REPLY

MAKING CLOUD COMPUTING MORE SUSTAINABLE

GREEN CLOUD

The (hyperscale) cloud provides many advantages over on-premises hosting or traditional data centers. One of them is a higher energy efficiency. Nevertheless, there are some basics that every person working in IT should keep in mind to make the cloud usage greener. Start using the cloud in a more environmentally friendly way today - we show you how.

REPLY [EXM, STAR: REY] is specialises in the design and implementation of solutions based on new communication channels and digital media. As a network of highly specialised companies, Reply defines and develops business models enabled by the new models of AI, big data, cloud computing, digital media and the internet of things. Reply delivers consulting, system integration and digital services to organisations across the telecom and media; industry and services; banking and insurance; and public sectors.

www.reply.com



CONTENT

The quest for Green Hosting 4

Cloud vs on premises: why is the cloud the greener choice? 6

Cloud architecture: key considerations for design decisions 10

Workload scheduling algorithms: reduce the cloud carbon footprint 13

Conclusion 18

Towards the Green Cloud with Reply 19

THE QUEST FOR GREEN HOSTING

The relentless digitalization of the economy and society is leading to a dynamic increase in the amount of data, stored in data centers, and more computational powers compared to the past to process all the new data and services. For instance, according to the market research reports from Fortune Business Insights[1], the global cloud storage market is projected to grow from $83.41 billion in 2022 to $376.37 billion by 2029, at a CAGR (Compound Annual Growth Rate) of 24.0% in the forecast period.

It follows without saying that the overall data centers resource and energy consumption (for processing, storage, cooling systems, lights, etc.) are key aspects to take into consideration for the next years. GreenOps can be very useful for this. By bringing development, finance, and business teams together, GreenOps drives business and financial accountability, and optimizes the carbon footprints of all cloud-based applications. The reduction and avoidance of emissions is particularly relevant here. The emissions can be divided into different categories. A particularly challenging one: indirect emissions. These occur along the entire value chain and also concern the activities of other players. Indirect emissions include, for example, those caused by the use of an external data center.

Those indirect emissions play an important role in this white paper, as chapter 2 explains how sustainable it is to run software in the hyperscale cloud in comparison to traditional data centers or on-premises hosting. It also explores how energy consumption can be reduced by using the hyperscale cloud. Furthermore, this whitepaper provides insights into best practices on how to use the public cloud in an environmentally friendly, sustainable way in day-to-day work. Chapter 3 illustrates how efficient cloud architecture can reduce energy consumption even more by optimizing the usage of the used resources. Chapter 4 investigates workload scheduling algorithms, which can further enhance the resource usage of the public cloud.



WHY IS THE CLOUD THE GREENER CHOICE?

CLOUD VS. ON PREMISES DATA CENTERS

The “Global data centers Energy Demand by Type 2015-2021” report from IEA2 related to data centers energy consumption, found that traditional data centers globally have decreased their energy demand, from around 97.6 Terawatt-hours in 2015 to around 50 Terawatt-hours in 2019. However, incidentally, hyperscale data centers have doubled their energy demand in the same period of time thanks to a wide adoption of new Public Cloud paradigms (like managed services and serverless architecture) together with numerous mass migration and modernization projects from traditional on-premises data centers.

Since the overall sum of data center energy consumption is still similar despite the huge increase of data and computational powers due to the digitalization growth, as depicted in the following Figure, it follows that the hyperscale data centers are more efficient and have been able to mitigate the power “greediness” of IT growth of the recent years.

Moreover, the diagram below shows that the energy consumption of traditional on-premises data centers and the Private Cloud (non-hyperscale, where the pricing model is typically not on-demand, and there are only very limited managed services), has been reduced as well. This is not because these traditional approaches have become more efficient, but because of the continuous workload migration towards hyperscale data centers. I.e., more and more enterprises abandon traditional approaches in favour of the hyperscale cloud.

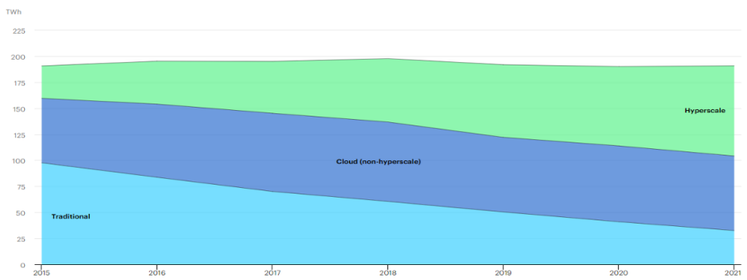


Figure 1: Global data center energy demand by data center type, 2015-2021, IEA report [2]

THE GREEN COMMITMENT OF THE HYPERSCALERS

In addition to that, the three biggest hyperscalers Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) have each set themselves goals to support the reduction of the environmental footprints of projects running on their infrastructure. Therefore, they have set for their future development they are trying to head in a greener direction: an essential element of all their strategies is the transition to 100% renewable energy.

All the hyperscalers have invested a huge effort in building and transforming their data centers. The main goal is to considerably increase the overall efficiency while reducing carbon footprint through green power sources.

Even though “green” energy is an element of all hyperscalers’ strategies, power consumption still represents a challenge for the future. To address this, the hyperscalers have built their data centers addressing the following aspects:

- **Hardware technology efficiency:** always investing a big amount of effort to renovate the hardware (e.g., adopting the new generation of Intel, AMD, and ARM processors). AWS, for instance, has developed a new processor based on ARM architecture that consumes up to 60% less energy than the same traditional virtual machines based on Intel processors.

- **Multi-tenant approach:** deeply relying on shared hardware (giving the customers the possibility to have dedicated ones, if requested) and best-of-breed hypervisors to achieve high performance, security isolation, and power consolidation.

- **Cooling system**: hyperscalers are generally quite new and are adopting efficient new cooling technologies. For instance: Microsoft is using a liquid immersion cooling infrastructure at Azure data centers in Washington, which saves water and lowers energy costs like warm water cooling. AWS and GCP are maximizing the use of free-air cooling systems that cool servers with outside air without using any additional water.

- **Software scalability:** developing solutions that can heavily rely on new paradigms like serverless, where the code is directly run into highly scalable ephemeral containers, and managed services, where the customers only see the platform or the software and, behind the scenes, the hyperscalers are able to efficiently scale the resources through custom solutions.

With these approaches, the hyperscale cloud has been able to satisfy the growing computational needs of the past ten years with only little increased energy demands. However, in the future, we cannot expect this trend to last. Most hyperscale data centers operate at near-optimal energy efficiency, so further improvements to the infrastructure can only have minor impacts. Additionally, to migrate to the hyperscale cloud, we also need to improve the software itself.

To understand this point better, imagine that the energy consumption of the cloud (and related workloads running on it) equals the fuel consumption of a typical speedway with cars, trucks and other vehicles running on it. If we want to reduce overall fuel consumption, we must build very efficient cars that leverage on green energy, we can improve aerodynamics, reduce ground friction, define better speedway tracks by reducing obstacles and crosses, etc.

However, the driver's behaviour and driving style are relevant variables that can be a real game changer for consumption. The same situation happens with the hyperscale-cloud, where, on the one hand, the highway and car improvements match the data center facilities, the hardware, and the virtual systems inside it. On the other hand, the behaviour of the driver is like the code running in the systems. If we can improve the behaviour of the code, making it very efficient, we can address this final aspect achieving the best solution possible. The field that is addressing this aspect to improve code efficiency is called Green Coding.

The main goal when it comes to efficient data centers is to increase the overall efficiency while reducing the carbon footprint through green power sources.

Hyperscale data centers are more efficient and can mitigate a part of the power "greediness" of growing computation demands.

AWS, Azure and GCP, as the largest hyperscalers, have set themselves sustainability goals including the shift to renewable energy for powering their data centers.

Efficiency gains in the hyperscale cloud infrastructure are achieved through: hardware technology efficiency, multi-tenant approaches, new cooling technology and improved software scalability (e.g., through serverless or managed servic

Green Coding is a key aspect to make running software more energy and resource-efficient.

AT A GLANCE

CLOUD ARCHITECTURE: KEY CONSIDERATIONS FOR DESIGN DECISIONS

As mentioned previously, hyperscalers can offer efficiency improvements over small-scale and traditional data centers. These efficiencies can be further improved upon by using best practice design choices that focus on minimising unnecessary resource utilisation and maximising the sharing of common tasks. It is very difficult, if not impossible, to get an accurate measurement of how design patterns affect the carbon emissions of general cloud usage with various cloud providers implementing their services differently, with different technologies and not releasing their power consumption data. Therefore, the following section will consist of points to consider when making design decisions and not concrete rules.

The choice of computing resources can make a huge difference to the carbon footprint of an organisation's operations. The aim here is to make the most efficient use of available resources with as little overhead as possible. Services such as managed Kubernetes can offer features such as full customisation, advanced networking capabilities and ease of scalability. However, these benefits mean that processes are constantly running to complete health checks, run automated tasks and various other routines that could be managed by a cloud provider and shared across multiple users. Even without Kubernetes, reserving compute resources and having them running idly will still contribute to overall emissions. This is also a waste of financial capital. By using the example of a simple web server that is irregularly receiving requests we can see that this may be mitigated by using a serverless solution such as GCP Cloud Functions or AWS Lambda instead of running a virtual machine constantly.

This is not to say that serverless is always the most efficient method. Using the example of a web server, if there is a constant stream of incoming requests then the overhead of provisioning a lightweight function and then releasing that resource could be less efficient than having an appropriately sized virtual machine. By benchmarking the performance of serverless computing, there is usually a cold start period whereby a process that has not been called for a while may take longer to start up and execute. This may result in processes upstream waiting for a response and again not making efficient use of resources. Along with constantly evolving technology across providers, this adds evidence to the point that making a design decision that minimises resource utilisation is extremely difficult.

Other resource usages can also be optimised by considering the above. Products such as cloud Big Table will spin up multiple compute resources to run on and have a minimum usage, which in this case is three instances – completely oversized for a lot of applications.

Networking is often a factor that is overlooked when considering resource usage. Transmitting data will generate emissions from routers, modems, and switches as well as the computer resources that are used to verify and authenticate the packages sent. If data is encrypted, then this will also require additional computing resources to encode and decode. Arguably, these usages are small but when they are multiplied by the vast amounts of data being transmitted then it adds up. An easy way to improve efficiency, and likely performance as well, is to host computer resources in the same location as the data that is being processed. This can be in the form of having everything hosted on the same cloud (private, public or the same vendor) or if this is already the case having both resources in the same geographical region.

Each cloud provider has their resources located in multiple places across the globe, a simple way of reducing carbon emissions in cloud projects is to choose a location that is using the latest green technology to run their data centers. For example, some locations are using power generated from hydrothermal sources which has a considerably lower impact on the environment than electricity generated from fossil fuels. These locations are usually well publicized by each provider and can be a good choice for those who have the goal of minimizing their carbon emissions. Bearing this in mind, consideration will have to be made about any practical or legal implications of using resources and storing data in certain areas.

Overall, it is very difficult to quantify energy usage accurately and it is likely that other design requirements will take priority over the reduction of a project's environmental impact. However, using the advice above, it is possible to make some significant improvements by considering the best practices of cloud architecture.

Green Cloud architecture best practices focus on design choices that minimise unnecessary resource utilisation and maximise the sharing of common tasks.

Choosing the greenest solution can be complex due to its dependency on the individual solution setup and requirements.

Managed computing services might be producing a lot of overhead due to constant monitoring (e.g., health checks) or high availability.

Serverless is not always the most efficient solution, because it has a cold start period whereby a process that has not been called for a while may take longer to start up and execute.

Transmitting data through networking generates lots of emissions. The same goes for data encryption.

To reduce network emissions, hosts compute resources in the same location as the data is being processed.

Depending on the cloud provider: choose a location that is using the latest green technology to run their data centers.

AT A GLANCE

WORKLOAD SCHEDULING ALGORITHMS: REDUCE THE CLOUD CARBON FOOTPRINT

WORKLOAD SCHEDULING ALGORITHM

This section aims to explore the effects of carbon-aware workload scheduling. These can be used to reduce the carbon emissions of the workloads when used by a cloud user. On a high level, there are two main dimensions to consider when scheduling workloads; temporal and spatial. Temporal scheduling (when used in the context of carbon emissions) is a term used to refer to the act of scheduling your workloads at a time when the energy being used by that workload is likely to come from low-carbon sources such as Biopower, Solar Energy, Geothermal Energy, Hydropower, Wind Energy and Nuclear energy. When talking about the spatial dimension the implication here is cleaner, run your workloads where it is most carbon efficient and the carbon intensity for power is lower.

TEMPORAL WORKLOAD SHIFTING[3]

IDENTIFY SUITABLE WORKLOADS

Shifting workloads over time is an interesting lever for making cloud computing greener. But not every workload is equally suitable for this. To use the possibility of shifting effectively, one should first be aware of the three relevant characteristics of workloads: duration, execution time and interruptibility.

The duration can be either short-running or long-running such as machine learning workloads or big data tasks. Tasks that do not fall into these two categories fall under the “continuously running workloads” category and can be things such as blockchain mining or brute force attacks, which often do not have a defined end time.

For the execution time category there is either “ad hoc” meaning the start time is unpredictable or random like for arbitrary

software compiles of developers vs. workloads of a scheduled nature such as nightly builds or scheduled batch jobs.

Lastly was the interruptibility aspect of workloads. Is the workload one, that can be split up into multiple runs because it could be paused and picked up later due to a state saving mechanism or is it one that cannot be interrupted because it might be part of a CI/CD pipeline or database backup that needed to be executed in one go to avoid inconsistency in data. Now that we know the distinctive characteristics of workloads, the authors took the next step to study the carbon intensity (gCO2/kWh) of the 4 regions over time to spot patterns.

CARBON INTENSITY OVER TIME

Depending on the location, a different time shift of the workloads is recommended. The following figure illustrates the carbon intensity during 24 hours for Germany, the UK, France and California:

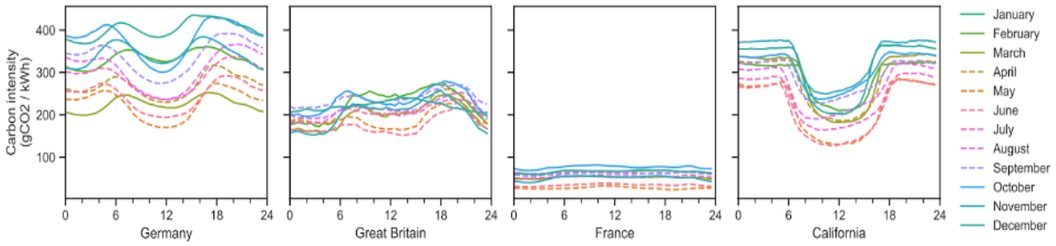


Figure 2: Carbon intensity comparison between Germany, Great Britain, France and California.[3]

Although Germany obtains about 30% of its energy by renewable means from wind and sun, it must be noted that the remaining energy demand is covered in a relatively “dirty” way by burning lignite and black coal and fossil gas. This conclusion about Germany is that the carbon intensity for power is usually lowest around mid-day when most solar energy is available and at 2 AM when there is less demand for electricity and the fossil power plants are operated at a lower capacity.

Great Britain is like Germany with less solar power however at around 18% nuclear energy the mean carbon intensity is lower and stays equal over the year. During the night time, the carbon intensity is the lowest but does not drop as much during the day due to lower amounts of solar energy. France has over 65% of its energy mix coming from nuclear and around 8% from hydroelectric power which both have very low carbon emissions. This leads to very steady and low mean gCO2/kWh throughout the day and year.

California’s graph is clearly shaped by the abundance of sunshine it receives. While the state generates 13.4% of its power from Solar energy, this figure jumps to 30.9% between 8 AM and 4 PM. Much of their other energy is imported from other “dirtier” states which is the reason for the high mean carbon intensity, but it makes for a great illustrative use case for temporal workload shifting.

SPATIAL WORKLOAD SHIFTING

Spatial workload shifting in the carbon context is the process of moving around workloads to data centers where the carbon intensity for power is low. Some data centers are directly attached to renewables, and some are just “less dirty “, i. e. running on less brown energy than the data center the workload is originating from. Lastly, data centers are often located in a region or country where the energy mix of the energy provider of that data center is made up of a larger share of renewable energy compared to your average data center.

This last point is nicely illustrated in Google’s “Google Cloud Region Picker” which suggests a data center based on your selected trade-offs and preferences.

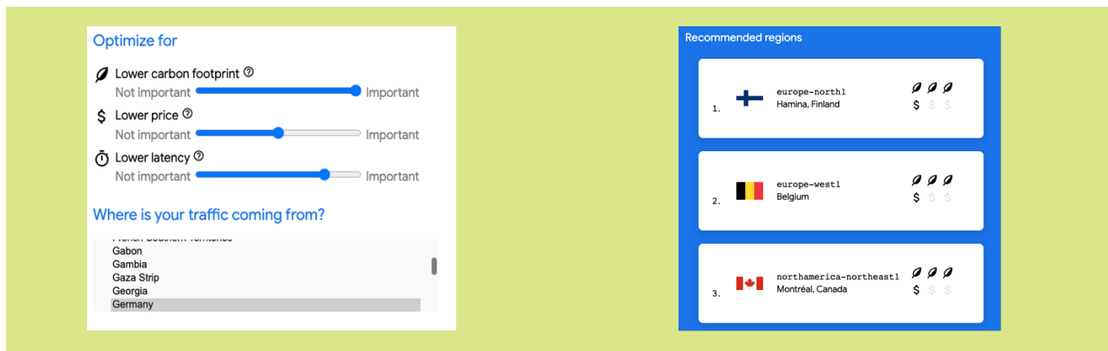


Figure 3: Google Cloud Region Picker.[4]

This is of course only an initial step when picking a data center. In an ideal world, workloads are so flexible and location agnostic that they can be moved seamlessly from one data center to another and maybe even across cloud providers at will, based on a list of preferences, rules and optimization decisions.

Some workloads do not have these properties. There are workloads like high-frequency trading algorithms that must run as close as possible to certain stock exchanges to execute trades profitably, these cannot be moved at will. Other workloads are restricted by data residency requirements and require their data to be stored in a certain geographical region or jurisdiction. These are just two aspects in an extensive list of latency, hardware and preferences that usually take precedence over the carbon footprint of that workload.

Let's imagine we live in a world where these are all clearly defined, and we also have a list of all data centers that match this criterion. We now are faced with the problem of data. Not literal data but data about the energy mix of the data centers. As of Q4 2022, none of the three main hyperscaler cloud providers provided accurate or rich data about their energy mix or carbon footprint of their services. This makes things exceptionally difficult.

Only now can we begin architecting and refactoring our workloads and application to be flexible and reschedulable to a point where uptime and latency are not heavily negatively impacted if they were to shift to a different geographic region. With entire virtual machines and virtual machine images this could be a bit easier. We mirror the virtual machine to another, more carbon-efficient region, and once it passes all health checks, we can reroute traffic to this new machine via something like a blue-green deployment as we shut off the old machine. On the other hand, if we have microservices that need to talk to each other, we need to make sure we are not breaking off a small integral piece of a larger application and reschedule it far away from its related services. This could lead to high latency. Open Telemetry and distributed tracing can help track down such inefficient calls.

As technology advanced, there have been many attempts at formulating optimisation algorithms that all have their merits however many remain theoretical due to the strict restraints and requirements inherent to virtual workloads. Carbon-aware workloads are the future and workloads are becoming more flexible and intelligent through orchestration systems like Kubernetes; however, every workload is different, no two data centers are the same and we lack rich data and transparency from cloud providers about the carbon impact of our workloads in the cloud.

Temporal scheduling (when used in the context of carbon emissions) is a term used to refer to the act of scheduling your workloads at a time when the energy being used by that workload is likely to come from low-carbon sources such as Biopower, Solar Energy, Geothermal Energy, Hydropower, Wind Energy and Nuclear energy.

Workloads can be categorized in 3 ways: duration (short-running or long-running), execution time (ad hoc or scheduled) and interruptibility.

Spatial scheduling refers to running your workloads where it is most carbon efficient and the carbon intensity for power is lower.

Some data centers are directly attached to renewables and some are just "less dirty ", i. e. running on less brown energy than the data center the workload is originating from..

Moving workloads around is often not possible due to certain restrictions (e.g., GDPR or low latency workloads).

AT A GLANCE

CONCLUSION

The hyperscale cloud offers many advantages. Additionally, many adjustments can be made and often it is highly dependent on the individual requirements of the respective workload. Nevertheless, there are some basics that every person working in IT should keep in mind when working on projects to make public cloud usage greener.

When it comes to a green cloud architecture, best practices focus on design choices that minimise unnecessary resource utilisation and maximise the sharing of common tasks. Serverless is not always the most efficient solution; depending on the individual workload a classic Compute VM or even a Managed Service might be a good choice. One key aspect of a green cloud architecture is how the workload itself has been programmed. Another one is where the workload is running geographically. Depending on the region, the carbon intensity varies due to the energy mix of the region and technical aspects, such as how efficient and modern the hardware or cooling or similar data center components are running.



TOWARDS THE GREEN CLOUD WITH REPLY

WE ARE READY TO SUPPORT YOU TO RUN YOUR IT SYSTEMS IN A MORE SUSTAINABLE WAY, REDUCE YOUR CARBON FOOTPRINT AS WELL AS COSTS AND CONTRIBUTE TO A GREENER PLANET.

Reply can support you in running your entire IT landscape in a more environmentally friendly way. From migrating to the cloud, optimizing existing cloud systems, and choosing the best architecture for new software projects, down to coding in a greener way – we've got you covered in every step of your IT sustainability transformation journey. Our experts have long-standing experience with migrating to the cloud, which can significantly reduce the carbon footprint associated with on-premises IT infrastructure.

For systems already running in the cloud, we offer to implement GreenOps practices to optimize resource usage and minimize waste. This can involve deploying automation tools to optimize infrastructure utilization, as well as monitoring and alerting systems to identify and remediate issues that may result in unnecessary energy consumption and costs.

Choosing the right cloud architecture is also critical in reducing environmental impact. Our experts can help you select cloud architectures that are optimized for efficiency, such as microservices-based architectures that allow for more granular scaling and utilization. They can also help to implement serverless architectures that reduce the need for physical servers and allow for more efficient resource utilization.

Finally, we promote green coding best practices that help to build software in a more resource and energy-efficient way. Examples of these techniques include using efficient algorithms and data structures, optimizing code for efficient resource usage, and minimizing unnecessary computation and data transfer.

Sources: [1] The global cloud storage market is projected to grow from $83.41 billion in 2022 to $376.37 billion by 2029, at a CAGR of 24.0% in the forecast period, 2022-2029: www.fortunebusinessinsights.com/cloud-storage-market-102773; [2] Global data centre energy demand by data centre type, 2015-2021: www.iea.org/data-and-statistics/charts/global-data-centre-energy-demand-by-data-centre-type-2015-2021; [3] Categorisation of workloads and concrete values in this chapter are based on: Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. Wiesner et al., 2021; [4] https://googlecloudplatform.github.io/region-picker