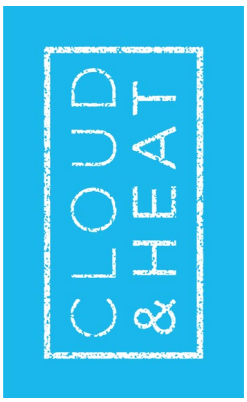


Cloud&Heat & Nyris

Nachhaltige künstliche Intelligenz
für einen grünen Planeten

Nachhaltige künstliche Intelligenz für einen grünen Planeten

Rechenzentren gehören aufgrund der in allen Bereichen ansteigenden Digitalisierung zu den größten Stromfressern. Mit der Digitalisierung geht auch ein steigender Bedarf an Rechenkapazität und damit eine Zunahme des Strombedarfs von Rechenzentren einher. Doch was verursacht eigentlich das immens hohe Aufkommen an Rechenjobs? Neben Web-Anwendungen, wie dem Streamen von Videos oder dem Browsen auf Online-Shopping-Plattformen gibt es weitere Anwendungen, die bestimmte Arten von Rechenzentren zu Stromfressern machen: Machine Learning, Künstliche Intelligenz oder das Internet of Things werden in Zukunft sogar noch deutlich mehr Strom verbrauchen und damit auch mehr schädliche Emissionen erzeugen als bisher.

Dabei verursacht beim Machine Learning der Betrieb der Online-Plattformen nur einen Bruchteil des Stromverbrauchs. Deutlich rechenintensiver ist das sogenannte Training, mit dem die Machine-Learning-Modelle trainiert, also im Prinzip "angelernt", werden. Aktuellen Studien zufolge kann bereits jetzt durch dieses Training pro Durchgang so viel CO2 emittiert werden wie von fünf amerikanischen Autos in deren gesamten Leben[1]. Dennoch bietet Machine Learning auch große Chancen, zum Beispiel durch Steigerung der Effizienz von Produktion und Verbrauch und damit einer nachhaltigeren Wirtschaft und Gesellschaft[2]. Es ist also notwendig, Rechenzentrumsinfrastrukturen effizienter und nachhaltiger zu gestalten, um die ökologischen Herausforderungen des wachsenden

Nyris - giving search the power of sight

Ein Unternehmen, welches solche Infrastrukturen für das tägliche Geschäft benötigt, ist nyris. Nyris strebt danach, die Fähigkeit des Sehsinns zu digitalisieren. Dazu wird künstliche visuelle Intelligenz genutzt, die Suche nach Ersatzteilen, Produkten oder Objekten natürlicher und intuitiver zu ermöglichen. Doch neben breitem Fachwissen, Einfallsreichtum, Kreativität und Leidenschaft erfordert dies auch sehr viel Rechenkapazität und damit verbunden sehr viel Energie, um die dafür benötigten komplexen Machine-Learning-Algorithmen zu entwickeln, zu trainieren und zu nutzen.

Cloud&Heat – the most energy efficient data centers

Die Rechenkapazität für diese aufwendigen Berechnungen stellt das Dresdner Green-IT Unternehmen Cloud&Heat bereit. Das Startup, welches ausgehend von der Idee, Server-Abwärme zum Heizen zu nutzen, gegründet wurde, baut und betreibt Rechenzentren mit einer innovativen Heißwasser-Direktkühlung. Der auf flexibel skalierbaren GPU-, CPU-und Storage-Angeboten basierende Infrastructure-as-a-Service (IaaS) bietet optimale Bedingungen für AI-Unternehmen wie nyris.

Der Use-Case

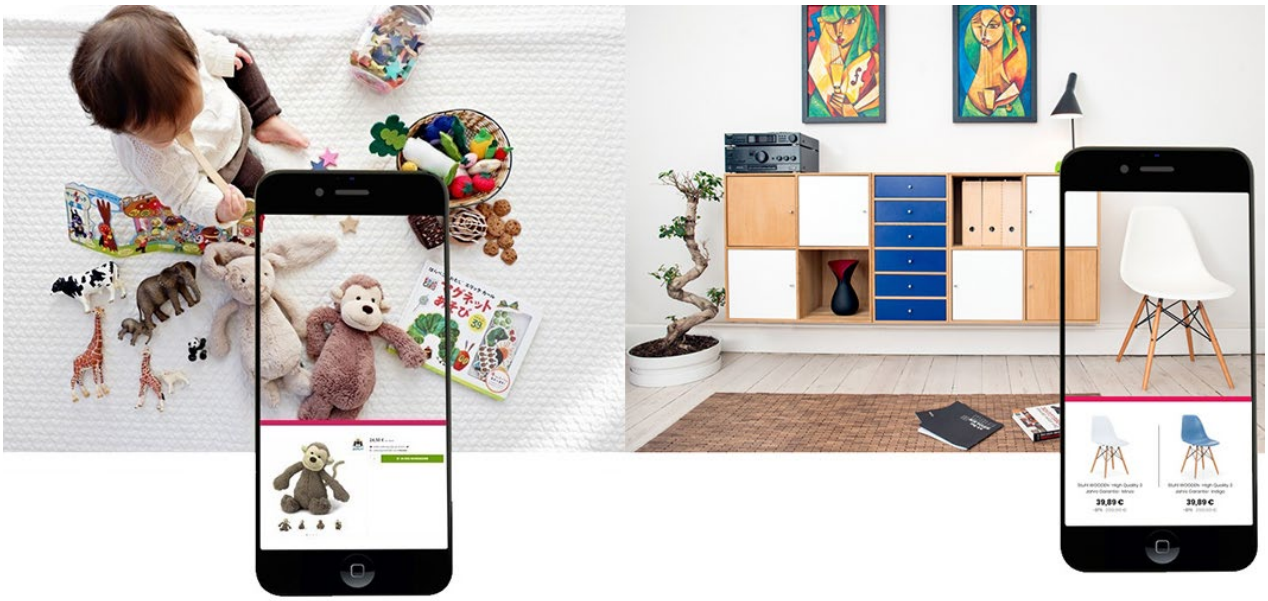
Dieses stand Ende 2019 vor der Frage, in welcher Form Compute-Infrastruktur zukünftig genutzt werden soll. Da nyris sich seiner Verantwortung als KI-Unternehmen bewusst ist und damit großen Wert auf Energieeffizienz und Nachhaltigkeit legt, fiel die Wahl nach einer ausgiebigen Evaluierung der Preis- und Leistungsverzeichnisse verschiedener Cloud-Anbieter auf Cloud&Heat. Hier wird bei geringem Preis aktuelle Hardware, ein hohes Maß an Rechenleistung und Flexibilität, sowie eine ökologisch sinnvolle, ganzheitliche Lösung angeboten, die nyris überzeugte.

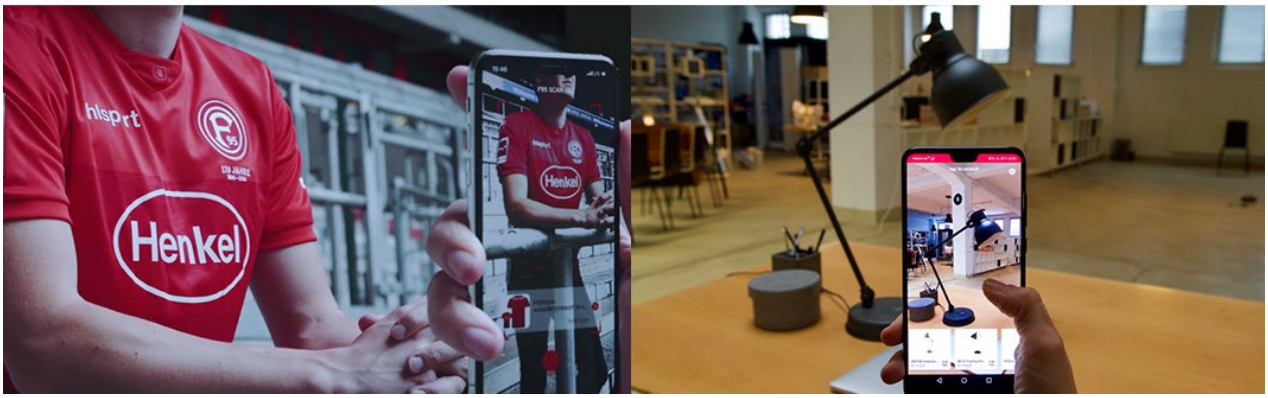
Die Architektur neuronaler Netze und die von nyris genutzten Trainings- sowie Optimierungsmethoden haben sich in den letzten Jahren enorm weiterentwickelt. Gerade der Bereich des Distance Metric Learnings, einem Unterbereich des maschinellen Lernens, welcher von besonderer Bedeutung für visuelle künstliche Intelligenz ist, stellt an die Hardware eine besondere Herausforderung dar. Diese Netze werden nicht mit einzelnen Bildern sondern mit Bilderpaaren oder Bildertriplets trainiert. Während des Trainings kommen viele Transformationsfunktionen hinzu, werden Stichproben zur Optimierung einer Zielfunktion in eine kompakte Darstellung umgewandelt oder komplexe Verlustfunktionen verwendet.

Eine weitere Notwendigkeit ist das optimale Zusammenspiel aus dem verwendeten Software Framework und der CPU-/ RAM- / GPU-Hardware. Obwohl viele Anbieter mit eigenen Chips in diesen Markt drängen, bleiben die GPU-Systeme von Nvidia aktuell weiter die am flexibelsten einsetzbaren Systeme, was gerade in der Forschung an neuen Architekturen ein entscheidender Vorteil ist. Aber auch hier müssen alle weiteren Hardwarekomponenten optimal aufeinander abgestimmt sein. Beim Einsatz

von 8 oder 16 GPUs kann der System-Arbeitsspeicher oder die CPU zu schnell zum Flaschenhals werden, falls nicht genügend Daten an die GPUs übertragen und diese damit nicht vollständig ausgelastet werden können. Dies führt nicht nur zu langen Trainingszeiten mit damit verbundenen hohen Gesamtkosten, sondern auch zu viel Frust bei den verantwortlichen Data Scientists. Es ist, als würde man für einen Porsche zahlen, um im Feierabendverkehr Berlins zu stehen.

Cloud&Heat arbeitet hier eng mit dem Kunden zusammen und findet maßgeschneiderte Lösungen für die jeweilige Anwendung. So können beispielsweise auch große Zahlen von GPUs auf speziellen Boards zusammengeschaltet werden, um eine maximale Auslastung zu ermöglichen.





Die Infrastruktur

Aber warum braucht man für das Training überhaupt GPUs? Beim Training von neuronalen Netzen sind viele wiederholende, einfache Rechenoperationen notwendig. Dies wird vor allem durch die Verwendung von spezialisierten Chips erreicht. In den meisten Fällen werden heute GPUs dafür verwendet, es werden aber auch eine Vielzahl von neuen Chips für das Training von neuronalen Netzen entwickelt (Google TPUs, Graphcore IPUs, Intel NNPs, Huawei NPUs etc.). Im Fall einer Nvidia V100 GPU besteht diese z. B. aus 5120 Kernen, welche einfache Rechenoperationen durchführen können. Eine Intel Xeon-Platinum-8180-CPU hat zum Vergleich nur 28 Kerne, die jedoch weitaus komplexere Operationen ausführen können. Dadurch sind auf einer GPU deutlich mehr gleichzeitige, aber dafür einfache, Operationen möglich.

Als Hardwarebasis wurde für das vorliegende Setup zunächst möglichst performante Hardware ausgewählt: Die Tesla V100 von Nvidia ist eine speziell für AI-Anwendungen entwickelte Grafikkarte und bietet exzellente Performance und Effizienz. Dadurch ist es möglich, auch komplexe Experimente viel schneller durchzuführen. Außerdem wurde die V100 mit Kühlkörpern ausgestattet, welche es ermöglichen, die von der Grafikkarte erzeugte Wärme direkt in den Heißwasserkühlkreislauf zu überführen. Von dort wird die Wärme direkt an heizungsseitige Abnehmer der Rechenzentren von Cloud&Heat, wie beispielsweise Hotels und Büroräumen im Frankfurter Eurotheum, weitergegeben.

Um alle Grafikkarten mit ausreichender Bandbreite ansprechen zu können, kommt ein Mainboard vom Typ “X11DGQ” der Firma Supermicro zum Einsatz. Dieses ist eingebettet in ein 1U Chassis, welches Platz für vier GPUs bietet und somit eine enorm hohe Leistungsdichte ermöglicht. Dieses Package kann dann den Anforderungen nach flexibel skaliert werden.

Aktuell werden die Grafikkarten vom Host-System direkt an die virtuellen Cloud-Instanzen von nyris durchgereicht. In einem nächsten Schritt soll diese Plattform allerdings um eine Abstraktionsschicht erweitert werden, welche eine größere Flexibilität ermöglicht. Als Technologie kann hier Kubernetes zum Einsatz kommen. Kubernetes ist ein Tool zur Orchestrierung von Anwendungs-Containern, mit dem diese einfach bereitgestellt, skaliert und verwaltet werden können. Nyris nutzt diese Form der Abstraktion, um Berechnungen flexibel und plattformunabhängig durchführen zu können.

Vorteile

Wo liegt nun der Vorteil dieser Infrastruktur? In einem in den nächsten Wochen zu veröffentlichen Whitepaper von Cloud&Heat will das Unternehmen transparent den Unterschied seiner Infrastruktur zu klassischen Rechenzentrumsinfrastrukturen darstellen. Wir werden anhand dieser Daten bereits jetzt einen Vergleich basierend auf dem beschriebenen Use-Case durchführen.

Als Berechnungsgrundlage dient das von Cloud&Heat im Frankfurter Eurotheum betriebene Rechenzentrum und dessen Infrastruktur. Es hat eine IT-Gesamtleistung von 500kW. Durch Nutzung der Cloud&Heat-Kühlung werden Einsparungen in den folgenden Bereichen erzielt:

- Reduktion Stromverbrauch durch Wegfall von Onboard Lüftern
- Steigerung der Energieeffizienz durch Wasserkühlung
- Reduktion der Emissionen durch Abwärmenutzung

Zur Berechnung der CO_2-Einsparung wird als Vergleichswert ein luftgekühltes Rechenzentrum ohne Abwärmenutzung mit gleicher IT-Gesamtleistung angenommen. Für die Modellrechnung wird außerdem angenommen, dass vier Chassis mit je vier V100 GPUs ausgestattet werden, sodass auf 4 Höheneinheiten ein extrem kompaktes Cluster aus 16 GPUs entsteht. Pro Chassis wird dabei mit einer IT-Leistung von 2kW gerechnet, sodass das Cluster eine IT-Gesamtleistung von 8kW hat. Weiterhin wird eine Auslastung von mindestens 50% angenommen, welche sich als sehr realistisch darstellt (laut Aussagen von nyris liegt die derzeitige Auslastung der Hardware bei ca. 80%).

Analog zur Berechnung im Whitepaper ergibt sich eine Einsparung von rund 11t CO_2 pro Jahr. Dies entspricht dem CO_2-Ausstoß von rund 6 Autos pro Jahr. Um diesen Ausstoß anderweitig zu kompensieren, würden ein Hektar Wald bzw. ca 900 Laubbäume benötigt. Insgesamt haben Cloud&Heat und nyris hier also eine Lösung entwickelt, die nicht nur CO_2-Emissionen, sondern auch Platz einspart - sowohl im Rechenzentrum, als auch bei der Flächennutzung zur CO_2-Kompensation.

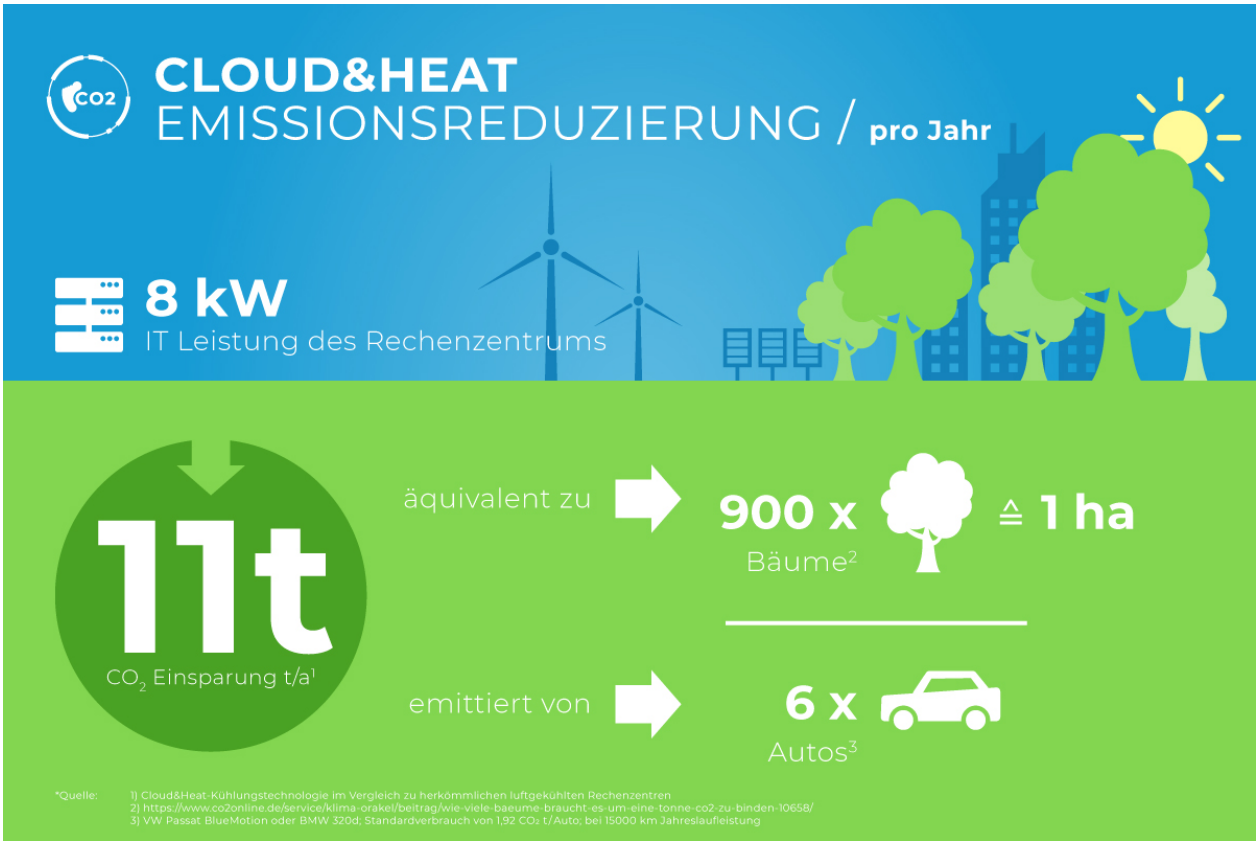
Vision

Auch wenn diese Maßnahmen bereits einen großen Schritt in Richtung Energieeffizienz und Nachhaltigkeit darstellen, bieten Rechenzentren weiteres Optimierungspotenzial. Wärme sollte beispielsweise nur dann durch Rechenlast produziert werden, falls sie am jeweiligen Rechenzentrumsstandort auch abgenommen werden kann und benötigt wird. Andernfalls müsste sie aufwendig heruntergekühlt und ungenutzt an die Umgebung abgegeben werden. Deshalb ist es meist am sinnvollsten, mehrere geografisch verteilte Rechenzentren zu betreiben und diese zu einem virtuellen Rechenzentrumsverbund zusammenzuschließen. Zum Management dieser komplexen Infrastruktur sowie zur energieeffizienten Verteilung der Rechenlasten entwickelt Cloud&Heat ein Tool, welches letztes Jahr in ein Open-Source-Projekt überführt wurde. Dieses führt containerisierte Anwendungen dynamisch genau an dem Standort aus, wo es energetisch am sinnvollsten ist und migriert diese bei Bedarf. Auch die beschriebenen Anwendungen von nyris könnten auf einer solchen verteilten Infrastruktur noch effizienter ausgeführt werden.

Machine-Learnig-Anwendungen sind aufgrund neuer Geschäftsfelder auf dem Vormarsch und verursachen einen unaufhaltsam steigenden Bedarf an Rechenzentrumspower. Durch innovative Lösungen und energieeffiziente, intelligente Infrastrukturen ist es jedoch möglich, diesem Trend nachhaltig zu begegnen.

Das skizzierte Szenario zeigt dabei basierend auf konkreten Zahlen die Notwendigkeit sehr deutlich, IT-Infrastrukturen nachhaltiger zu gestalten. Obwohl das Bewusstsein für energieeffiziente Lösungen vor allem auf Provider-Seite zunehmend geschärft wird, scheint das Upgrade zur grünen Cloud aus verschiedenen Gründen noch auf sich warten zu lassen. Im Zuge des Umdenkens sind jedoch auch die Cloud-Nutzer angehalten, abhängig von anderen Faktoren, vornehmlich auf grüne Lösungen zu setzen, um den weltweiten CO2-Fußabdruck ganzheitlich zu senken. Der Fokus ist klar: Es braucht grüne IT-Lösungen, um unseren Planeten zu erhalten.

Wenn wir das clever anstellen, können wir die immer weiter steigenden Emissionen von Machine Learning eindämmen und mithilfe dieser Technologie unser Leben, unsere Gesellschaft und unseren Planeten nachhaltig positiv beeinflussen. Dem Ziel, die unglaublichen Fähigkeiten des menschlichen Auges zu digitalisieren, ist nyris zusammen mit Cloud&Heat in jedem Fall bereits effizient und nachhaltig nähergekommen.



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