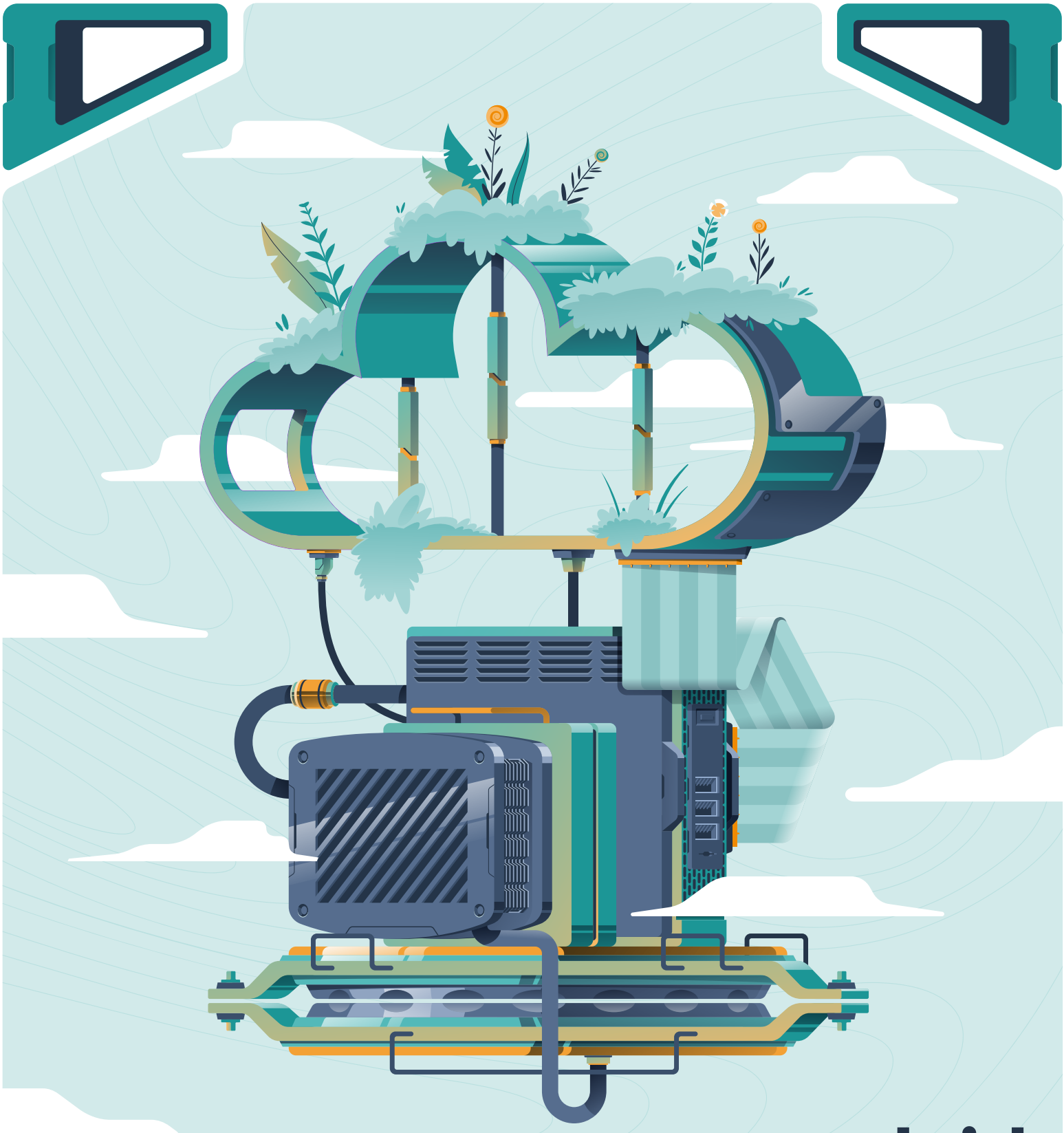
CAN PLATFORM ENGINEERING ADDRESS DATA CENTER ENERGY SHORTAGES?



cycloid

TABLE OF CONTENTS

INTRODUCTION

Data centres are the backbone of the digital world. They power everything, from streaming websites to the now ubiquitous AI services. Their potency, however, comes at a price: their power demand can make both their environmental impact and operational costs skyrocket. Soon, data centres might become a massive burden on the global ecology, and, in some cases, something of a financial drain.

Getting to grips with data centres' voracious energy consumption requires a good look into the latest research. Thankfully, there is plenty - at Cycloid, we've got the stats on electricity and water waste, future power demand projections, and potential solutions at our fingertips.
In this ebook, we put all of these to use. We go over the data on the current state of affairs, break down the pros and cons of different approaches to data centre sustainability, and take a look at the prognoses for the near future. Above all, we explore how that platform engineering can help keep resource consumption down, all while improving performance and costs-efficiency.

CHAPTER 1

Understanding Data Centre Energy Consumption

The global data centre landscape

According to an estimate by the International Energy Agency (IEA)[1], in 2022, data centres world-wide consumed between 240 and 340 TWh. That would represent approximately 1-1.3% of global electricity demand that year.

If that does not sound like much, add the 110 TWh or so consumed through cryptocurrency mining (another 0.4%) - and just like that, we are at almost 2% of global demand in 2022[1]. By 2030, the number is projected to climb to about 4%[2].Concerning? We think so.

Surprisingly slow growth in energy consumption: what's the deal?

Despite strong demand for data centre services, their energy consumption (excluding crypto) has not soared since 2010. A few innovations have helped keep the growth within limits: for instance, hardware and thermal management has grown more efficient.

Another factor is that an ever-increasing number of companies opt to rely on a cloud provider such as Amazon, instead of purchasing infrastructure and servers for their own use. As a result, on-premise data centres are largely being replaced by cloud and hyperscale facilities. However, it is not all sunshine and rainbows. This data does not take into consideration the rebound effect on public clouds, where the increased efficiency of adopting cloud services boosts the overall demand. So much so, in fact, that it is offsetting the original benefits.

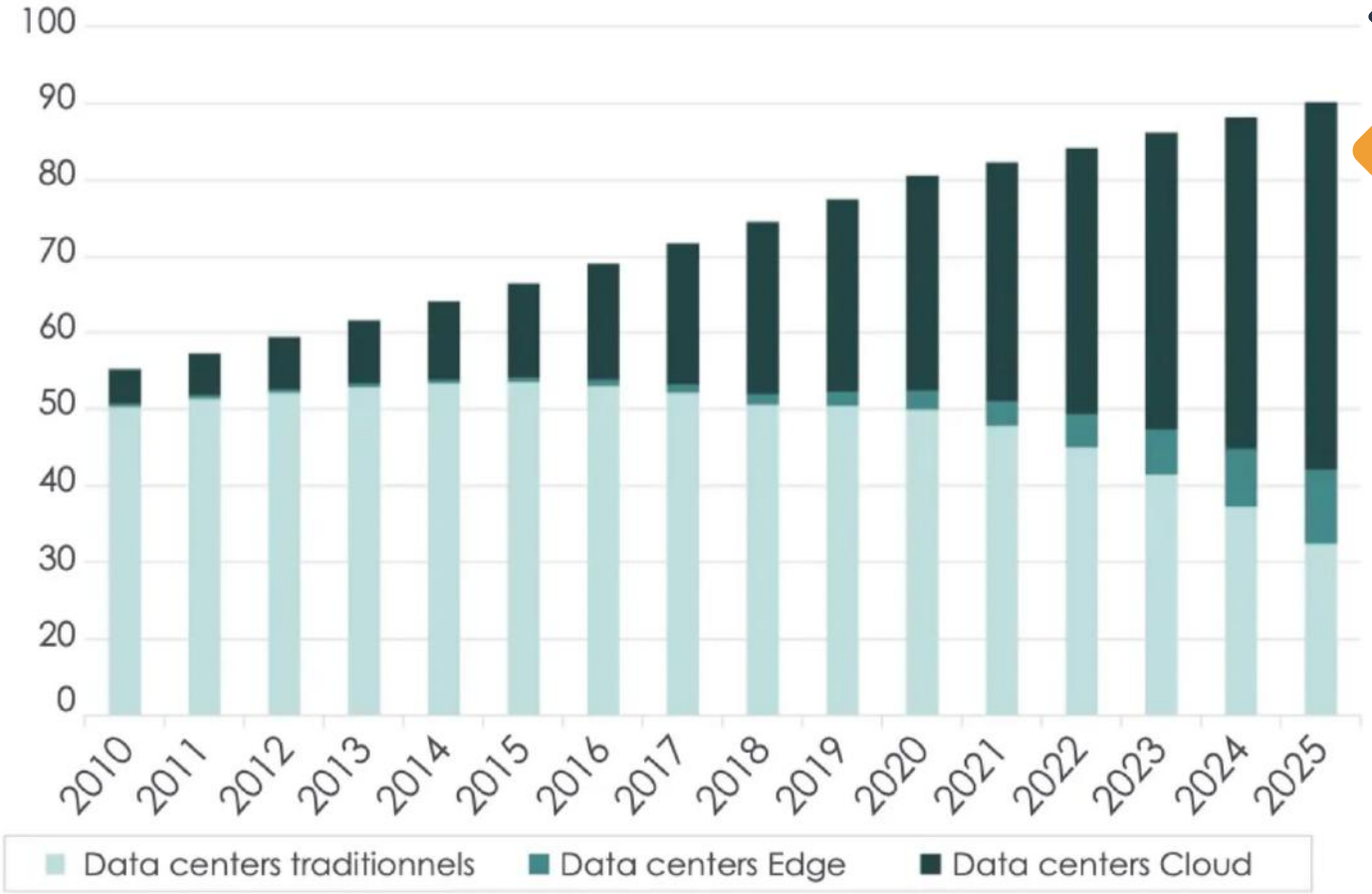
On top of that, we are seeing AI spread to just about every corner of today's digital ecosystem, including on-premise data centres. These have already resulted in a spike in energy consumption within the segment, with annual growth of 20 to 40%. The combined power consumption of Amazon, Microsoft, Google and Meta more than doubled between 2017 and 2021, reaching approximately 72 TWh in 2021[1]. So why is that happening - and what are our options?

The overlooked environmental toll of AI

The extensive use of AI, particularly through language models such as GPT-3 and its later versions, is taking a toll on energy and water usage in data centres. Facilities' water waste in particular has often been swept under the rug.

Indeed, water may not be the first thing you think of when the subject of data centre sustainability comes up. But think of it this way: large volumes of water are being used to generate electricity that is then supplied to data centres. This is also referred to as indirect water consumption, or IWC. Next comes DWC, direct water consumption, which mostly results from the need to neutralise the vast amounts of heat emitted by the computing centres. This is mostly done using on-site water cooling systems. To top it off, data centres keep cropping up in areas of high water stress, their staggering water consumption unheeded: an average data centre uses up the water equivalent of a city with a population of 30,000 to 40,000.[3]

More data is available. One case study[4], for instance, looks at the environmental footprint of operational water consumption of GPT-3, a 175 billion parameter language model used by popular online services such as ChatGPT. It turns out that the training of a model like that, carried out in data centres such as those of Microsoft, used up a whopping 1287 MWh[4]. While water consumption for AI training varies from centre to centre, it is significant enough to be considered when assessing the environmental impact of AI.



Source : Borderstep Institute

Table 1: Estimate of GPT-3's average operational **water consumption footprint**. "*" denotes data centers under construction as of July 2023, and the PUE and WUE values for these data centers are based on Microsoft's projection.

Location	PUE	WUE (L/kWh)	Electricity Water Intensity (L/kWh)	Water for Training (million L)			Water for Each Inference (mL)			# of Inferences for 500ml Water
				On-site Water	Off-site Water	Total Water	On-site Water	Off-site Water	Total Water	
U.S. Average	1.170	0.550	3.142	0.708	4.731	**5.439**	2.200	14.704	**16.904**	**29.6**
Wyoming	1.125	0.230	2.574	0.296	3.727	**4.023**	0.920	11.583	**12.503**	**40.0**
Iowa	1.160	0.190	3.104	0.245	4.634	**4.879**	0.760	14.403	**15.163**	**33.0**
Arizona	1.223	2.240	4.959	2.883	7.805	**10.688**	8.960	24.259	**33.219**	**15.1**
Washington	1.156	1.090	9.501	1.403	14.136	**15.539**	4.360	43.934	**48.294**	**10.4**
Virginia	1.144	0.170	2.385	0.219	3.511	**3.730**	0.680	10.913	**11.593**	**43.1**
Texas	1.307	1.820	1.287	2.342	2.165	**4.507**	7.280	6.729	**14.009**	**35.7**
Singapore	1.358	2.060	1.199	2.651	2.096	**4.747**	8.240	6.513	**14.753**	**33.9**
Ireland	1.197	0.030	1.476	0.039	2.274	**2.313**	0.120	7.069	**7.189**	**69.6**
Netherlands	1.158	0.080	3.445	0.103	5.134	**5.237**	0.320	15.956	**16.276**	**30.7**
Sweden	1.172	0.160	6.019	0.206	9.079	**9.284**	0.640	28.216	**28.856**	**17.3**
Mexico*	1.120	0.056	5.300	0.072	7.639	**7.711**	0.224	23.742	**23.966**	**20.9**
Georgia*	1.120	0.060	2.309	0.077	3.328	**3.406**	0.240	10.345	**10.585**	**47.2**
Taiwan*	1.200	1.000	2.177	1.287	3.362	**4.649**	4.000	10.448	**14.448**	**34.6**
Australia*	1.120	0.012	4.259	0.015	6.138	**6.154**	0.048	19.078	**19.126**	**26.1**
India*	1.430	0.000	3.445	0.000	6.340	**6.340**	0.000	19.704	**19.704**	**25.4**
Indonesia*	1.320	1.900	2.271	2.445	3.858	**6.304**	7.600	11.992	**19.592**	**25.5**
Denmark*	1.160	0.010	3.180	0.013	4.747	**4.760**	0.040	14.754	**14.794**	**33.8**
Finland*	1.120	0.010	4.542	0.013	6.548	**6.561**	0.040	20.350	**20.390**	**24.5**

Estimated average operational water consumption footprint of GPT-3[4]

Of course, training is not the sole energy-intensive component of AI usage. Inference, which is the act of generating content from these models, also has a significant environmental footprint, with electricity consumption of around 0.004 kWh per request. A 2023 study, aptly titled "Gone with the clouds: Estimating the electricity and water footprint of digital data services in Europe," offers an alarming assessment. According to the research, by 2030, the water consumption linked to internet usage in Europe will clock in at 273.4 - 820.1 million m^3 per year. Electricity consumption, at the same time, will range from 56.3 to 169 TWh.

Nor are the authors of "Gone with the clouds" alone in their concern. Researchers at the University of Washington[5] have looked into the power consumption of intensive AI usage, and their findings reveal that the **hundreds of millions of entries received by ChatGPT daily can consume approximately 1 GWh. The same daily amount of energy is needed to power 33,000 American homes[5]**. We might think ChatGPT is only spilling the tea when we use it for a quick personal query, but what actually gets spilt is water. And lots of it.

Apart from transforming the vague idea of "needing to go green" into concise, handy figures, these findings offer a reminder: taking steps to mitigate the environmental consequences of the technology is a pressing need indeed.

After all, generative AI models have seen significant democratisation: AI has already been integrated into a variety of software, including Office, Microsoft's Copilot, as well as Google's Gmail and Docs and Github's Copilot, and more is to come. This trend could exponentially increase the Internet's environmental footprint in a very short time.

According to Gartner[6], by 2025, without sustainable artificial intelligence practices, AI will consume more energy than human labour. This will render carbon neutrality all but unattainable. As it becomes more pervasive and requires increasingly complex machine learning models, AI consumes more data, more resources, and, consequently, more power. If current practices remain unchanged, the energy required for AI training, associated data storage, and processing could represent up to 3.5% of global electricity consumption by 2030 - almost double the 2021 estimate.

Fortunately, current practices need not stay unchanged. In fact, in the long run, challenging them benefits every party involved. But how can we make sure that change bears fruit? We have gone through some of the latest research on AI sustainability to find out.

At least two recent studies [4] [7] point out the crucial role that the choice of location and timing of AI training plays in regulating its environmental footprint. By exploiting the spatial and temporal diversity of water and electricity consumption, one could strategically schedule and plan the processes involved in training and running AI models, effectively reducing their environmental impact. In other words, adjusting when and where AI processes happen based on how efficiently resources are used can lead to significant savings in water and energy. Better for the planet - yes, and much less costly.

One drawback is that there is often no transparency on real-time effectiveness of different setups. Thus, the 2023 report recommends increasing visibility of real-time water and electricity efficiency, as well as increased transparency from AI model developers and data centre operators.
The same study highlights another potential issue: between carbon, energy, and water usage, minimising one may increase the others. The solution, the research suggests, lies in striking a balance between performing power-intensive tasks during the hours when more solar energy is available, or the cooler hours of the day for adequate water usage.

To ensure that this effort is successful, the findings of both studies must be carefully considered. It is true that data centres' hunger for resources other than energy is sometimes left out of the spotlight. Today, however, the need to address it is clear, and a strategic approach to the timing and location of key AI processes can help optimise resource usage. While this requires some research, the benefits still outweigh the drawbacks by a good margin.

It goes without saying that the positive outcome need not come out of careful planning alone: it must also be driven by innovative thinking, including the contributions of platform engineering. So, in the next chapter, we have a good look at the solutions that the field can offer and just what benefits they yield.

CHAPTER 2

7 ways Platform Engineering can help boosting Resource Efficiency

Platform engineering can greatly aid in addressing energy shortage issues in data centres. There are a few tools in its toolbox: prioritising efficiency, optimising resource utilisation, promoting sustainable practices, and more. Self-service portals and automation are among the most promising: as users are empowered to manage resources efficiently, the need for manual intervention is reduced considerably. Below, we break down the key benefits of the two - and we don't shy away from explanations, so follow along for the details.

1 Fighting Resource Overconsumption, One Task at a Time

User at the Wheel: Self-service portals allow users to provision, de-provision, and scale resources like virtual machines or on-demand storage. This reduces the likelihood of over-provisioning resources, which can lead to unnecessary power consumption. Users can choose to request only the resources they need for a specific task and easily decommission them once the task is complete.

- **Automated Right-Sizing:** Automation tools integrated with self-service portals, such as autoscaling, can monitor resource usage and automatically adjust allocations to ensure resources are not over- or under-utilised. This dynamic adjustment helps minimise energy waste by ensuring that only the necessary amount of computational power is used.

2 Seamless Workload Management

- **Intelligent Scheduling:** Automation can be used to schedule resource-intensive tasks during off-peak times when energy costs and demand are lower, thereby reducing the total cost of computing infrastructure.

- **Energy-Savvy Orchestration:** Self-service portals coupled with automation can incorporate power-aware scheduling algorithms that prioritise running workloads on the most energy-efficient servers or data centre locations, reducing overall power use.

3 Cutting Waste with Elasticity - Sounds Like a Stretch? Take heart, it isn't!

- **On-Demand Scaling:** Automation enables auto-scaling, where resources automatically scale up or down based on real-time demand. This means that during periods of low demand, unnecessary servers can be powered down, and during peak times, resources can be scaled up efficiently. This elasticity ensures that energy consumption is aligned with actual resource needs, avoiding waste.

- **No More Idle Resources:** Automated systems can detect and shut down idle or under-utilised resources, which might otherwise continue to consume power unnecessarily. This reduces the baseline power usage of the data centre. In other words, no need to keep your eyes peeled - the platform does it for you.

4 Streamlined DevOps and Continuous Integration/Continuous Deployment (CI/CD)

- **Reduced Human Error:** By automating repetitive tasks, the potential for human error—such as leaving resources running inadvertently—is minimised. This greatly reduces unnecessary energy consumption caused by oversight.
- **Boost for Dev Efficiency:** Automated build and test environments simplify developers' workflows, speeding up feedback, reducing errors, and making the development cycle more efficient and productive.

5 Data at Your Fingertips

- **Real-Time Monitoring:** Self-service portals often include dashboards that provide real-time visibility into resource usage and energy consumption. Automation can trigger alerts or actions based on this data: think shutting down under-utilised resources or redistributing workloads to more power-efficient environments.
- **Boost for Dev Efficiency:** Automated build and test environments simplify developers' workflows, speeding up feedback, reducing errors, and making the development cycle more efficient and productive.

6 Power to the User

- **Informed Decision-Making:** Self-service portals provide users with the information and tools needed to make energy-efficient decisions. For example, users can see the energy impact of their resource requests and choose more efficient options when available.
- **Great Power, Great Responsibility:** By integrating energy consumption data with financial metrics (FinOps), users are made aware of the costs of their resource usage. This serves as motivation to optimise their resource requests and reduce unnecessary power consumption.

7 Getting to Grips with Shadow IT (and Its Power Consumption)

Shadow IT refers to the use of unauthorised technology resources by employees - you guessed it, often without the knowledge or approval of the IT department. Now, shadow IT can sometimes drive innovation and problem-solving. But it may also engender uncontrolled and redundant resource use, which leads to inefficiencies and increases energy consumption.

Over-provisioned or idle resources, redundant systems, and inefficient configurations are all potential outcomes of shadow IT, and each contributes to unnecessary power use. More often than not, these resources lack optimization otherwise provided by central management, which exacerbates the energy shortage problem.

Control Strategies:

- **Enhanced Visibility:** Self-service portals can reduce the reliance on shadow IT by providing a sanctioned and efficient way to access necessary resources. By offering transparency and control, these portals greatly simplify management and monitoring tasks for IT departments
- **Incorporating Automation:** Automated monitoring and alerting can help identify unauthorised resources, enabling the IT department to address shadow IT issues proactively
- **Education and Awareness:** Informing users about the risks of shadow IT and promoting the use of approved, energy-efficient tools can further reduce the prevalence of shadow IT

The impact of shadow IT on power consumption is by no means negligible - but neither is it insoluble. Solutions like self-service portals and automation offer enhanced visibility, control, and user education: all of these have the power to reduce, if not eliminate, the issue of shadow IT overconsumption. And the cherry on top is that in this scenario, meeting efficiency goals also means supporting broader sustainability in the field.

CHAPTER 3

Understanding Sustainable Platform Engineering

Sustainable platform engineering is a guiding principle that integrates environmental responsibility directly into the design, development, and operation of platforms. Its principal aim is to minimise the platforms' carbon footprint, but it also takes performance and efficiency seriously - no compromises necessary.

A key component of sustainable platform engineering is the integration of **GreenOps** and **FinOps**—two critical operational practices—into the orchestration layer of platform engineering. Do you want to make sure that sustainability and cost-effectiveness are not mere afterthoughts but core components of system management and optimisation? Embedding GreenOps and FinOps can lift this weight off your shoulders.

GreenOps: Operationalizing Sustainability

The key task of **GreenOps** is implementing sustainability within IT operations. It does so through the systematic management of energy consumption, resource utilisation, and environmental impact across the entire lifecycle of platforms and applications.

In other words, GreenOps is there to help you minimise the environmental footprint of digital infrastructure. Let us go over some of the practices that make that possible:

- **Energy-Efficient Resource Management:** Automatically scaling resources up or down based on real-time demand, thereby ensuring that only the necessary resources are in use. The result? Noticeable reductions in energy waste.
- **Carbon Footprint Monitoring:** Implementing tools and processes to track and reduce the carbon emissions associated with data centre operations, including the sourcing of renewable power.

All in all, using a GreenOps application shows a strong commitment to reducing an organisation's environmental footprint. The app will make recommendations based on regional specificities and consider the optimal time to execute the workload before integrating these considerations into its suggestions. As a result, GreenOps will favour data centres with the smallest environmental footprint at a given time, promoting a more responsible use of IT resources. Sounds familiar? Rightly so: GreenOps' approach is in line with the recommendations of the 2023 "Gone with the Clouds" study, showcasing it as a powerful tool to reduce AI's environmental footprint.

By incorporating GreenOps at the orchestration layer, platform engineers can automate and enforce sustainability practices across all layers of the stack, from infrastructure to applications. That way, rather than a mere operational concern, energy efficiency becomes a built-in feature of the platform's architecture.

FinOps: Financial Optimization with Sustainability in Mind

FinOps focuses on the financial management of cloud resources, ensuring that organisations maximise their return on investment while controlling costs. In the context of sustainable platform engineering, FinOps plays a crucial role in balancing cost efficiency with sustainability goals. Its key practices include:

- **Cost-Aware Resource Allocation:** Using AI and automation to optimise the allocation of cloud resources, reducing wasteful spending while also minimising energy usage.

- **Sustainable Cost Reporting:** Providing detailed reports that not only track financial costs but also correlate these with power consumption and carbon emissions, enabling more informed decision-making.

At the orchestration layer, FinOps can be integrated to provide real-time insights and automated adjustments that align financial management with sustainability objectives. This ensures that cost-saving measures do not inadvertently lead to increased energy consumption or environmental impact.

Orchestration Layer: The Integration Point

Within sustainable platform engineering, the orchestration layer is the ideal integration point for GreenOps and FinOps. This layer, which manages the deployment, scaling, and operation of applications and services, is where decisions about resource allocation, load balancing, and system efficiency are made.

Embedding GreenOps and FinOps at this specific layer enables you to do a few things:

- **Automate Sustainable Practices:** Automate the enforcement of energy-efficient and cost-effective operations, ensuring that sustainability is maintained even as systems scale.
- **Enable Real-Time Adjustments:** Implement real-time monitoring and adjustments to balance performance, cost, and environmental impact dynamically.
- **Foster Cross-Disciplinary Collaboration:** Encourage collaboration between sustainability teams, financial operations, and platform engineers, ensuring that all aspects of the platform's operation contribute to overall sustainability goals.

You can think of FinOps & GreenOps as a “force de proposition,” making informed recommendations to optimise processes, and of orchestration as the executive force. In other words, FinOps & GreenOps recommend action - and orchestration makes it happen.

Conclusion

Building sustainable platforms by combining GreenOps and FinOps at the orchestration level is a reliable, leading-edge way to manage today's IT infrastructures. Placing sustainability and financial responsibility right at the core of platform operations kills two birds with one stone: on the one hand, it reduces the environmental impact, and on the other, sets the foundation for a sustainable, cost-effective future.

CHAPTER 4

The Future of Data Centres: Uncertainty as an Asset

Neutralising AI's impact

We've discussed the drawbacks of adopting AI, but with proper use, it also holds incredible potential to enhance the positive impact of data centre sustainability practices. In sustainable platform engineering, AI and ML are used to optimise resource allocation, improve cooling efficiency, and enhance power usage effectiveness (PUE) in data centres. They could enable dynamic management of computing resources, predictive analytics, and smart cooling systems, all of which help reduce power consumption and improve the overall efficiency of data centre operations. They can also support the integration of renewable energy sources, which ensures a lower carbon footprint.

Still, AI and ML do require a lot of energy, particularly during the training of large models and the continuous operation of complex algorithms. The computational power required for these processes can be substantial, potentially offsetting the energy savings achieved through AI-driven optimizations. Operational costs, including data processing and real-time inference, can add insult to injury.

Energy-intensive though they are, AI and ML are excellent tools to ensure efficient resource management within data centres. However, it is essential that associated energy costs do not outweigh the benefits. One way to achieve that is to implement the technology in areas where it offers the most significant power savings and in line with sustainable practices like transfer learning and optimised hardware usage. Another possibility is to drop the large generic ML models in favour of task-specific ones, which, according to a 2024 study[8], are considerably more energy-efficient.

What's next?

With AI expected to amplify the global data centre power demand by 160% by 2030[9], we can hardly afford to trifle with responsible AI deployment practices. Some of them are almost intuitive: for instance, implementing AI to regulate the resource consumption of power-intensive processes is obviously more fruitful than adopting it indiscriminately. Others are subject to the development of ML models with a narrowed focus, which would allow companies to opt for the power-saving task-specific models. While these solutions can turn AI and ML into powerful energy-saving tools, research on how to prevent them from becoming energy drains is growing—and we'll be keeping a close watch on it for you.

Heedful AI implementation, development of task-specific ML models, deepening our understanding of the nuts and bolts of AI energy consumption… The common denominator? Humans. We are the driving force for positive change. AI is powerful, but it does not work miracles.

It is up to us to minimise waste, both financial and environmental. Humans are in charge of shutting down obsolete projects that run forgotten in data centres, gulping down valuable resources. Research, too, is human-driven, and learning to monitor facilities' resource consumption will supply precious insight. With up-to-date figures at our fingertips and someone like you at the wheel, platform engineering can drive positive change.

The future of data centres is in our human hands. Unsettling it might be, but also empowering - and hopeful.

Conclusion

In recent years, the environmental cost of data centres has risen - and the numbers are projected to surge in the near future. In 2023, their carbon footprint clocked in at 43 million tons of CO2 per year, increasing at an annual rate of 11%[10]. Their carbon dioxide emissions are projected to more than double between 2022 and 2030, and their power consumption, to make up 3-4% of the global demand[11]. Not very hopeful, we agree - but every projection is subject to change, and some of it is in our hands.

One approach that can help address data centres' voracious use of resources is leveraging the spatial and temporal diversity of water and electricity consumption. By scheduling energy-intensive tasks strategically, depending on location and solar energy available, both carbon and water waste can be minimised. However, to become bulletproof, this method needs a better insight into runtime water efficiency and a lifecycle view of AI's water footprint, which could come from more data from AI model developers and data centre operators.

Among your most potent allies in combating data centres' excessive power consumption is sustainable platform engineering. Embedding environmental responsibility into the design, development, and operation of platforms, it aims to reduce software development carbon footprint while boosting performance and team efficiency. Some of its key tools include automation and self-service portals, which can optimise workloads and resource usage, and deliver real-time insights into energy consumption.

Sustainable platform engineering also involves the integration of GreenOps and FinOps into the orchestration layer. Doing so allows you to automate energy and cost-efficient operations to maintain sustainability as systems scale, and to enable real-time adjustments to balance cost, performance, and environmental impact. Integrating GreenOps and FinOps at this specific layer means that their suggestions are implemented, paving the way for cost-effectiveness and lowering the environmental footprint.

Today, the future of data centres' energy efficiency is somewhat opaque. Prognoses are often pessimistic, but for this very reason, sustainability innovations thrive, and the methods in use get updated continuously. Google, for instance, announced having decreased their data centre's cooling bill by up to 40% using their DeepMind AI lab[12]. Studies like the 2024 "Innovating Sustainability" propose well-researched, promising ways to put AI at the service of sustainability efforts[13]. A variety of handy, easy-to-deploy methods are now accessible - and with mindful use, they might just turn things around.

FINAL THOUGHTS

Far be it from us to suggest that you solely rely on platform engineering to curtail data centres' ravenous power consumption.There is no need to put all eggs in one basket: combining it with other solutions generated by green IT, from managing CPU's efficiency to "green coding," will probably work best. However, sustainable platform engineering is a key area to consider when putting in place IT sustainability strategies, and for good reason. Many of its practices, such as optimising resource usage or automating scaling, are relatively easy to deploy and yield considerable results.

In many ways, platform engineering is as much about people as it is about software. At its core is the idea of using technology mindfully, of balancing efficiency and sustainability goals, rather than relying on technological solutions alone. This is what drives us at Cycloid, empowering us to deliver solutions that prioritise both sustainability and performance.

Looking to exceed your sustainability goals? Hop on & enjoy the ride!

REFERENCES

[1] Data Centers and Data Transmission Networks , IEA, July 2023

[2] *"AI is poised to drive 160% increase in data center power demand"*, Goldman Sachs, 14 May 2024.
https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

[3] Sattiraju, Nikita. *"The Secret Cost of Google's Data Centers: Billions of Gallons of Water to Cool Servers"*, Time Magazine, 2 April 2020.
https://time.com/5814276/google-data-centers-water Accessed 22 September 2024

[4] *"Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models"*, University of California, Riverside, October 2023

[5] Q&A: UW researcher discusses just how much energy ChatGPT uses , University of Washington, July 2023
https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/

[6] Gartner Unveils Top Predictions for IT Organizations and Users in 2023 and Beyond , Gartner, October 2022
https://www.gartner.com/en/newsroom/press-releases/2022-10-18-gartner-unveils-top-predictions-for-it-organizations-and-users-in-2023-and-beyond

[7] Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning , Stanford, McGill, November 2020
https://jmlr.org/papers/volume21/20-312/20-312.pdf

[8] Luccioni, Sasha, Yacine Jernite, Emma Strubell. *"Power Hungry Processing: Watts Driving the Cost of AI Deployment?" FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency*, June 2024. Doi: 10.1145/3630106.3658542

[9] *""AI is poised to drive 160% increase in data center power demand"*, Goldman Sachs, 14 May 2024
https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

[10] Buyya, Rajkular, Shashikant Ilager, Patricia Arroba. *"Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads" Software: Practice and Experience*, 2024; 54(1): 24–38. doi: 10.1002/spe.3248.
https://onlinelibrary.wiley.com/doi/full/10.1002/spe.3248

[11] *""AI is poised to drive 160% increase in data center power demand"*, Goldman Sachs, 14 May 2024
https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

[12] Evans, Richard, Jim Gao. *"DeepMind AI Reduces Google Data Centre Cooling Bill by 40%" Google DeepMind*, 20 July 2016.
https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/

[13] Chen, Yanyu & Li, Qian & Liu, JunYi.*"Innovating Sustainability: VQA-based AI for Carbon Neutrality Challenge", Journal of Organizational and End User Computing. Vol. 36*, 2024, 1-22.10.4018/JOEUC.337606.

About Cycloid

Cycloid is the sustainable platform engineering company with a mission to promote efficient infrastructure and software delivery alongside digital sobriety. Cycloid optimizes platform engineering, alleviates cognitive load on IT teams, and enhances Green IT and FinOps practices. Placing sustainability at the orchestration layer, the Cycloid engineering platform is a comprehensive solution for platform engineering teams and end users, delivering optimal UX with modular self-service portal access to project lifecycle, resource management, FinOps and GreenOps capabilities. With a zero lock-in, GitOps first approach, Cycloid encourages a culture of digital sobriety that flows through an organization, making DevOps and cloud delivery more efficient and cost-effective.