




CONTENTS

Welcome	4
The Center	5
Faculty Members	7
Industry Affiliates	10
Current EcoCloud Affiliates	13
We need to dramatically change the way we build and use computers	14
A true measure of datacenter consumption	16
Harnessing light	19
Getting control over our data centers: Heating Bits	22
Divide and conquer: breaking memory up into compartments	24
A large language model for medical knowledge	26
Predicting the future with CloudProphet	28
Contact	30



WELCOME TO ECOCLOUD



Welcome to EcoCloud, the only academic center of its kind, promoting the development of IT infrastructure and cloud computing technology as key enablers for a sustainable society.

On the one hand, this mission requires the effective use of cloud computing to preserve biodiversity, natural resources, and climate in the years to come. We need digital twins for smart cities, sustainable artificial intelligence technologies, energy-aware scientific computing.

On the other hand, we are targeting the definition of a new circular economy of large IT infrastructures or datacenters to minimize carbon footprint and long-term environmental damage.

Our mission is to provide support in networking the EPFL community with key industrial players in the IT sector to enable environmental sustainability for our digital world, thanks to IT. In so doing, we aim to address the major national and global IT challenges that affect us all. Moreover, our success synergistically builds on strong support for, and a growing appreciation by, private industry.

With our strong emphasis on industry collaboration and technology transfer, the Industry Affiliates Program (IAP) aims to build long-term partnerships founded on research collaborations, large-scale University-Industry partnered research grants, PhD student dissemination activities with industry, internships, fellowships, executive education and more.



THE CENTER AND ITS FACILITIES

On January 1st, 2022, Professor David Atienza took over the direction of the EPFL EcoCloud Centre from predecessor, Professor Babak Falsafi. EcoCloud's scientific mission has been expanded with a strong new focus on fundamental research and education in the domain of sustainable cloud computing.

"Historically, EcoCloud's main focus has been to deliver technologies jointly with top companies in the information technologies (IT) sector to help them optimize the large cloud computing infrastructure of public cloud systems," says Atienza. "We are now focusing on the whole IT ecosystem to develop sustainable multiscale computing from the cloud to the edge," he adds. "Our goal is to rethink the whole ecosystem and how we can provide IT solutions that can make computing more sustainable. In particular, the goal is to optimize the used resources for computing to minimize the environmental and social impact of IT infrastructures and practices. This includes the monitoring of materials, energy, water as well as other rare resources, and the creation of a circular economy for IT infrastructure, considering the impact of electronics on the environment from production to the recycling of cloud computing components."

IT infrastructure as enabler for a sustainable society

"In collaboration with the School of Engineering (STI), the School of Computer and Communication Sciences (IC), the School of Architecture, Civil and Environmental Engineering (ENAC), and the School of Basic Sciences (SB) we have defined multi-disciplinary IT application pillars or directions that are strategic for them," says Atienza.

Four multi-center discussions, and multiple projects kicked off in 2022 in the following research areas: energy-constrained and sustainable deep learning (in collaboration with the Center for Intelligent Systems (CIS) and the Center for Imaging), computational and data storage sustainability for scientific computing (in collaboration with the Space Center and the Energy Center), sustainable smart cities and transportation systems (in partnership with the FUSTIC Association, CIS and CLIMACT Center) and energy-constrained trustworthy systems, including Bitcoin technology (in collaboration with the Center for Digital Trust).

In addition to its multi-center research projects on specific applications, EcoCloud is also working on fundamental technologies to enable sustainable IT infrastructures, such as minimal-energy computing and storage platforms, or approaches to maximize the use of renewable energy in data centers and IT services deployment.

Moreover, EcoCloud will keep developing and strengthening, in this new era of sustainable cloud computing research,

THE CENTER AND ITS FACILITIES

its collaboration of many years with IT partners through its Industrial Affiliates Program (IAP), such as Hewlett Packard, Intel, IBM, Huawei and Oracle, who have confirmed their interest in continuing to collaborate with the center on its new research topics.

A new research facility on sustainable computing

We are creating an experimental facility dedicated to multi-disciplinary research on sustainable computing at EPFL," says Atienza. In this facility, EcoCloud provides a detailed IT monitoring infrastructure (e.g., performance, power, energy, temperature, etc.), supported by specialized IT personnel, to assist and support the EPFL laboratories in performing tests related to multi-center IT research projects and cloud infrastructures. Work on UrbanTwin and HeatingBits (see page 22) is underway, SEAMS (a collaboration between the EPFL Space Center and the SKAO) started in early 2024.

"This year, research activities are focussed on the agreed projects with the different schools and centers at EPFL, but in the future, we expect to make open calls for anyone at EPFL interested in research related to sustainable computing to be supported by EcoCloud."



Inside the CCT Datacenter

Best practices for IT infrastructure

The dissemination of best practices for sustainable IT infrastructure is another core mission of EcoCloud. "In

cooperation with the Vice-Presidency for Responsible Transformation (VPT), we are going to develop a course about the fundamentals of sustainable computing for EPFL students at the master level, which will be offered by the Section of Electrical Engineering (SEL) and the Section of Computer Science (SIN) for the complete campus," says Atienza.

"Continuous education for professionals is also important. We plan to offer training to companies to support and assist them in their digitalization processes and help them understand how to implement the most sustainable IT technologies and processes possible."

"IT is the engine of our digital world. With a compound annual growth rate of more than 16 %, cloud computing must embrace a strategy of digital responsibility to support economic progress and societal development without compromising the future of our planet," concludes Atienza.

The key pillars of EcoCloud activity

<< IT infrastructure as enabler for a sustainable society

- Energy-constrained and sustainable deep learning
- Sustainable smart cities and transportation systems
- Computational and data storage sustainability for scientific computing
- Energy-constrained trustworthy systems

<< Sustainable IT infrastructure

- Minimal-energy computing and storage cloud platforms
- Sustainable use of renewable energy in IT infrastructures

<< Dissemination of best practices for IT infrastructure in a sustainable society

<< Preparation of courses and focused programs on sustainable computing, for IT professionals

<< Annual EcoCloud event on sustainable computing trends and forward-looking research

FACULTY MEMBERS

AND LABS - IN ALPHABETICAL ORDER



Anastasia Ailamaki

*Data-Intensive Applications
and Systems Laboratory*

Enabling discoveries in scientific domains through automating physical database design, revolutionizing exploration algorithms in very large data repositories



Alexandre Alahi

*Visual Intelligence
for Transportation*

Socially aware AI applying computer vision, deep learning and human-robot interaction to transportation applications



David Atienza

*Embedded
Systems Laboratory*

Efficient machine-learning based resource management in servers and data centers. Low-power design of edge AI and heterogeneous server architectures



Antoine Bosselut

*Natural Language
Processing Lab*

Natural language processing, machine learning, artificial intelligence



Thomas Bourgeat

*Verification and Computer
Architecture Laboratory*

We leverage the power of formal methods and high-level hardware programming languages to ensure the correctness and the security of tomorrow's computers



Maria Brbic

*Machine Learning for
Biomedicine Lab*

Developing machine learning methods that solve real-world data challenges, paving the way for new biomedical discoveries



Edouard Bugnion

*Data Center
Systems Laboratory*

Data center efficiency, infrastructure support in network and data planes for OLDI applications. System security, Trusted Execution Environments in hardware



Andreas Burg

*Telecommunications
Circuits Laboratory*

Design of technology systems, prototypes and demonstrators for the development of robust, reliable and energy efficient systems



Volkan Cevher

*Laboratory for Information and
Inference Systems*

Robust machine learning and optimization, reinforcement learning, game theory, and deep learning.



Drazen Dujic

*Power Electronics
Laboratory*

Ensuring reliable, compact, and efficient power electronics-based power supplies for data centers. From the power grid to the chip



Babak Falsafi

*Parallel Systems
Architecture Laboratory*

Computer architecture, vertically integrated data center systems, post-Moore server design



Olga Fink

*Intelligent Maintenance and
Operations Systems Laboratory*

Development of intelligent algorithms for complex infrastructures and industrial systems. Deep learning and hybrid algorithms for intelligent maintenance systems

FACULTY MEMBERS

AND LABS



Pascal Frossard
*Signal Processing
Laboratory*

Computer vision, medical imaging, network machine learning, robust machine learning, deep learning



Rachid Guerraoui
*Distributed
Computing Lab*

Distributed machine learning with Byzantine resilience and privacy. Distributed algorithms for new technologies: RDMA and NVRAM



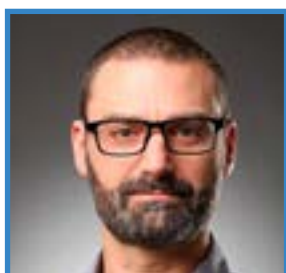
Paolo Ienne
*Processor
Architecture Laboratory*

Computer and processor architecture, FPGAs and reconfigurable computing, electronic design automation, computer arithmetic



Martin Jaggi
*Machine Learning
and Optimization Laboratory*

Machine learning, optimization algorithms and text understanding, as well as several application domains



Colin Jones
*Automatic
Control Laboratory*

Theory and practice of optimization-based, or model predictive control with a particular emphasis on problems arising from renewable energy challenges



Sanidhya Kashyap
*Robust Scalable
Systems Software Lab*

Robust and high-performance software for heterogeneous hardware: concurrency, scheduling, networks, analytics and fuzzing



Anne-Marie Kermarrec
*Scalable Computing
Systems Laboratory*

Large-scale distributed systems, failure resilience, performance and privacy-preservation, frugal distributed learning systems



Jean-Paul Kneib
*Laboratory
of Astrophysics*

Reliable transport and precise integration of a flow of 707 Petabytes per year of data from large arrays of radiotelescopes



Christoph Koch
*Data Analysis Theory and
Applications Laboratory*

Efficient and scalable massively parallel real-time analytics engines, complex expressive declarative and domain-specific languages in databases



Viktor Kuncak
*Lab for Automated
Reasoning and Analysis*

Precise automated reasoning techniques: tools, algorithms and languages, for the construction of reliable computer systems



Zhengmao Lu
*Energy Transport
Advances Laboratory*

Towards a deeper understanding of phase change phenomena, creating sustainable energy and water technologies by optimizing interfacial transport



Gabriele Manoli
*Laboratory of Urban and
Environmental Systems*

Analysis and conceptualization of complex urban and environmental dynamics, to guide the design of greener and more sustainable territories

FACULTY MEMBERS AND LABS



François Maréchal

*Industrial Process and Energy
Systems Engineering*

Process and energy system engineering for efficient use and reuse of energy, efficient energy conversion, integration of renewables and complex system integration



Elison Matioli

POWERlab

Microchannel liquid cooling of data center components, ultra-efficient purpose-built cooling solutions



Giovanni De Micheli

*Laboratory of
Integrated Systems*

Modelling of hardware with dedicated languages, co-design of software and hardware, system-level optimization with efficient performance, energy consumption and yield



Christophe Moser

*Laboratory of Applied
Photonics Devices*

Non-linear transformation in fiber optics to simplify machine learning tasks



Martin Odersky

*Programming Methods
Laboratory*

The design and implementation of Scala, to achieve a fusion of object-oriented and functional programming, compatible with platforms such as Java and .NET



Mario Paolone

*Distributed Electrical
Systems Laboratory*

Developing smart grid concept solutions to efficiently deliver sustainable, economic and secure electricity supply



Mathias Payer

HexHive Laboratory

Software testing to discover security bugs. Sanitization for memory, type, and API violations. Fuzzing of complex code to trigger bugs



Clément Pit-Claudel

Systems and Formalisms Lab

Can we leverage compilers, languages, and proofs to build more robust, more efficient, and more trustworthy systems?



Demetri Psaltis

Optics Laboratory

Optical systems such as spatiotemporal nonlinearities in multimode optical fibers, used as neuromorphic neural networks



Jürg Schiffmann

*Laboratory for Applied
Mechanical Design*

Design and experimental investigation of small scale turbomachinery for decentralized energy conversion



Mirjana Stojilovic

*Parallel Systems
Architecture Laboratory*

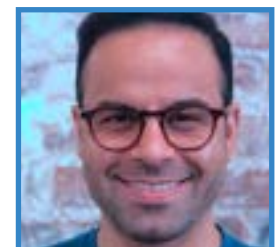
Electronic design automation, reconfigurable computing, electromagnetic-compatibility and signal-integrity issues, hardware security



Carmela Troncoso

SPRING Lab

Designing strong, embedded security and privacy guarantees in complex systems. Quantification of the information an adversary can infer from acquired data



Amir Zamir

*Visual Intelligence and
Learning Lab*

The development of computer vision models that can function as part of larger intelligent systems

INDUSTRY AFFILIATES

MISSION

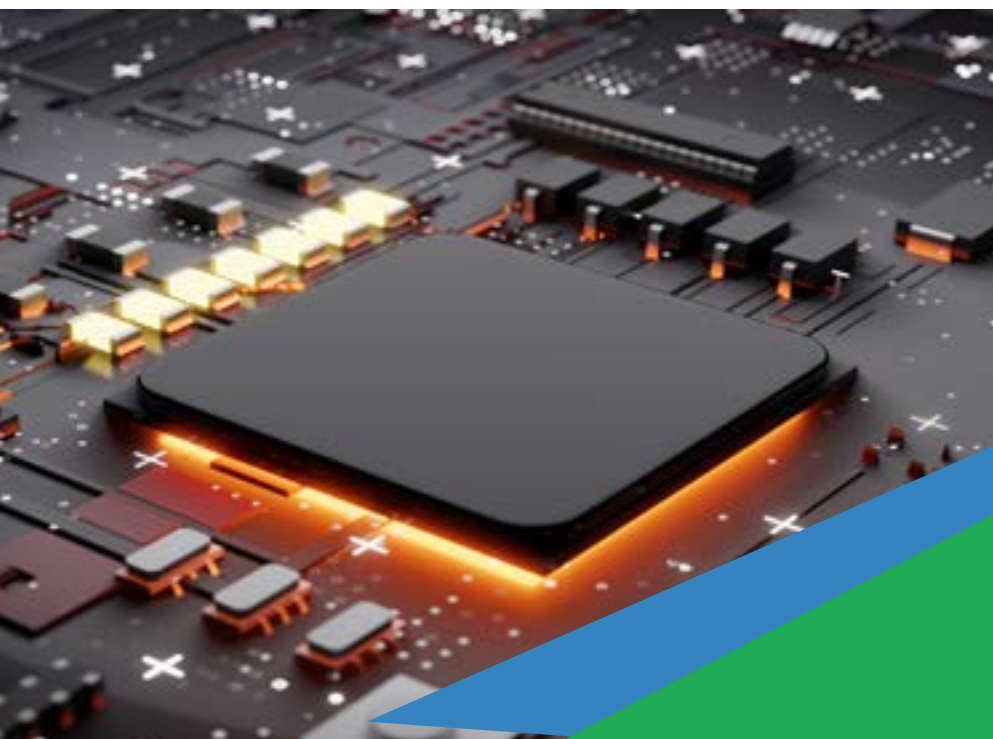
The EcoCloud Industry Affiliate Program (IAP) offers companies a unique opportunity to collaborate with EPFL faculty, students and researchers. Affiliates are given unparalleled access to new technologies and ideas as they move from laboratory to marketplace. The ideal platform for communication and discovery between the research and corporate communities, the program catalyzes collaborative research, customizes educational programs and facilitates graduate recruiting.

The EcoCloud IAP was created to enable connections and strengthen collaborations between EcoCloud and industry. While there are numerous benefits to joining the IAP and becoming an affiliate, the advantages boil down to three key reasons:

<< To gain early awareness of the latest research - Through meetings, visits and online resources, EcoCloud Industry Affiliates get to preview the latest research findings from across our labs before they are published.

<< To explore potential research collaborations and sponsorships - Companies can get much more by sponsoring EcoCloud research directly. Becoming an affiliate allows you to see how we work and what we do, giving you the insights you need to identify research partnerships.

<< To recruit EcoCloud students - Our students are one of our most valuable assets. They can add substantial value to your company as interns or employees. We post student profiles on the members-only website, and you can meet them at our Annual Event. EcoCloud can also host talks geared towards students, and distribute job and internship announcements.



INDUSTRY AFFILIATES

MEMBER BENEFITS

The EcoCloud IAP is a corporate membership program whereby companies pay an annual membership fee in return for facilitated access to the research programs, the researchers and the graduate students, offering the ability to capitalize on the unique, dynamic trans-disciplinary innovation culture at EPFL. At the heart of EcoCloud's IAP activities are entrepreneurial startups emerging from our research, the delivery customizable executive education, industry collaboration and technology transfer.

Specific benefits of becoming an Affiliate include:

<< Annual Event - An annual conference exclusively designed for our existing and prospective Industry Affiliates and research colleagues to showcase the activities of the center. The event's program includes activities such as presentations by EcoCloud faculty and researchers on the latest research results organized around chosen research themes; student poster sessions; lab tours; demonstrations; discussions on grand challenges, applications and technology roadmaps; and opportunities to meet and network with EcoCloud researchers, students and colleagues.

<< Research Monitoring - Throughout the year, EcoCloud will enable Affiliates to remain engaged by means of newsletters, seminars, talks and virtual meetings. When feasible, these events will be broadcast live for our IAP member companies, so participants can join in remotely to hear from different thought leaders at EcoCloud and keep informed about the latest research. Affiliates also have access to the comprehensive information about research outcomes, EcoCloud events, video and publication archives, and other research outputs.

<< Graduate Student Recruiting - EcoCloud will organize events (including the Annual Event), during which Affiliates have access to soon-to-be graduating students to facilitate recruiting. We work with our Affiliates to facilitate recruiting activities throughout the year, including advertising job and internship announcements, hosting talks and seminars, and other student-targeted networking events.

<< Occasional Visits - Affiliates may arrange visits to EcoCloud's affiliated laboratories and the experimental facility created in the new data center of EPFL. Visits enable previews of EcoCloud's research programs and results and demonstrations of emerging technologies. We work with our Affiliates to identify appropriate EcoCloud researchers, ongoing projects and potential opportunities for collaboration.

INDUSTRY AFFILIATES

MEMBER BENEFITS

<< Joint Research Projects - Member companies have opportunities to engage with EcoCloud in research projects and collaborations into deployable technology. These include, but are not limited to: the opportunity to contribute and formally participate in EcoCloud research projects, customization of educational programs and the opportunity to develop and sponsor structured research programs.

<< Advertising - Our affiliates have the opportunity to promote their company's brand on EcoCloud's website and reports.

<< Outreach and Executive/Continuing Education - EcoCloud develops and hosts at least two outreach programs per year. Past events included a summer workshop on cooling technologies, a winter school on data-centric systems, and co-hosting the 3D silicon integration conference. Moreover, EcoCloud is eager to develop special executive courses, as well as continuing education courses, to address the specific needs of our affiliates.

<< Technical Advisory Board - EcoCloud Affiliates designate a technical staff to the EcoCloud Technical Advisory Board which meets once a year to discuss grand challenges, research and industrial trends, and EcoCloud research direction. This annual meeting can vary from a one-hour meeting to a full-blown two-day retreat to present research and solicit feedback.

<< Visiting Scholars and Fellows Program - The EcoCloud Visiting Scholar and Fellows Program stimulates and supports our research by engaging promising scholars and practitioners in order to foster exchange. Each year, a number of distinguished academics (Visiting Scholars) and junior faculty and students (Fellows) will be selected on the basis of their qualifications, the quality of their research plans, and the relevance to both EcoCloud's mission and targeted research objectives. EcoCloud's Visiting Scholars and Fellows will work on projects that offer joint collaborative opportunities.

CURRENT ECOCLOUD AFFILIATES

IN ALPHABETICAL ORDER



IBM Research Europe

IMMO Ventures

infomaniak

intel®

JJ COOLING
INNOVATION

ORACLE



WE NEED TO DRAMATICALLY CHANGE THE WAY WE BUILD AND USE COMPUTERS

Is sustainable computing compatible with business as usual? Not according to David Atienza, Director of EcoCloud, who says we need a culture change in how we build and use computing systems.

David, haven't we always needed lots of energy to power data centers? What's changed now?

In the past, for reasons related to Moore's Law, the number of elements in a data centre would double while the energy consumed on the computing side was roughly the same. In recent years, it's been harder to get the same efficiency. We are seeing the effects of humans' energy use on the planet, and it's getting harder and harder to get energy at the levels we need.

That's why we need more sustainable data centres and, in general, to develop a completely different way of building them. It's not just about locating massive data centres needing 50 MW of energy in the middle of nowhere: you still need to provide this energy, and that will affect the whole planet. Today it is estimated that data centres use 2-3% of the energy generated in the world. If current trends continue, this could reach around 7-8% by 2030.

What's driving this growth? New applications or heavier use of established applications?

Both. Many of us are using applications that were used by a minority in the past – social media, for example, or online streaming – and the COVID-19 pandemic created major growth in these areas. There are also new applications in areas like artificial intelligence (AI) that are growing dramatically in the use of resources, such as large language models.

How does EcoCloud go about building smarter data centres?

EcoCloud brings together academic and industrial

partners to work on solutions. Among our industrial partners are major information-technology (IT) infrastructure providers, like HPE, Microsoft and Huawei, along with users of data centres, such as Swisscom and AXA. The technologies we develop, both hardware and software, are open source.

There are also 28 labs affiliated with EcoCloud, with a total of around 250 researchers, from the engineering, computer science, basic science and environmental engineering schools at EPFL. The main idea is to create multidisciplinary, public-private projects to deliver IT as a service to society.

One project, UrbanTwin, is creating a digital twin of the Swiss towns Aigle and Lausanne in partnership with ETH Zürich and others. EcoCloud acts both as coordinator and as cloud-computing infrastructure provider. In another project, Heating Bits, led by Professor Mario Paolone at EPFL, we're building a new 5MW data centre from scratch in such a way that we extract as much as possible from every bit of energy that comes out. For example, we have technology to extract electricity from the waste heat produced by cooling data centres. The remaining heat is then used to heat university buildings at EPFL.

EcoCloud has an educational component, too. We organize hands-on courses for companies, prepare courses for under-graduate students at EPFL and hold regular seminars.

Speaking of education, are there any myths about sustainable computing you'd like to bust?

WE NEED TO DRAMATICALLY CHANGE THE WAY WE BUILD AND USE COMPUTERS

Very often what people consider 'sustainable computing' is actually just 'energy-efficient computing'. In fact, the largest part of the sustainability cost for data centres is building the servers. Servers are typically replaced every three to five years, even though research has shown that it would take (at least) 10-12 years of service for the carbon footprint involved in building them to be balanced out. So this replacement scheme does not work for the true concept of sustainable computing.

In Heating Bits, the goal is to take servers that are being replaced and change to new liquid cooling (integrating a new technology developed at EPFL, linked closely to my ERC Consolidator project Compusapien). This allows servers to be cooled down while operating in an overclocking regime on a continuous basis when performance is needed. Doing this can make them run for another six to seven years and deliver even better performance for certain workloads.

What developments would you like to see in this field over the next few years?

I'd like to see an awareness of what sustainable computing really means. We need to develop a completely new culture, dramatically changing how we build computers and how we use them. This aligns with EcoCloud's mission to create a new sustainable computing research domain.

At the moment, when we create computing systems, we tend to think in a very limited way, focusing on consuming as little energy as possible while maximizing performance. In reality, you need a much broader understanding of the full lifecycle to be able to really address sustainable computing. A large set of competences is necessary to properly understand how people use computers, and we need to educate users to recognize that calculations are not for free.

To do this, at EcoCloud, we realized we needed to expand the labs involved from those focusing on computing to include others covering a whole range of disciplines. We've also increased the number of companies involved; by including datacentre providers, hyper scalers and companies who use computing, we can evaluate the compute requirements for different applications.

For example, a luxury goods provider like Rolex could delay the most accurate (and computationally intensive) simulations of new models until they have a clearer understanding of the final design. Or in the case of banking, scheduling calculations at a certain time of day could maximize the use of renewable energy resources.



Similarly, with the new generation of digital twins, as in the Urban Twin project, it may not be necessary to analyse every detail; mobility models, for example, could be much coarser and still provide valuable insights, and you can then create detailed models for particular problems. This involves a change of mindset; it's about being application or goal driven.

All of our large projects are driven by PhD students and postdocs, and we encourage them to exchange information and come up with new ideas. Moreover, we are developing a new sustainable computing experimental facility in the new EPFL data centre, managed by EcoCloud. Once it's ready, we'll be able to work on master's and PhD theses where we can really analyse the sustainability of different computing methods. For example, we'll be able to calculate the carbon footprint of a particular algorithm running on our servers at a particular time of day. This kind of analysis can help create awareness and develop new IT sustainability courses that can drive the culture change necessary for more sustainable ways of computing.

Originally published in HiPEAC High Performance Embedded Architecture and Compilation Info magazine.

Main image: Dr. Miguel Peón, EcoCloud Scientist and Prof. David Atienza, Director of EcoCloud

Image above: Professors Alex Alahi, David Atienza, Edouard Bugnion, Mario Paolone and Babak Falsafi, all of EcoCloud



A TRUE MEASURE OF DATACENTER CONSUMPTION

Prof. Babak Falsafi has prepared the following text as a review of the state-of-the-art in datacenter energy efficiency.

The datacenter market is constantly evolving – and so is energy consumption. Until now, energy efficiency has often been expressed in terms of the PUE (Power Usage Effectiveness) indicator, but this is increasingly out of touch with reality. How can datacenter efficiency and emissions be accurately determined?

Electricity markets have been extremely volatile over the past two years, due to concerns about energy supply. With increasing digitalization and the introduction of artificial intelligence (AI) in all sectors, concern is growing not only about electricity consumption in datacenters, but also about the growing demand and ecological impact of IT. A study by Schneider Electric forecasts a 5% annual increase in electricity consumption for the entire IT sector between 2023 and 2030, 75% of which is expected to be attributed to datacenters (driven by AI) and mobile networks (due to the move to 5G).

Electricity consumption by datacenters worldwide has remained relatively stable over the past ten years, constituting around 1.5% of global electricity consumption. In service-oriented economies, this percentage is somewhat higher: in Switzerland, for example, it is 4%. This stability can be explained by the fact that companies are increasingly turning to the cloud, and cloud providers are keeping a close eye on electricity consumption to maximize the return on their investment. Colocation datacenters, which house customers' IT equipment, are continually developing more efficient infrastructures for cooling, power distribution and heat recovery.

A 2022 study by the Uptime Institute (see graph below) shows that the energy efficiency of datacenters worldwide as measured by the conventional energy efficiency indicator PUE, obtained by dividing the total

energy consumed by the datacenter by the total energy used by the IT equipment has been declining only slowly in recent years.

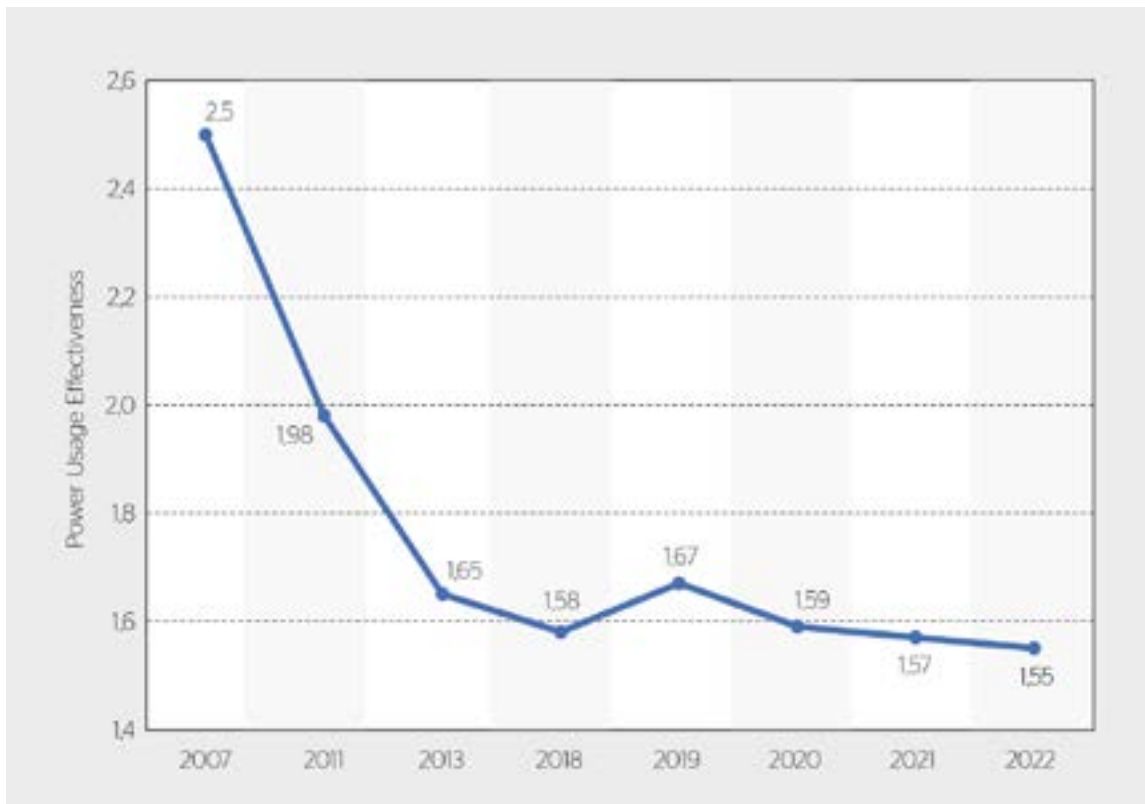
What are the limits of PUE?

Like many other indicators, PUE has its limitations. First of all, it does not take into account the total environmental impact of datacenter operations. It only considers the share of electricity consumed in the building's infrastructure, including cooling and power distribution, and does not take into account the various ways in which the overall flow of energy within the datacenter can contribute to sustainability, or reduce emissions. Modern datacenters rely on both waste heat recycling technologies and on-site renewable energies.

Secondly, the PUE value is subject to fluctuations depending on factors such as the season, current datacenter load and even the time of day. This makes it an unreliable measure of efficiency, especially for datacenters operating under variable environmental conditions and loads.

One of the biggest limitations of the PUE value is that it is not very relevant for measuring IT efficiency. Ironically, inefficient servers can make the PUE value look surprisingly low. This is because the more energy IT devices consume, the better the PUE. This encourages the provision of surplus IT resources to artificially improve PUE values. Even if IT equipment is efficient, which is probably the case in newly built datacenters, the degree of utilization of IT systems also has a considerable influence

A TRUE MEASURE OF DATACENTER CONSUMPTION



Historical development of the PUE energy efficiency indicator

on operational efficiency – but the PUE value doesn't tell you whether servers are being used at 20% or 80%.

In Switzerland, thanks to exemplary work in the field of sustainable datacenters, we are aiming for a PUE value of 1.15. This means that over 80 % of the electricity in these datacenters is used for IT equipment (servers, memory, network). The question of energy consumption and efficiency in datacenters therefore revolves around IT.

Where is IT heading?

Technological forecasts point to spectacular growth in electricity consumption in IT. On the one hand, silicon manufacturing technologies have benefited for four decades from a doubling of chip density every two years (Moore's Law). This increase in chip density has been accompanied by a corresponding improvement in energy efficiency, so that denser chips have been able to operate at higher frequencies, without increasing overall energy consumption.

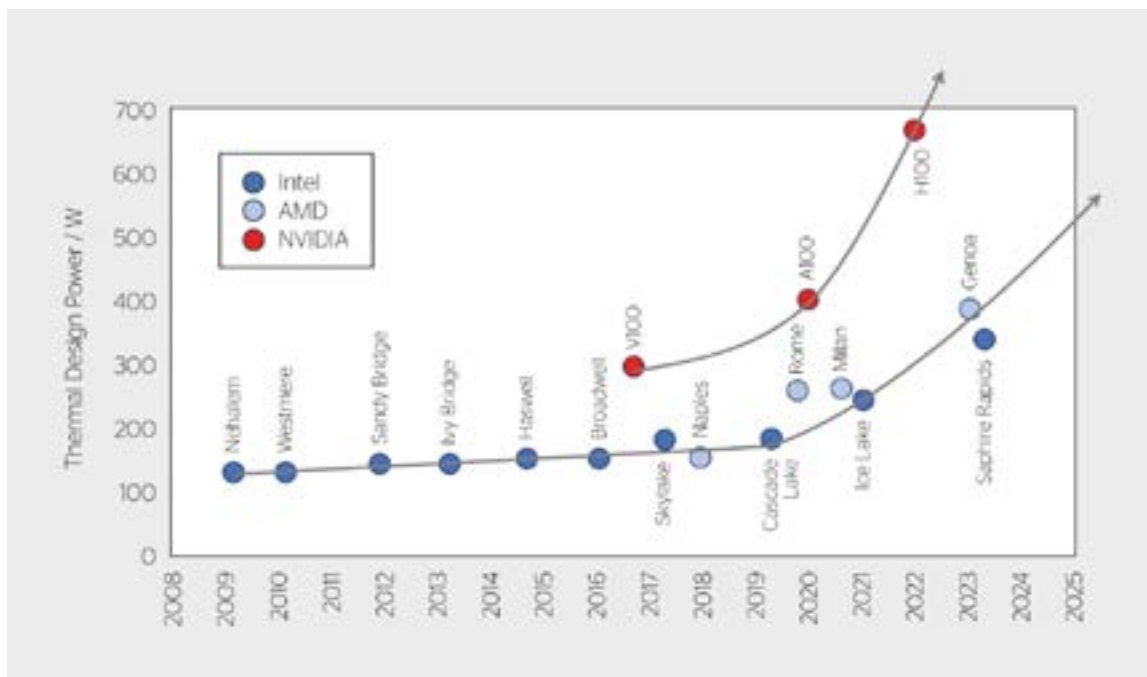
But progress in silicon density has meanwhile reached physical limits. While there are improvements in algorithms, software and chip design that enable platform

specialization, none of these will lead to an exponential increase in density for all datacenter services.

On the other hand, demand has increased by an average of six times every year over the past ten years due to the rapid growth of artificial intelligence (AI). This is happening at the same time as Moore's Law is slowing down, meaning that new computing devices now have to be developed and deployed faster than ever before.

The graph below illustrates the rapid increase in Thermal Design Power (TDP) - the maximum heat dissipation in watts - of server processors (CPUs), the basic computing units in datacenters. CPU TDP rose from single-digit values in the 1990s to around 100 W in 2000, then stabilized for a decade thanks to energy-efficient designs. However, as gains in efficiency and density get smaller and smaller, TDPs increase rapidly for the latest CPUs. Graphics processors, the GPUs that form the basis of AI, have a dramatically higher TDP value: 300 W for an Nvidia A100 card, and the latest product, the H100 card, can even reach 700 W.

These trends call for appropriate methods and indicators to assess the energy efficiency of IT devices and loads.



Increase in maximum heat dissipation of server and graphics processors in recent years

What are the right metrics for IT?

Given the predicted growth in IT energy consumption and the advances being made in datacenter infrastructure, new indicators and methods for assessing datacenter energy efficiency and emissions are needed. This applies to both infrastructure and IT equipment. These indicators must not only take into account heat recycling and the use of renewable energies in the building's infrastructure, but also the efficiency of the various components of IT equipment, including computing logic (e.g., CPUs, GPUs and accelerators), memory, data storage and network equipment. In addition, precise measurement methods and appropriate software and hardware instruments are needed to determine these indicators.

It is also essential to look at other indicators which, until now, have tended to be in the background, but are becoming increasingly important. Workload utilization, for example, offers a more nuanced view of how efficiently IT resources are being used. A server that is only 20% utilized most of the time is not just an unused resource, but also flagrant inefficiency, leading directly to wasted energy and higher operating costs.

Technological quality is also essential. Although not a traditional indicator, it serves as a benchmark for evaluating the devices and methods used in a datacenter. When choosing technologies, operators must ensure maximum performance and focus on the most modern and efficient options. For example, opting for flash memory instead of hard disks can significantly reduce power consumption and cooling requirements, while speeding up data access. Similarly, choosing fiber optic cables instead of copper cables for networks not only

increases speed, but also minimizes energy consumption. This choice also extends to servers, power supplies and power distributors, which stand out for their energy efficiency, reliability and longevity, reducing both energy consumption and total cost of ownership.

Finally, the maximum permissible operating temperature is another important indicator. Traditionally, datacenters operated at lower temperatures to minimize the risk of overheating. Modern devices, however, have been designed to operate safely at higher temperatures. By adapting maximum permitted operating temperatures to these higher limits, companies can drastically reduce energy consumption for cooling, which has a considerable impact on overall data center efficiency.

The SDEA label is comprehensive

In order to measure datacenter efficiency and emissions holistically, the Swiss Datacenter Efficiency Association (SDEA) – a consortium of sustainability pioneers from industry and science - created the SDEA label in 2020. SDEA's KPI (key performance indicator) tool features a calculator that records heat recycling, use of renewable energies, consolidation and virtualization of workloads, use of servers, storage and network components, data compression, first-class component technology and permissible operating temperature. According to a recent study by the International Energy Agency (IEA), the SDEA label is the only certification for datacenters that offers a quantitative ranking and does not simply make recommendations.

Originally published in bulletin.ch

Main photo: Prof. Babak Falsafi at the EPFL CCT Data Center



HARNESSING LIGHT

It is not often that professors from EcoCloud are featured in the pages of The Economist. In an article from December 2022, entitled "Artificial intelligence and the rise of optical computing", the author had this to say:

"The idea of turning neural networks optical is not new. It goes back to the 1990s. But only now has the technology to make it commercially viable come into existence. One of the people who has observed this transition is Demetri Psaltis, an electrical engineer then at the California Institute of Technology (Caltech) and now at the Swiss Federal Institute of Technology in Lausanne. He was among the first to use optical neural networks for face recognition."

Although Demetri Psaltis was indeed among the first to research optical neural networks, even before the 1990s, many of his most recent publications deal with this same topic. This includes one from earlier this year: an article in Nanophotonics, which was already cited by another author in the same journal within a month. However, that does not mean that Prof. Psaltis has been focused solely on this subject for the last forty years: it is a lot more complicated than that, and a lot more interesting.

The dawn of optical neural networks

Neural networks are a technology that gather data in a way inspired by our nervous system – a lattice of facts (nodes) and connectors (edges). They use artificial intelligence to increase the depth of their own knowledge, for the benefit of the user. According to ChatGPT (which is itself a neural network), the largest neural network in the world is Google's GShard, with over a trillion parameters. This is a far cry from Prof. Psaltis' ground-breaking work on optical neural networks in the 1980s:

"Our first neural network had thirty-two nodes. Not much, but it was a start. It is not a coincidence that the rise of neural networks happened just as the home computer revolution was taking off!"

ChatGPT is a neural network, which uses machine learning to gather information on a range of subjects, increasing its knowledge autonomously, and using natural language processing to provide answers to users' questions.

Wikipedia is not a neural network, because it is maintained by human volunteers, who provide, edit and correct the information stored in its databases. It is much more reliable than ChatGPT – for the moment, at least.

The optical neural networks revolution had barely begun, but – as pointed out in the Economist – Psaltis was one of the pioneers. At the bridge between the domains of physics and computing, this ground-breaking concept involves harnessing light, and using it to create neural networks. Optical computers transport data with photons rather than electrons, offering the possibility to make use of the amazing properties of light – its parallelism and its speed – as a form of parallel processing. This not only results in fast data processing, but runs with lower power consumption than a traditional computer.

In an article published in Applied Optics in 1985, Psaltis described the basis of this new technology:

"Optical techniques offer an effective means for the implementation of programmable global interconnections of very large numbers of identical parallel logic elements. In addition, emerging optical technologies such as 2D spatial light modulators, optical bistability, and thin-film optical amplifiers appear to be very well suited for performing the thresholding operation that is necessary for the implementation of



Illustration taken from a 1987 paper on using optical neural networks for facial recognition – with self-portraits

the model.”

Implementing this model was a breath-taking challenge, and one of many that made the reputation of Demetri Psaltis, who was then a researcher at Caltech. In order to get an idea of how important his research was to the development of optical computing, one only has to look up the phrase “optical neural network” in Wikipedia. The first reference in the corresponding article is to a 1988 paper by D. Psaltis, the title of which – Adaptive Optical Networks Using Photorefractive Crystals – conjures up images of physicists wrangling beams of light, like frontiersmen taming wild horses.

There was, however, a problem: “Back then my thing was building optical computers. We built them then, and we’re building them now. The difference is that back then we needed enormous databases, massive memory, super-fast networks. We didn’t have any of that.”

The secret is in diversity

According to an article published in 2021 in APL Photonics:

“The pioneer himself, Psaltis, declared in the 1990s that he was abandoning optical neuromorphic computing for two reasons: lack of practical devices that can be integrated and insufficient knowledge of complex neural networks.”

Chemical advances in optical components, reconfigurable optical circuits, low-energy applications for optical devices – all of these concepts feature in the work Prof. Psaltis carried out at Caltech up until 2007, and at EPFL thereafter. Another dominant theme is microfluidics. In 2006 he was working on “devices in which optics and fluidics are used synergistically to synthesize novel functionalities. Fluidic replacement or modification leads to reconfigurable optical systems, whereas the implementation of optics through the microfluidic toolkit gives highly compact and integrated devices.”

After a long absence, however, neural networks made a return. Perusing his list of publications, the phrase “neural networks” is not to be found from 2003 until 2019. Then it is back – with a vengeance.

The return of the neural network

In 2016, Prof. Psaltis wrote the following in an article entitled “Optical Computing: Past and Future”:

“All-optical information processing has a checkered past: but technological developments, tougher problems and the rise of big data are all prompting a new look.”

Psaltis confirms that changes to the landscape have occurred: “What we have now is Google, supercomputers, fibre optic networks. So everybody is talking about optical neural networks again. In industry, over a billion dollars have been invested in this technology!”

During the course of this interview, doctoral students would occasionally drop in to the office of Prof. Psaltis, and his enthusiasm for their discoveries is immediately visible and audible. So what work are he and his researchers doing now, and how do neural networks feature in it? We can look briefly at two examples.

Using optics to build neural networks

The neural networks of the 1980s have given way to the much larger Deep Neural Networks (DNN) of the modern era, but there is a certain amount of continuity in what Prof. Psaltis was pioneering then, and what he is aiming to achieve now.

One invention in particular – Multimode Fibers (MMF) – has been a game changer. In a paper in Nanophotonics (2022), written in collaboration with Prof. Christophe Moser of the Laboratory of Applied Photonics Devices (LAPD), Psaltis wrote: “In what follows, we review

HARNESSING LIGHT

recent works that use modern data-driven deep neural networks (DNNs)-based methods for imaging, projection in scattering media and specifically MMFs. We show that modern data-driven deep neural networks considerably simplify the measurement system and experiments and show that they can correct external perturbations as well. We also show recent works that MMF can be used as a medium to do optical computing."

Multimode Fibres use multiple rays of light simultaneously, with each ray of light running at a different reflection angle. They can therefore be used to transmit masses of data over the short distance between one part of a computing processor and another. Putting MMF to work in building deep neural networks means going really big, really fast – with lower energy use than traditional binary technology.

At this point, the professor turns things around: as well as using optical computing to build deep neural networks, he is building neural networks that can help design optical components.

Using neural networks to build optics

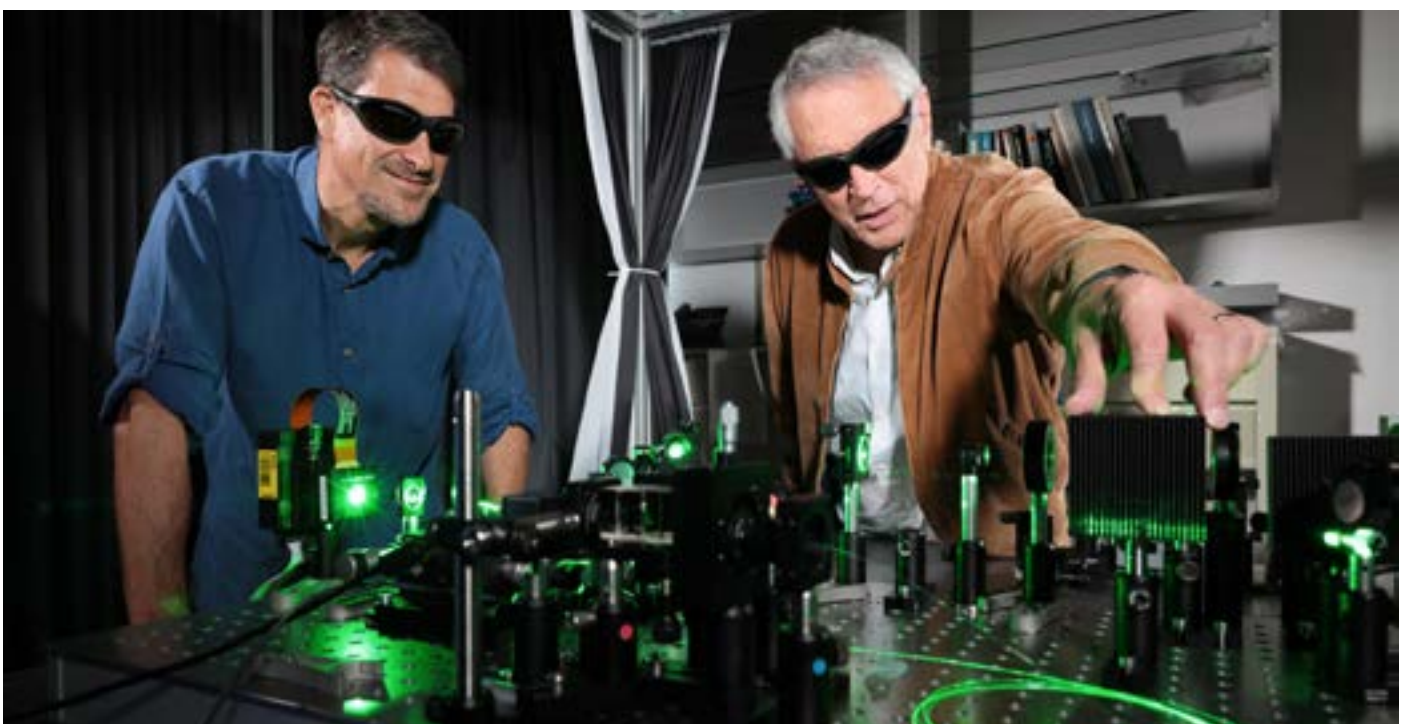
As neural networks grow in depth and speed, they are being utilised to produce work that only humans could previously do. Dall-E 2 is used to generate images, ChatGPT can churn out essays – but the work is arguably poor quality compared to that of a true professional. MaxwellNet is different.

"We built a deep neural network out of Maxwell's equations. With MaxwellNet the laws of physics are the rules. So you build your model, and if these rules are not met there will be an error. You're using the physics described in Maxwell's equations in order to test a simulation."

MaxwellNet was tasked with helping to design lenses, components for optical computers, as described in a paper from last month in APL Photonics: "MaxwellNet maps the relation between the refractive index and scattered field through a convolutional neural network. We introduce here extra fully connected layers to dynamically adjust the convolutional kernels to take into account the intensity-dependent refractive index of the material. Finally, we provide an example of how this network can be used for the topology optimization of microlenses that is robust to perturbations due to self-focusing."

So here is a deep neural network with a difference, as Prof. Psaltis explains: "You don't have to compile a database, you don't need to train the model, so much as to let James Clerk Maxwell decide if your model will work or not."

It sounds like science fiction, but it is just another chapter in this lifelong, ongoing adventure: harnessing light, surfing photons, conjuring matrices of data and watching them dance.



Prof. Christoph Moser and Prof. Demetri Psaltis

A man with glasses and a dark sweater stands in a server room, gesturing with his hands. The room is filled with rows of server racks and complex wiring. A large green and blue arrow graphic points from the top left towards the center of the page.

GETTING CONTROL OVER OUR DATA CENTERS: HEATING BITS

The cloud is ever-growing and, as a consequence, data centers are multiplying in size and number. The situation is difficult to manage. To meet demand, outdated and inefficient technology is often being used to provide quick solutions. Cooling systems are frequently unwieldy and energy inefficient, with little or no use made of the wasted heat. The carbon footprint of data centers is soaring. The situation is becoming ungovernable.

Within the context of the initiative funding "Solutions for Sustainability", EPFL has granted a project called Heating Bits, in an attempt to impose some order on this unruly situation. A multi-talented team of professors are setting their labs to work, defining and implementing new technologies and optimal control strategies to data centers.

"The Heating Bits project develops multiple technologies and methodologies to minimize the carbon footprint of data centers," explains Prof. Mario Paolone. "To do this we are managing different flows: information flows, electricity flows – sources and use – and thermal flows. Heat will be extracted from the CPUs of the center for external use: supplying local district heating, as well as for the generation of electricity by means of an Organic Rankine Cycle."

Heating Bits gathers together the expertise of a range of laboratories. Prof. Elison Matioli (PowerLab) is retrofitting standard server chipsets with on-chip liquid cooling. Prof. Jörg Schiffmann (LAMDA) is developing optimal Organic Rankine Cycles to generate electricity for the recuperated heat. Prof. Drazen Dujic (PEL), is building a microgrid to distribute flows of electricity between energy storage systems, photovoltaic generation and power supplies. Prof. David Atienza (ESL/EcoCloud) will develop a scheduling scheme for multiple virtual machines inside server to perform accurate predictions of the applications running inside the virtual machines. Prof. François Maréchal (IPESE) is working on the multi-energy system integration, including supply of heating,

cooling and electricity services aimed at minimising the carbon dioxide emissions of the overall energy system. As for Prof. Paolone himself, he is building the multi-time-horizon-control that will coordinate all of these systems, so that they play together in time to minimise the data center carbon footprint.

As a six-year project, Heating Bits will be the sole focus of a number of PhD students during their studies at EPFL. "Speaking for my lab, some people are working only on Heating Bits from Day 1. Many will be reoriented to Heating Bits, but all partners are recruiting for this project."

Already deployed are a 60 kWh battery, integrated into a microgrid, and a dedicated space in the new EPFL data center, the CCT building (Centrale de Chauffage par Thermopompe), with its large array of solar panels. A dedicated space for experimentation will be provided in EcoCloud's new experimental area, located in the CCT building. A demonstrator is also being developed with a 50 kW power supply: conversion devices will be sized accordingly.

The importance of laboratories working together is paramount to a project with such a wide range of challenges. Of the labs listed above, all who were here at the time were involved in Nano-Tera. Two are involved in the UrbanTwin initiative, which upscales similar technology for an entire city, and three are members of the EPFL EcoCloud Center.

"In these collaborative projects you are exposed to different disciplines and the know-how of people who are in varied cultural and technical backgrounds,"

GETTING CONTROL OVER OUR DATA CENTERS: HEATING BITS

explains Prof. Paolone, "this is extremely beneficial when developing solutions – and these solutions are often the most successful. Many of the ideas we pioneered within the Nano-Tera project have been deployed throughout the EPFL campus as industry grade products for the operation of our internal electrical grid. Much of this infrastructure will be used for Heating Bits."

There was also a spin-off. To quote the website of Zaphiro Technologies SA: "Based on the PhD-thesis of the three founders, Zaphiro is still keeping innovation at the core of the company strategy, continuously bringing superior solutions to the energy market." Prof. Paolone confirms this: "Zaphiro is commercializing many of the solutions that came out of S3Grid, our Nano-Tera project."

As for Heating Bits, the possibilities for spin-offs abound: "We have been contacted by many leading industry

Participating labs:

Elison Matioli, PowerLab

Jörg Schiffmann, Applied Machine Learning

Drazen Dujic, Power Electronics

David Atienza, Embedded Systems

François Maréchal, Industrial Process and Energy Systems

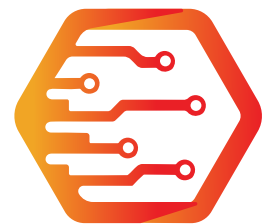
Mario Paolone, Distributed Electrical Systems

players," explains Paolone. "One of them asked to see me right away - one week after the project's kick-off meeting, their representatives were sitting in my office." It sounds dramatic, but this is a problem that requires urgent action.

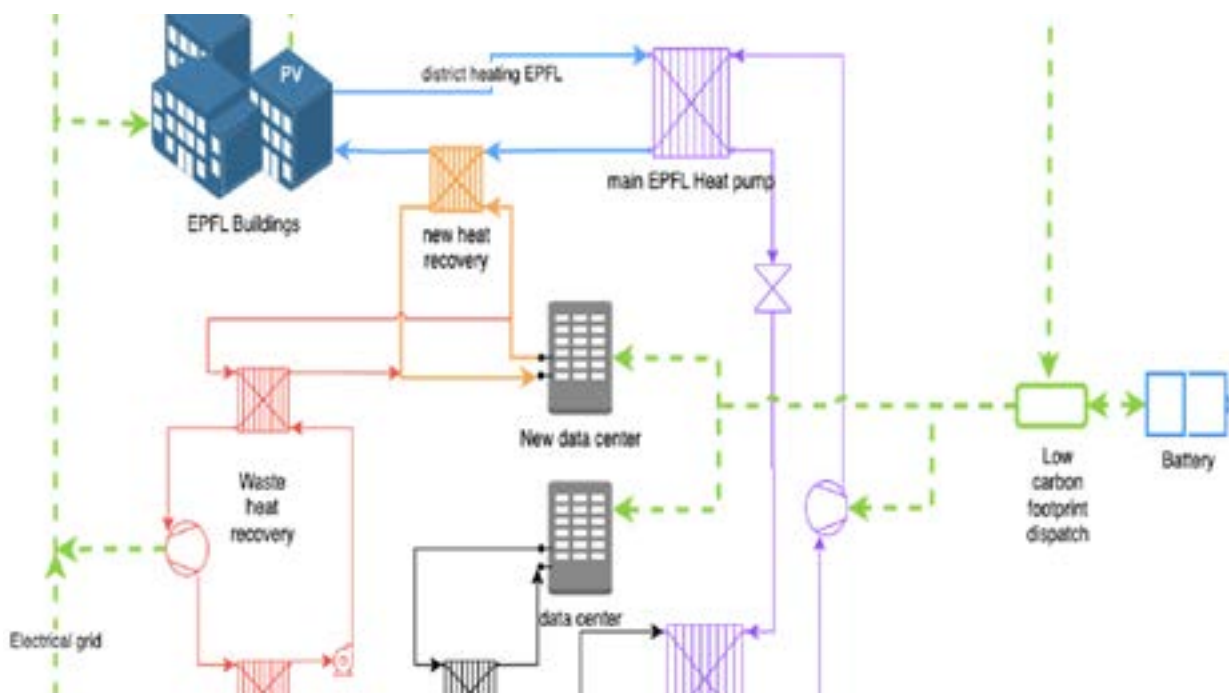
"Data centers nowadays use 2-3% of global energy. It is projected to go to 10% in 2030. The ease with which we generate all this information has exploded," explains Paolone, who is doing his best to get some control over the situation.

Control is key here. Orchestras depend on order, if they are to play with the right tempo, the correct volume and the smooth deployment of the sections of an orchestra. In this case Prof. Paolone is like an opera conductor, controlling the flows of electricity, heat and information, not only with several labs working in conjunction, but also in synchronizing the educational, research and industrial outputs of Heating Bits. For some people, luckily, establishing control comes naturally.

Photos: Prof. Mario Paolone in the EPFL energy storage facility and in his office.



HEATING BITS



Heating Bits heat and information flows



DIVIDE AND CONQUER: BREAKING MEMORY UP INTO COMPARTMENTS TO BOOST SECURITY

For computer security, enforcing the principle “do not let your left hand know what your right hand is doing” is vital.

Modern computers and smartphones do many things at the same time. For example, you may have a browser with one tab playing a song while in another you are logged into your bank account, paying a bill with a QR code that you scan with your camera. Under no circumstances should the audio player’s tab be able to use the privileges that you have granted to the banking tab. These privileges include access to your camera and access to your bank account, so long as you are logged in. With that level of access the audio player could access your bank account and issue illicit transactions.

In this example the browser is a single program handling multiple tasks at the same time. If all tasks run in the same isolation domain then any bug in any task would allow the abuse of the privileges of all other tasks. Each task (and their privileges) must therefore be kept separate and isolated from all others.

Today, the only viable solution is process isolation. Each task runs in an independent process with tight limits on how it can interact with the operating system. This interface, born in the 80s, takes tens of microseconds to switch between any two tasks. Compared to the 80s, today’s programs are many orders of magnitude more complex. A browser consists of around 100 million lines of code (printed on A4 paper, the stack of source code would reach 400m). Given the high switching cost between processes, this isolation is limited to coarse-grained tasks such as individual browser tabs. Instead, it would be better to isolate fine-grained tasks such as restricting an ad or social media interaction from the main web page that is displayed, or to isolate the library that scans your QR code from the rest of the banking app.

Researchers at Prof. Mathias Payer’s HexHive Lab, in collaboration with Prof. Babak Falsafi’s PARSA

Lab, have come up with a surprising approach to this problem, achieving fine-grained isolation of sub-tasks in nanoseconds, thereby providing much tighter security guarantees at low performance cost.

On a computer, each program runs in a so-called virtual address space, a flat array of memory where all code and data is stored. SecureCells, like its related project Midgard, supercharges the address space, and repurposes it: in this case, for compartmentalization. Traditionally, all sub-tasks have access to all code and all data. SecureCells uses compartmentalization to break up the address space into logical compartments which enforce isolation between each one. Code from one compartment can only access data it is given access to.

Compartmentalization works at a finer level than traditional process isolation. First conceived in the 1980s, it is a principle by which software is broken into clearly defined and differentiated areas of memory: compartments, which are separated from each other. Originally, compartmentalization worked by separating address spaces and running each process in its own address space. Separating address spaces is costly and switching between them takes time, in the order of microseconds. Instead, the researchers at HexHive are breaking a single address space into compartments and enforcing security checks at compartment transitions, not address space transitions.

“SecureCells is designed as a secure mechanism inside a single address space, you don’t have to switch address spaces, saving valuable time,” explains Prof. Payer. “All the architectural and microarchitectural costs of switching

DIVIDE AND CONQUER: BREAKING MEMORY UP INTO COMPARTMENTS TO BOOST SECURITY

are dispensed with. Switching compartments takes place at the nanosecond scale, with very little overhead."

The results are dazzling. To use the address space in this new way drastically reduces the cost of compartmentalization while guaranteeing security properties.

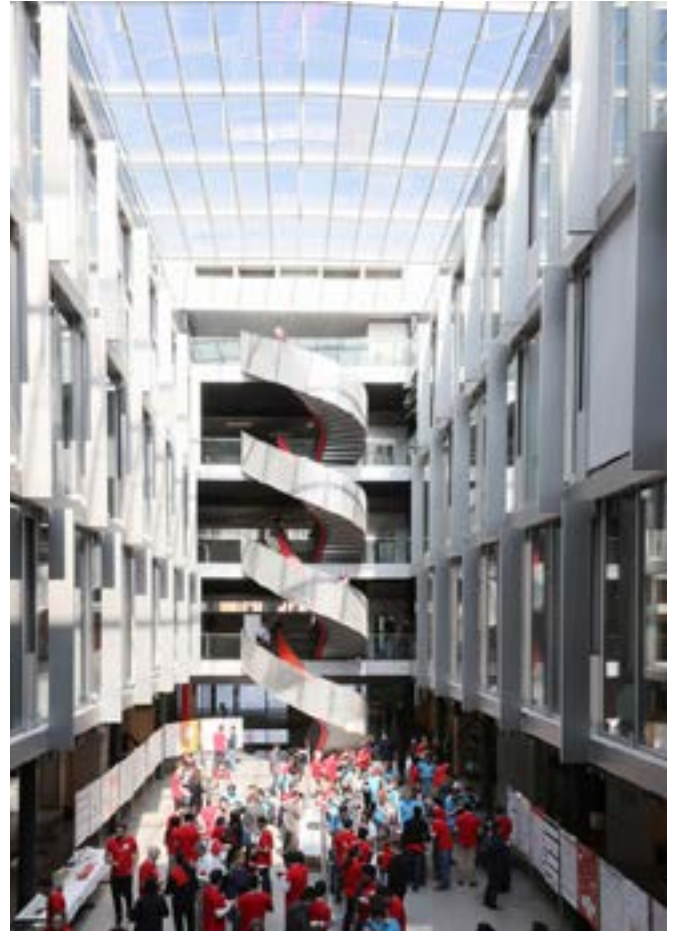
Doctoral Assistant Atri Bhattacharyya, who is jointly advised by Mathias Payer and Babak Falsafi, is clear about this: "SecureCells is a secure mechanism on which to build secure applications. It is a mechanism with an isolation guarantee, and prohibits privilege escalation. We base our guarantee against privilege escalation, with checks implemented for each unprivileged operation in our mechanism."

Industry players have also attempted compartmentalization, with varying degrees of success, according to Bhattacharyya. "Arm and Intel are trying similar techniques. Intel's MPK is a step in this direction, but lacks strong security properties for code. There is a general push in this area, across industry and academia."

"Qualcomm and Intel have both shown an interest in SecureCells," according to Prof. Payer. "With today's complex software we need to find ways to break software into small compartments, enforcing security guarantees between each one. Taking a large address space and dividing it into smaller pieces, with architectural extensions to allow data to go back and forth much faster - this is the best answer to this complex problem."

And where did the idea come from? "Three years of tracking upcoming problems in the security space, noticing common patterns, and defining the challenge. With a well-defined problem statement, the solution just emerges from the problem," explains Bhattacharyya. Which goes to show that three years of thinking can achieve amazing things: if it is the right people, in the right lab.

Photos: Prof. Mathias Payer in his lab



Security conference outside the lab of HEXHIVE





A LARGE LANGUAGE MODEL FOR MEDICAL KNOWLEDGE

EcoCloud researchers have just released Meditron, the world's best performing open source Large Language Model tailored to the medical field designed to help guide clinical decision-making.

Large language models (LLMs) are deep learning algorithms trained on vast amounts of text to learn billions of mathematical relationships between words (also called 'parameters'). They are familiar to most of us as the algorithmic basis for chatbots like OpenAI's ChatGPT and PaLM, used for Google's Bard. Today's largest models have hundreds of billions of parameters, also costing in the billions of dollars to train.

While massive-scale generalist models like ChatGPT can help users with a range of tasks from emails to poetry, focusing on a specific domain of knowledge can allow the models to be smaller and more accessible. For instance, LLMs that are carefully trained on high-quality medical knowledge can potentially democratize access to evidence-based information to help guide clinical decision-making.

Many efforts have already been made to harness and improve LLMs' medical knowledge and reasoning capabilities but, to date, the resulting AI is either closed source (e.g. MedPaLM and GPT-4) or limited in scale, at around 13-billion parameters, which restricts their access or ability.

Seeking to improve access and representation, researchers in EPFL's School of Computer and Communication Sciences have developed MEDITRON 7B and 70B, a pair of open-source LLMs with 7 and 70-billion parameters respectively, adapted to the medical domain, and described in their pre-print *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models*.

Building on the open-access Llama-2 model released by Meta, with continual input from clinicians and biologists,

MEDITRON was trained on carefully curated, high-quality medical data sources. This included peer-reviewed medical literature from open-access repositories like PubMed and a unique set of diverse clinical practice guidelines, covering multiple countries, regions, hospitals, and international organizations.

"After developing MEDITRON we evaluated it on four major medical benchmarks showing that its performance exceeds all other open-source models available, as well as the closed GPT-3.5 and Med-PaLM models. MEDITRON-70B is even within 5% of GPT-4 and 10% of Med-PaLM-2, the two best performing, but closed, models currently tailored to medical knowledge," said Zeming Chen, lead author and a doctoral candidate in the Natural Language Processing Lab (NLP) of Professor Antoine Bosselut who is the Principal Investigator of the project.

In a world where many people are suspicious, or even fearful, of the rapid advance of artificial intelligence, Professor Martin Jaggi, head of the Machine Learning and Optimization Laboratory (MLO), emphasizes the importance of EPFL's MEDITRON being open-source, including the code for curating the medical pretraining corpus and the model weights.

"There's transparency in how MEDITRON was trained and what data was used. We want researchers to stress test our model and make it more reliable and robust with their improvements, building on the safety of the tool in the long and necessary process of real-world validation. None of this is available with the closed models developed by big tech," he explained.

A LARGE LANGUAGE MODEL FOR MEDICAL KNOWLEDGE

Professor Mary-Anne Hartley, a medical doctor and head of the Laboratory for Intelligent Global Health Technologies, hosted jointly in the MLO and Yale School of Medicine, is leading the medical aspects of the study. "We designed MEDITRON from the outset with safety in mind. What is unique is that it encodes medical knowledge from transparent sources of high-quality evidence. Now comes the important work of ensuring that the model is able to deliver this information appropriately and safely."

One of these sources of high-quality evidence is the International Committee of the Red Cross clinical practice guidelines.

"It is not often that new health tools are sensitive to the needs of humanitarian contexts," says Dr Javier Elkin, who heads the Digital Health Program at the International Committee for the Red Cross. "The ICRC is a key custodian of humanitarian principles and we are excited to collaborate with this EPFL initiative that allows us to incorporate our guidelines into the technology."

Through an Humanitarian Action Challenge grant

coordinated by the EssentialTech Centre at EPFL, in early December a joint workshop in Geneva will explore the potential - as well as the limitations and risks - of this kind of technology, with a special session on MEDITRON from the authors.

"We developed MEDITRON because access to medical knowledge should be a universal right," concluded Bosselut. "We hope that it will prove to be a useful starting point for researchers looking to safely adapt and validate this technology in their practice."

The release of MEDITRON aligns with the mission of the new EPFL AI Center that focuses on how responsible and effective AI can advance technological innovation for the benefit of all sectors of society. The EPFL AI Center leverages the extensive existing expertise of faculty and researchers to nurture multidisciplinary engagement in AI research, education, and innovation as well as broader partnerships with different actors in society.

Main photo: Prof. Martin Jaggi, 2021 EPFL / Alain Herzog - CC BY-SA 4.0

Image below: Prof. Antoine Bosselut



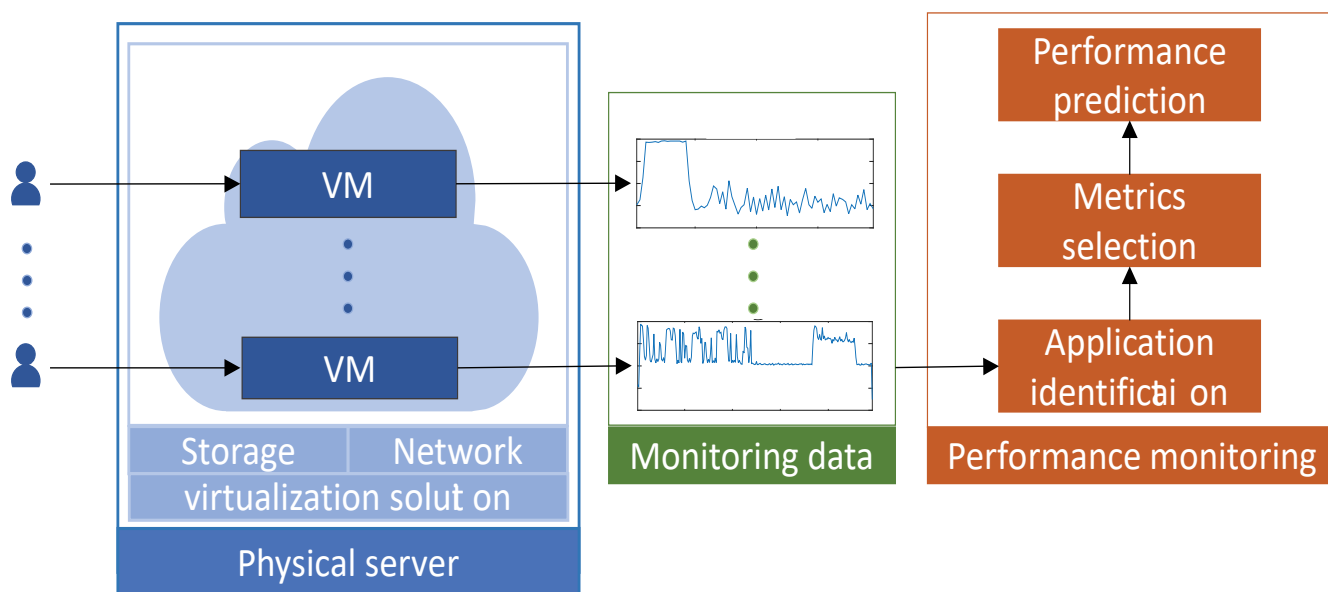
PREDICTING THE FUTURE WITH CLOUDPROPHET

If we are going to reduce the carbon footprint of data centers we need to use computing resources more efficiently. If processes always made use of data center facilities in a regular way, it would be an easy game. However, the resources of a data center are used by customers (more often customers of customers) in ways that are not only unpredictable, but lacking in transparency - even in real time. Data center staff are not permitted to observe the processes being used by their customers. So how do you allocate resources efficiently, when you are being handed black boxes?

To solve a complex question like this, it would appear that a crystal ball is necessary. The crystal ball being

developed at EPFL is called CloudProphet.

"Companies like Google and Amazon provide virtual machines for customers," explains Prof. David Atienza, head of the Embedded Systems Lab in the School of Engineering. "But these customers do not tell you anything about what they are actually doing, and we are



System description and workflow

PREDICTING THE FUTURE WITH CLOUDPROPHET

not permitted to look inside. Therefore, the behavior of applications is hard to predict – they are black boxes.”

By identifying application processes from the outside, and basing performance prediction only on hardware counter information, CloudProphet learns to anticipate an application’s demands on resources. Neural networks multiply in a balletic form of machine learning, building up a picture of the predicted requirements.

“Data centers do have diagnostic tools for the identification and performance prediction of applications, but they can only claim an improvement rate of 18 %. We are achieving results that are orders of magnitude above that,” explains ESL PhD student Darong Huang. “These results have now been published in IEEE Transactions on Sustainable Computing.

“We hope that CloudProphet will pave the way to a more intelligent resource management system for modern data centers, thus reducing their carbon footprint.”

Atienza explains that PhD students and postdoctoral fellows are funded by industry to work on these projects, maintaining systems at a distance.

“Industry project managers are invited to give advice on the physical and logistical constraints in place so that we

get as close as possible to a real-world application,” he says.

The new EPFL data center, located in the CCT building (Centrale de Chauffage par Thermopompe), presents another opportunity to apply this technology to the real world. In collaboration with Mario Paolone’s Distributed Electrical Systems Lab, the EcoCloud Center, and the EPFL Energy Center, a framework will be set up to put these systems to the test.

“Once we can see up to what extent the carbon data footprint is reduced, we can look at whether the next step is to license the software, or to start a spin-off company,” concludes Atienza.

All this is in the future, and nobody can predict the future. Although with the benefits of machine learning, collaboration between research teams and dialogues with industry, we can get close.

Main photo: Prof. David Atienza and fellow Embedded Systems Laboratory researcher Darong Huang

Below: EcoCloud discussion at EPFL Engineering Industry Day 2023



CONTACT

ECOCLOUD



Exploring new paradigms in data center hardware technology, pioneering strategies of cloud data management and innovative security techniques, in collaboration with industrial partners

EcoCloud

EPFL ECOCLOUD
INJ 234 (Bâtiment INJ)
Station 14
CH-1015 Lausanne
Tel: +41 21 693 13 24
contact.ecocloud@epfl.ch
<https://ecocloud.epfl.ch>

Prof. David Atienza

Chief Scientific Officer
EPFL STI IEM ESL
ELG 130 (Bâtiment ELG)
Station 11
CH-1015 Lausanne
Tel: +41 21 693 11 31
david.atienza@epfl.ch

Valérie Locca

Administrative Assistant and IAP Manager
EPFL ECOCLOUD
INJ 234 (Bâtiment INJ)
Station 14
CH-1015 Lausanne
Tel: +41 21 693 13 24
support.ecocloud@epfl.ch



CREDITS

AND RIGHTS

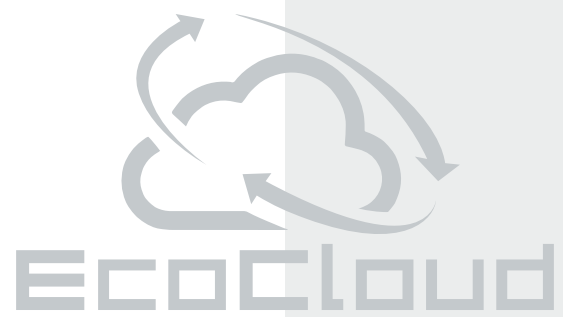


Texts copyright EcoCloud 2022, written by

4	David Atienza
5-6	Leila Ueberschlag
10-13	Babak Falsafi and David Atienza
14-15	HiPEAC Info magazine
16-18	Babak Falsafi
19-21	John Maxwell
22-23	John Maxwell
24-25	John Maxwell
26-27	Tanya Petersen
28-29	John Maxwell

Image Credits - All Rights Reserved

1	Peach-adobe, Adobe Stock: 470470740
4	David Atienza, EcoCloud
5, 11	Eduardo Garlant, Adobe Stock: 317397107
6	Xavier Ouvrard, EcoCloud
7-9	EPFL
10	cherezoff, Adobe Stock: 467673383
12	Issaranow, Adobe Stock: 280047831
14, 15, 16	Alex Widerski, EcoCloud
17, 18	bulletin.ch
19	Demetri Psaltis, EPFL
20	Optical Engineering 26(5), 265428 (1 May 1987) https://doi.org/10.1117/12.7974093
21	Alain Herzog, EPFL - CC-BY-SA 4.0
22	Alex Widerski, EcoCloud
23	Mario Paolone, EcoCloud
24	Alex Widerski, EcoCloud
25, 26, 27	Alain Herzog, EPFL - CC-BY-SA 4.0
28	Alex Widerski, EcoCloud IEEE Transactions on Sustainable Computing 2024-01-23
29	Alex Widerski, EcoCloud
30	Ipopba, Adobe Stock: 480306328



EPFL