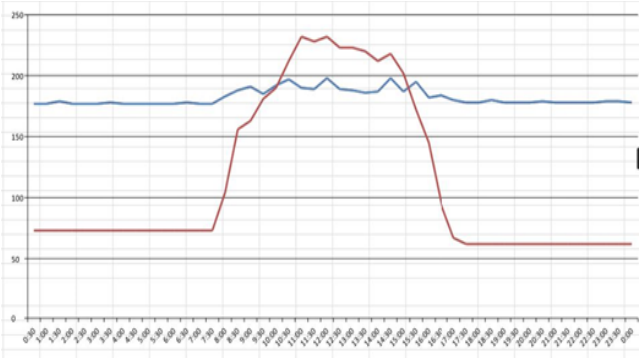

Figure 2 Citrix VDI on Monday, blue line: Power (Watt) and red line users (number)

What is clear from the above profiles is that in the current environment there is very little influence of the number of users on the power draw of the physical server. This load profile is not unique to this use case, a survey amongst various datacenters (unpublished results) conducted by the researchers shows that many installations show similar user load profiles, with loading dwindling to near zero outside office hours. The OPERA project therefore has 2 targets, namely lowering power draw in (near) idle situations and increase the number of users per Watt in loaded situations.

The power recorded is used within the OPERA project for calculation of the baseline energy efficiency:

$$EE_{vdi} = \frac{7\ userweeks}{kWh}$$

The user load profile abbreviated to 64% idle and 36% max user load is used to configure load generators that are used to show the effect of workload management in the OPERA solution.

V. The OPERA heterogeneous solution

The OPERA research project is still underway, but the target solution for the VDI environment will have the following macro design: as shown in the figure below, we want to achieve the cross-ISA configuration using the Mellanox ConnectX-5 card supporting the RDMA protocol to move containers through the post-copy migration method.

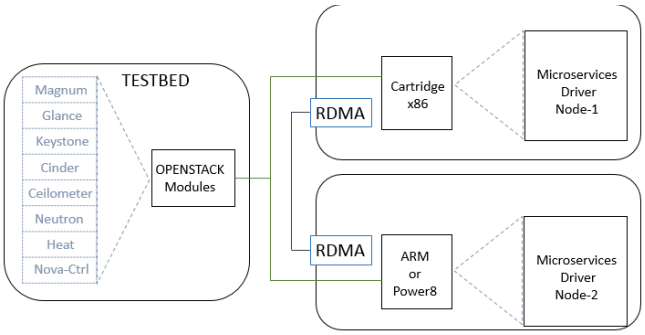

Figure 3: Macro design of the OPERA system

In the design, shown in figure 3, the VDI services will be delivered as a SaaS model wherein the components are microservices, each running in their own containerized environment. Depending on load characteristics, these microservice containers will be moved between node 1 and node 2 to obtain the lowest energy point while maintaining the service levels that are needed to satisfy the users. The actual ISA of node 1 and 2 are not important, thanks to the cross ISA compiler developed for the project, container migration will become possible between X86, ARM, and Power8. In order not to impede performance and allow for quick migrations, OPERA developed a post copy migration system in which the binary on one node can access memory still residing on the other node [8]. The post copy migration is important due to the fact that migration in future implementation will not only happen on the basis of workload, but also on the basis of the phase an application is in. Post copy migration sets stringent demands on the interconnection between the nodes. OPERA developed drivers for the FPGA cards that support RDMA to create a high bandwidth and low latency connection between the nodes.

VI. First results

The current status of the OPERA project resulted in measurements on a single node of the solution. This node consists of a M510 cartridge (Intel Xeon D-1587, 1.7 GHz (16) x86 cores) placed in an HPE Moonshot chassis.

The measurements yielded a loaded power draw of 117W and an idle power draw of 26W. given the load profile these measurements result in the following energy efficiency:

Table Head	**First results**		
	CPU load	***# concurrent end users***	***EE***
CSI Citrix - DL380 (baseline)	55%	240	7
LXC – M510 (single node)	55%	1400	142

Table 1: Energy Efficiency improvement on single node.

The example stated above shows a factor of 20 improvement of energy efficiency, under the assumption that the peak workload does not increase, this translates in a potential 95% energy reduction, Future developments will likely impact the energy efficiency even further through 2 effects. In low load conditions, the workload manager will stop microservices that are no longer needed and migrate (near)idle containers to a node with less capabilities and lower electrical power requirements, this reduces power draw in idle situations. Likewise, the energy efficiency in the loaded conditions are dependent on the optimal loading of the hardware. In high load conditions, the OPERA energy aware

workload manager [7] will either consolidate workload or distribute it over the heterogeneous nodes to ensure sufficient resources are allocated to a service thus lowering power draw under load.

Past studies give an insight in the potential of the heterogeneous workload manager; Chun et. al. showed that with a single ISA, a mix of relative strengths of machines [15] or cores in a single chip [16] can outperform a homogeneous configuration by nearly 40% in terms of energy proportionality and energy-delay product. Subsequent studies show that by combining multiple ISAs in a single chip, energy consumption can be further reduced by an additional 21% [17]. The combined effect would further improve the energy efficiency of datacentre services by a factor of 2.

VII. The Impact on power and energy in datacenters

A survey conducted by the consortium partner Certios showed that the adoption of workload and power management techniques in datacenters is currently low. The survey also confirmed the 5 years average operational lifetime of servers indicated in the American report [13]. Both these observations support the conclusion that the potential for energy savings is enormous.

The measurements shown indicate that huge energy savings are attainable through adoption of workload management in combination with heterogeneous servers. The potential constitutes 95% of the server energy use, translating into over 70% of total datacenter energy. These estimations are done on the basis of the current IT workload, growth in this workload is highly probable, damping the effect on the energy use. It is important to realize that the savings projected are on energy, while a major design criterium for datacenters is Power. Peak power draw is much less impacted, and workload management will cause power draw to closely follow user behavior over time and thus result in power fluctuations that are much stronger than currently the case. Since these technologies can also be adopted by users, datacenter operators need to closely monitor the datacenter infrastructure and prepare for deeper fluctuations in electrical power draw for the coming future.

Acknowledgment

The authors wish to acknowledge the .consortium partners for providing the data that was used for the estimations in this paper. We invite interested parties to visit the OPERA website [6], results and documents will be published there after the project closes at the end of 2018.

VIII. References

[1] European Commission, „Energy Efficiency,” 10 26 2016. [Online]. Available: : https://ec.europa.eu/energy/en/topics/energy-efficiency.

[2] European commission, „2018 Best Practice Guidelines for the EU Code of Conduct on Data Centre Energy Efficiency: Version 9.1.0,” 2018. [Online]. Available: https://ec.europa.eu/jrc/en/publication/2018-best-practice-guidelines-eu-code-conduct-data-centre-energy-efficiency-version-910.

[3] The Green Grid, „WP49-PUE A Comprehensive Examination of the Metric,” The Green Grid, https://www.thegreengrid.org/en/resources/library-and-tools/20-PUE%3A-A-Comprehensive-Examination-of-the-Metric, 2014.

[4] P. B. a. L. C. Maria Avgerinou, „Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency,” *Energies 2017, 10, 1470,* p. 18, 2017.

[5] S. N. Jonathan Koomey, „moores-law-might-be-slowing-down-but-not-energy-efficiency,” IEEE spectrum, 2015. [Online]. Available: https://spectrum.ieee.org/computing/hardware/moores-law-might-be-slowing-down-but-not-energy-efficiency. [Geopend 24 8 2018].

[6] OPERA, „OPERA home page,” 1 march 2018. [Online]. Available: http://www.operaproject.eu/.

[7] P. S. A. N. J. R. Ruiu, „Workload Management for Power Efficiency in Heterogeneous Data Centers,” *The 10th International Conference on Complex, Intelligent, and Software Intensive Systems,* 2016.

[8] J. Y. R. Nider, „Remote page faults with a CAPI based FPGA,” *SYSTOR ’17 Proceedings of the 10th ACM International Systems and Storage Conference – Haifa, Israel,* 2017.

[9] EC, 2020 climate & energy package, „2020 climate & energy package,” 2018. [Online]. Available: https://ec.europa.eu/clima/policies/strategies/2020_nl.

[10] O. Andersen, A. S. Andrea en H. J. Walnum, „Life Cycle Assesment of Electronics. Ugelstad-particles in Ball Grid Array and Chip Scale Packaging,” Sogndal, 2010.

[11] L. T. Peiró en F. Ardente, „Environmental Footprint and Material Efficiency Support for product policy,” Ispra, 2015.

[12] M. Stutz, „Electronics Goes Green 2012+ (EGG), 2012,” in *Carbon footprint of a dell rack server*, Berlin, 2012.

[13] A. e. a. Shehabi, „United States Data CenterEnergy Usage Report LBNL-1005775,” Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, 2016.

[14] ITU, „Energy efficiency metrics and measurement methods for telecommunication equipment,” ITU, Geneva, 2014.

[15] B.-G. Chun, G. Iannaccone, G. Iannaccone, R. Katz, G. Lee en L. Niccolini, „An energy case for hybrid datacenters,” 2010.

[16] R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan en D. M. Tullsen, „Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power Reduction,” Washington, 2003.

[17] A. Venkat en D. M. Tullsen, „Harnessing ISA diversity: design of a heterogeneous-ISA chip multiprocessor," Proceeding of the 41st annual international symposium on Computer architecture (ISCA '14)., 2014.

[18] EPA, „understanding global warming potentials," 2018. [Online]. Available: https://www.epa.gov/ghgemissions/understanding-global-warming-potentials.