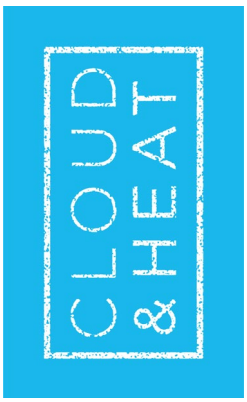


nyris

Cloud&Heat & Nyris

Sustainable AI for a greener planet

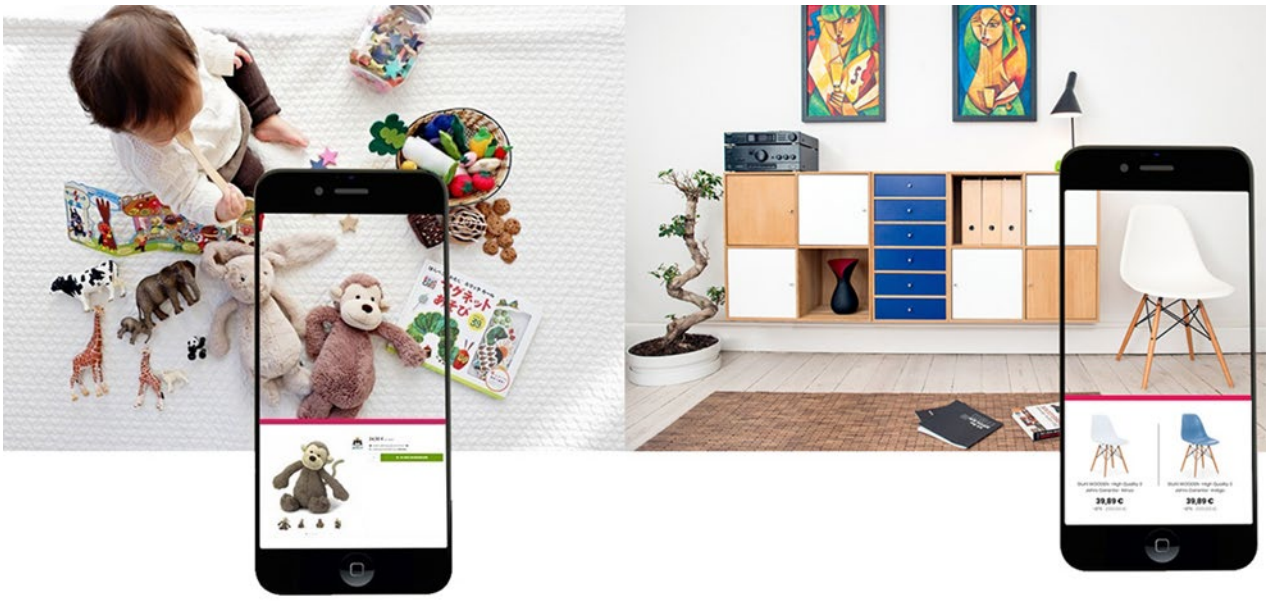
Sustainable AI for a greener planet

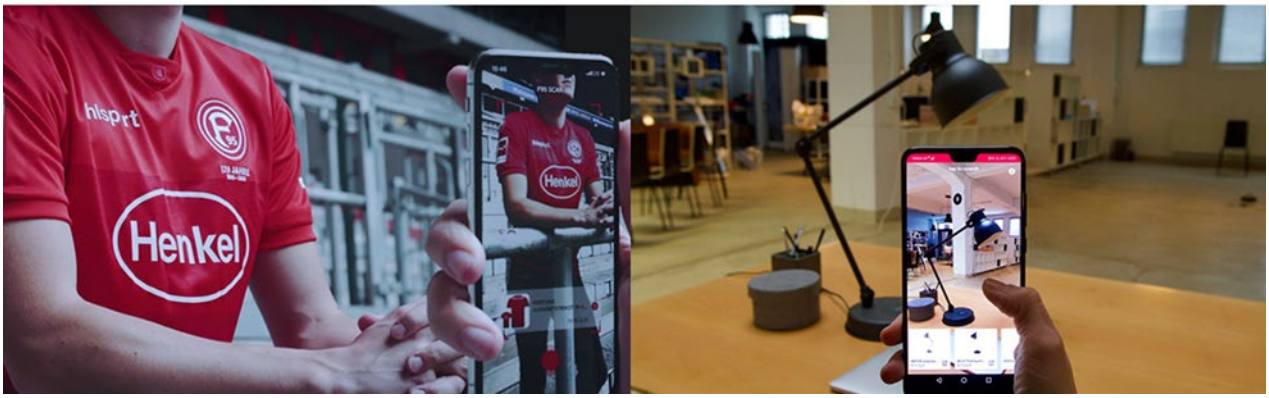
Due to the increasing digitization in all areas of life, data centers are amongst the largest energy consumers around the word. Since more compute capacity is required to drive digitalization, the power consumption of data centers will continuously increase. But what is causing the high amount of workloads in our data centers? Besides web applications such as video streaming or online shopping there are other applications that make some data centers consume a considerable amount of energy: Machine Learning, Artificial Intelligence and the Internet of Things will increase their energy consumption and their emissions in the future.

In the case of Machine Learning, using the algortihms for serving applications constitutes only a very small part of the whole power consumption. The so-called Training, a phase during which Machine Learning Models are "educated", is far more compute-intensive. Recent studies claim that each training can emit as much CO2 as five American cars emit during their whole lifetime[1]. Still Machine Learning provides huge advantages, for example by increasing the efficiency of production and consumption which can lead to a more sustainable society[2]. Hence, it is necessary to increase the efficiency and sustainability of data center infrastructures to overcome the ecological challenges of the increasing demand for compute capacity.

Nyris - giving search the power of sight

A company that uses these infrastructures daily is nyris. Nyris aims to digitize the incredible capability of human sight. For this, artificial visual intelligence is used to enable the nyris visual search engine for spare parts, products or objects and allow users to find those more naturally and efficiently. Besides profound knowledge, creativity and passion this requires a huge amount of compute capacity and energy to develop and train the associated complex Machine Learning algorithms.





Cloud&Heat – the most energy efficient data centers

This compute capacity for these complex calculations is provided by the Green-IT company Cloud&Heat from Dresden. The startup was founded based on the idea of using server waste heat for heating buildings and is now building and operating data centers utilizing an innovative hot water direct cooling technology. The flexibly scalable offers for GPU-, CPU-, and storage are offered on an Infrastructure-as-a-Service (IaaS)-base and provide perfect conditions to AI-companies like nyris.

The Use-Case

In 2019 nyris was faced with the question how compute infrastructure shall be provided and used in the future. Since nyris is aware of their responsibility as AI-company, they set a high value on energy efficiency and sustainability. Hence, after comparing several cloud providers, they choose Cloud&Heat. There they found recent hardware, a high amount of flexible compute capacity and an ecologically meaningful and holistic solution at a reasonable price.

The architecture of neural networks and the training and optimization methods used by nyris have advanced tremendously during recent years. Especially in the area of distance metric learnings, a subset of Machine Learning of particular importance for visual artificial intelligence, the hardware poses a major challenge. These networks are not trained with single images but rather with pairs or triplets of images. During training, several transformation functions are added, random samples for optimizing the objective function are drawn and complex loss functions are applied.

This requires a perfect cooperation of the used software framework and the CPU-/ RAM-/ GPU-hardware. Although many providers are entering the market with special chips, the most flexible applicable hardware are the GPU-systems from Nvidia. This is especially useful when researching new architectures. But still all other hardware components need to be carefully aligned. When using 8 or 16 GPUs, the main memory or the CPU can pose a bottleneck to the training since not enough data reaches the GPUs to fully load them. This leads to long training times, higher overall costs as well as frustration among the responsible data scientists. It is like paying for a Porsche while being stuck in evening rush hour traffic.

Cloud&Heat closely works together with their customers to sort out these types of problems finding tailor-made solution for each application. Hence it is also possible to connect a huge amount of GPUs on special boards for reaching maximum utilization.

The Infrastructure

But why are GPUs even necessary for training Machine Learning Models? During training of neural networks, many repeating, simple calculations have to be executed. This is mainly achieved by utilizing specialized chips. In most cases, GPUs are used nowadays but there are several manufacturers introducing their dedicated chips for neural network training like Google TPU, Graphcore IPU, Intel NNP and Huawei NPU. Looking at a Nvidia V100 GPU, it consists of 5120 cores which are able to perform easy arithmetic operations. Contrary, an Intel Xeon-Platinum-8180-CPU has only 28 cores, but they are able to perform much more complex operations. Hence, on a GPU are more cores that simultaneously can do simple arithmetic operations which is far better suited for Machine Learning.

As platform for the nyris setup, first of all a high-performance hardware was picked: The Tesla V100 from Nvidia is a GPU especially developed for AI-applications and provides excellent performance and efficiency[3]. Thus it is possible, to speedup even complex experiments.

In addition to that, the V100 was equipped with heat sinks that facilitate transmitting the heat generated by the GPU directly to the hot water cooling system. From there, the heat is directly handed over to heating consumers of Cloud&Heat's data centers. For example, this can be hotels and offices as it is the case in the Eurotheum in Frankfurt[4].

To be able to use all GPUs at reasonable bandwith, a board of type "X11DGQ" from Supermicro is used. This is embedded into a 1U Chassis, that accommodates space for four GPUs and, hence, enables a huge power density. In addition, this package can be flexibly scaled on demand.

Currently, the GPUs are directly handed over from the host system to the virtual cloud instances of nyris. In a next step, this platform shall be extended by one layer of abstraction that then enables further flexibility. Technologically this could mean the usage of Kubernetes, which is a Tool for orchestration of

containerized applications: It provides options for easily deploying, scaling and managing application containers[5]. Nyris already uses this form of abstraction to carry out operations flexibly and platform-independent.

Benefits

But what are the benefits of such an infrastructure? In a whitepaper that will be published during the next weeks, Cloud&Heat transparently shows the difference between their infrastructure and a traditional data center infrastructure. Based on this data, we will already give an insight on the comparison based on the use case described above.

The foundation for our calculation is the data center and it's infrastructure operated by Cloud&Heat in the Eurotheum in Frankfurt. It has a total IT-power of 500kW. By utilizing Cloud&Heat technology, savings in the following areas can be achieved:

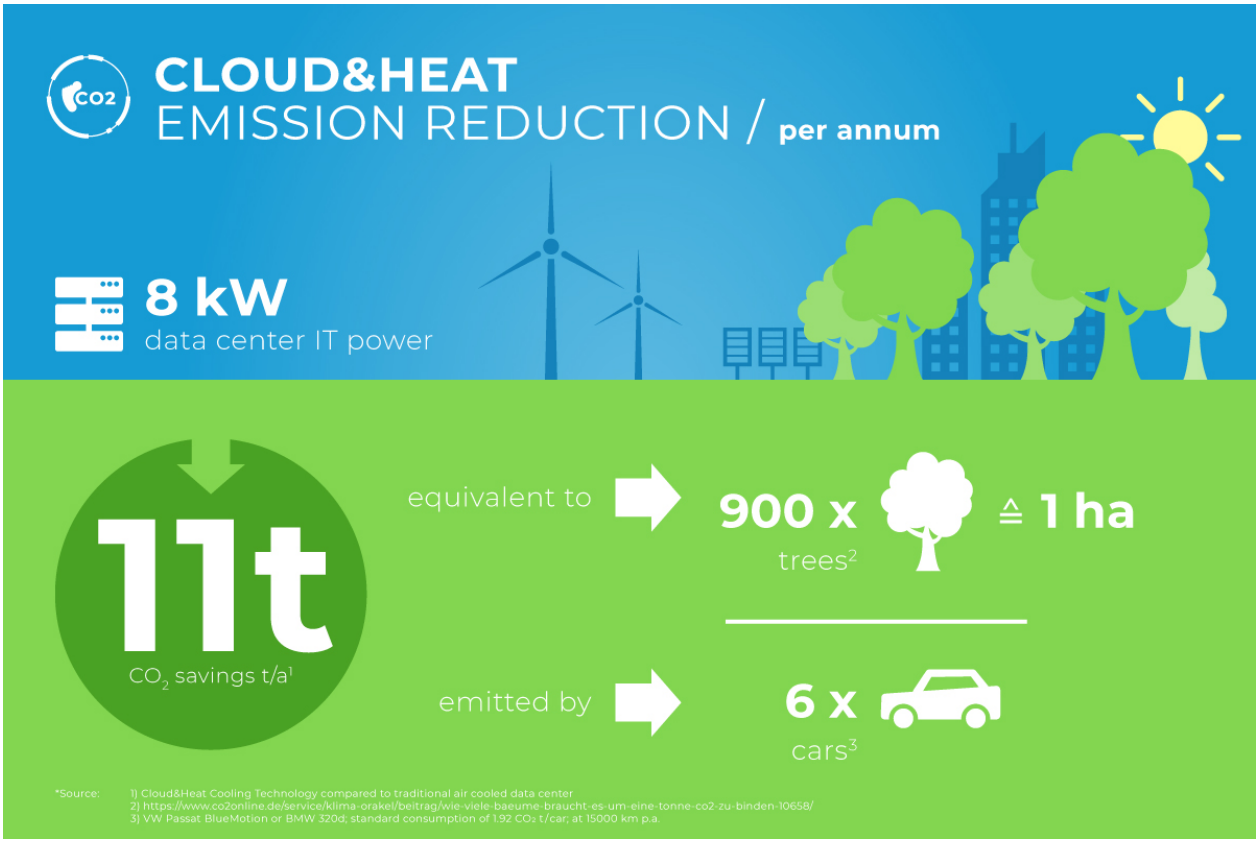
- reduction of power consumption by eliminating onboard-fans
- increased energy efficiency by water cooling technology
- reduction of emissions by reusing waste heat

To calculate the CO2-saving, the Cloud&Heat data center is compared to an air-cooled data center without waste heat reuse and the same total IT-power.

In addition, it is assumed that four chassis are equipped with four V100 GPUs each, such that an extremely compact 16 GPU cluster is created on only four rack units. This cluster has an overall IT-power of 8kW, since every chassis has an IT-power of 2kW. Furthermore, we assume an utilization of 50% which is a conservative estimation since nyris currently reports a utilization of their hardware of roughly 80%.

On the lines of the calculation done in the whitepaper, a saving of roughly 11t CO2 per year can be achieved. This is equivalent to the emissions of 6 German cars per year[6]. To compensate this emissions, one hectare forest[7] (10000m2) or 900 deciduous trees[8] would be required.

Altogether, Cloud&Heat and nyris developed a solution, that does not only save CO2 emissions, but also saves space: Both in a data center and when compensating emissions.



Vision

Although these measures are a big step towards energy efficiency and sustainability, data centers still provide huge potential for optimizations. For example, heat should only be generated by workload if it can be used at the corresponding data center site. Otherwise it would need to be cooled down and released to the environment costly. Hence, in most of the cases it makes sense to operate several geographically distributed data centers and to combine them into a virtual data center federation. For managing of such a complex infrastructure and for deploying workloads to these distributed sites, Cloud&Heat is developing a tool[9], which was transferred to an open source project last year[10]. This tool ensures that containerized applications are run at this data center site, where it is most meaningful from an energetic point of view and even migrates applications on demand. The applications of nyris could also be run on such an infrastructure even more efficiently.

Machine Learning applications are on the rise while new business areas are explored, and, they cause an inevitably rising demand for compute capacity. Whith innovative solutions and energy efficient, intelligent infrastructures we can come up against this trend in a sustainable way.

The sketched scenario based on practical numbers clearly shows the necessity to improve the sustainability of IT infrastructures. Although awareness for energy efficient solutions is increasing on the provider-side, the upgrade to a green cloud seems to be on hold for several reasons. In the course of rethinking, the cloud users are requested to do their part as well: Everyone should consider moving to green solutions for fulfilling their compute demand to reduce the global CO2 footprint holistically. The goal is clear: We need green IT solutions to preserve our planet.

If we manage to do this wisely, we will be able to cut the increasing emissions coming from Machine Learning and to sustainably influence our lives, our society and our planet using this technology. Together with Cloud&Heat, nyris came a bit closer towards the goal to digitize the amazing capabilities of the human eye more efficiently and sustainably.

1) https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/
2) https://www.pwc.co.uk/services/sustainability-climate-change/insights/how-ai-future-can-enable-sustainable-future.html
3) https://www.boston-it.de/blog/2019/04/15/test-drive-tesla-v100.aspx?utm_source=rss&utm_medium=syndication&utm_campaign=rss
4) https://www.datacenter-insider.de/cloudheat-uebernimmt-ehemaliges-rechenzentrum-der-ezb-in-frankfurt-a-613373/
5) https://de.wikipedia.org/wiki/Kubernetes
6) VW Passat BlueMotion oder BMW 320d; typische Emissionen 1,92tCO2/a bei 15.000km/a Laufleistung
7) https://www.baysf.de/de/wald-verstehen/wald-kohlendioxid.html
8) https://www.co2online.de/service/klima-orakel/beitrag/wie-viele-baeume-braucht-es-um-eine-tonne-co2-zu-binden-10658/
9) https://gitlab.com/rak-n-rok/krake
10) https://www.cloudandheat.com/release-the-krake/