C’est l’accueil café ☺

Eco-responsabilité avec le cloud Azure

Le vendredi 21 octobre de 10h à 12h

Nicolas Clerc, Narjes Majdoub, Stanislas Quastana

Agenda

- Introduction
- Datacenters Microsoft : ce que fait Microsoft
- GreenOps : ce que vous pouvez faire avec Azure
- Développement éco responsable
- Green AI
- Questions/Réponses

Ce qu'on fait de notre côté

Datacenters Microsoft

Impacts d'un Datacenter

- Consommation de ressources
  - Pour sa construction
  - Pour la fabrication du matériel informatique
- Consommation d'énergie
  - Pour opérer le bâtiment (éclairage, sécurité, refroidissement)
  - Pour faire fonctionner les équipements IT
- Consommation d'eau
  - Pour la climatisation, pour la construction
- De manière générale, la discussion finit souvent autour de l'empreinte carbone

Empreinte carbone

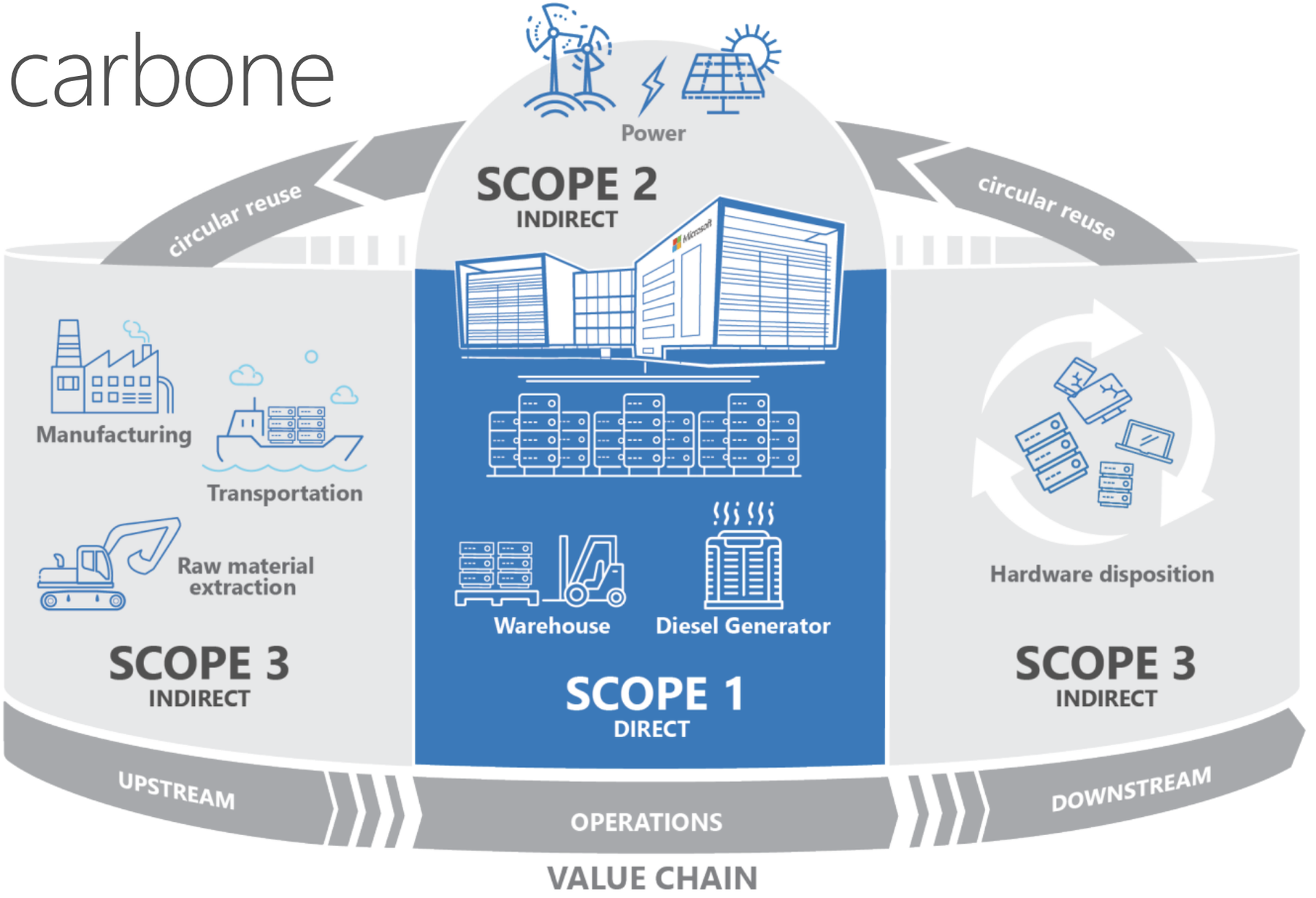


Figure 7: Examples of Scope 1, 2, and 3 carbon emission types in Microsoft cloud value chain

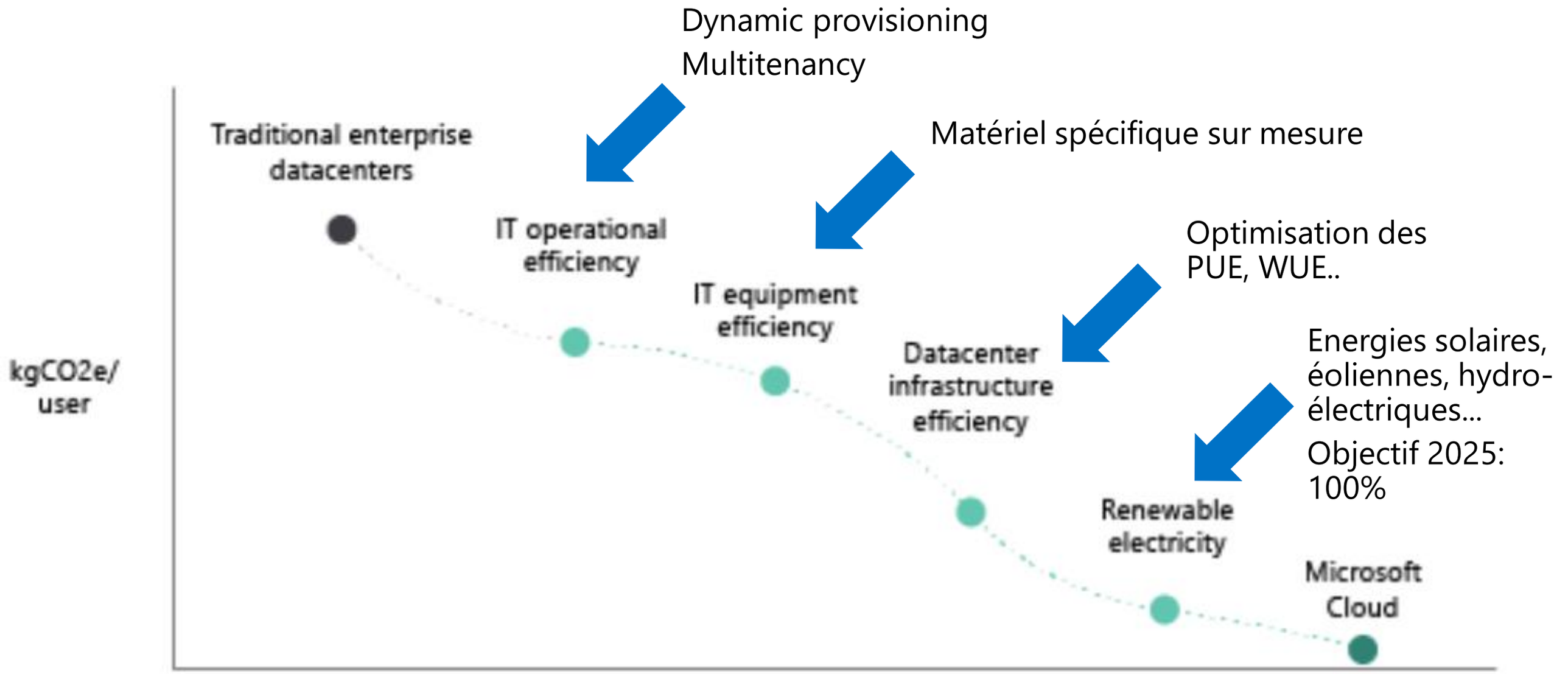
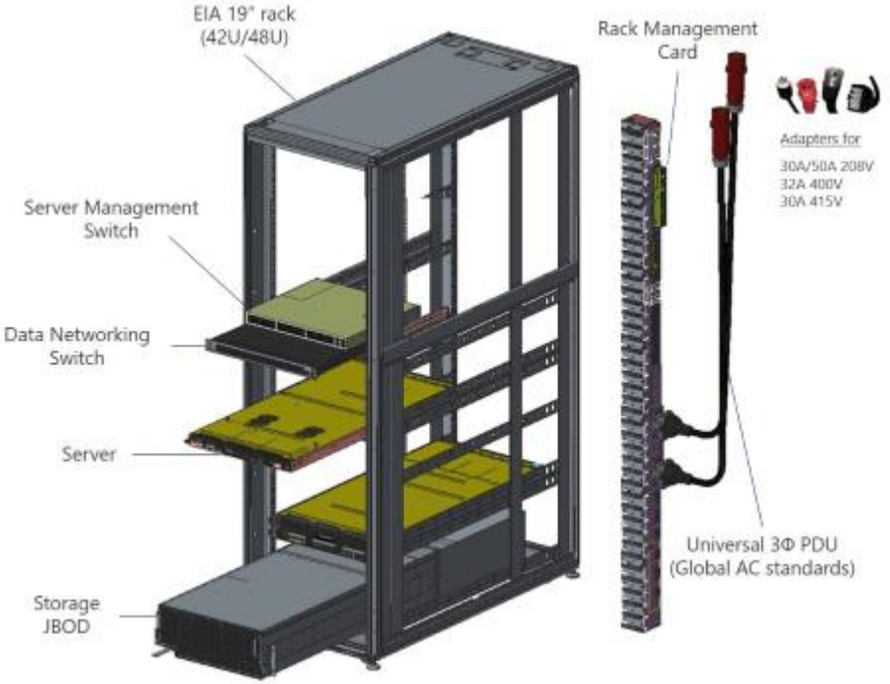


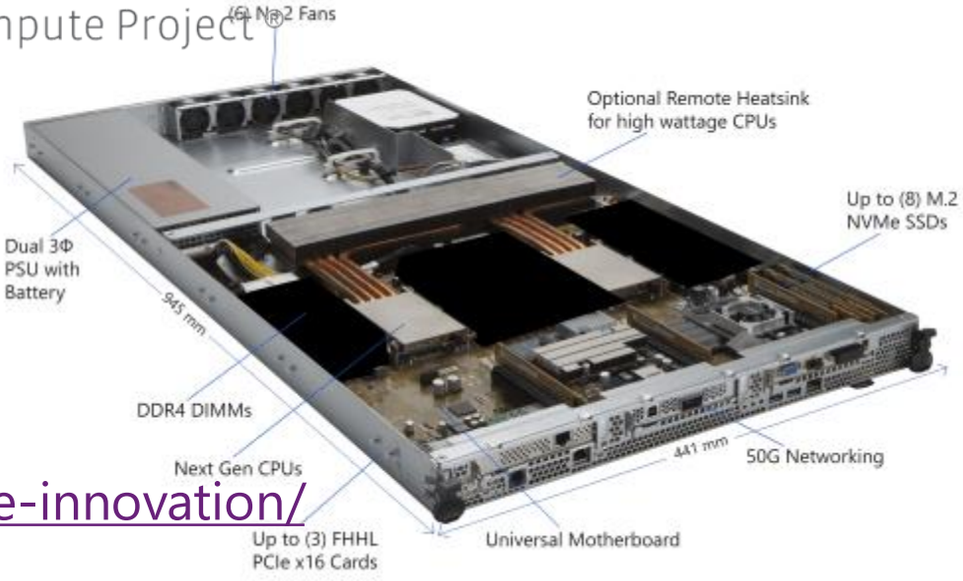
Figure 1*: The four features of the Microsoft Cloud that reduce environmental impact.

**kgCO2e = kilograms of carbon dioxide equivalent*

Matériel spécifique



OPEN Compute Project



https://azure.microsoft.com/en-us/global-infrastructure/hardware-innovation/
https://www.opencompute.org/wiki/Server/ProjectOlympus

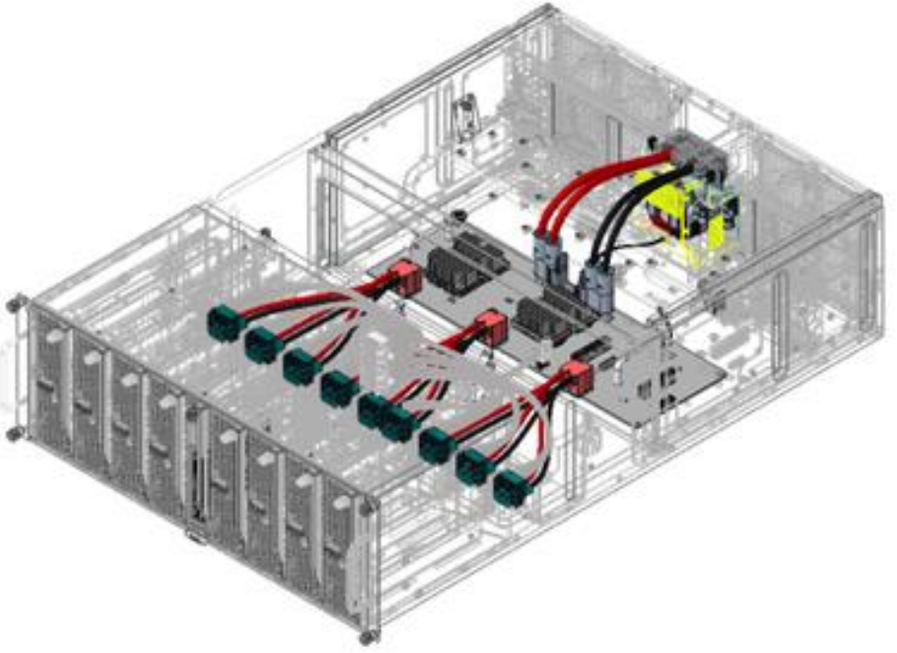
Modular Chassis — common chassis design for future of accelerators

By Gopa Parameswaran

Published Oct 18 2022 08:50 AM

A community-driven approach to infrastructure innovation is vital – not just for continued advancements in trust, efficiency, and scalability, but in service of a larger vision of empowering the ecosystem towards building for the computing needs of tomorrow.

We are contributing a new hardware design - modular chassis (Mt. Shasta) to Open Compute Project Foundation (OCP), whose mission is "to apply the benefits of open source and open collaboration to hardware and rapidly increase the pace of innovation in, near and around the data center." In the world of high-performance computing (HPC), artificial intelligence (AI) and video codecs, there is a need for a converged architecture that brings form factor, power and management interface into a modular design. Mt. Shasta is a modular chassis that enables multiple accelerators to fit into a single chassis supporting 48V power feed. Designed in partnership with Quanta and Molex, Mt. Shasta is designed to be fully compatible with Open Rack V3, with a modular form factor that provides support for high power devices, and hot swappable modules with flexibility in changing module-module connectivity. Mt Shasta also supports additional flexibility of a host connecting to the chassis.



Majority of the devices we deploy in data centers today are general purpose CPUs. To deploy these in the data center, we design 1RU or 2RU systems and mount them in one of our standard racks like Mt. Olympus. We have BMC based out of band management infrastructure and host CPU based in band management infrastructure for automation in our data center. Process and tools are well established to support this effort. Automation tools support deployment, health monitoring and diagnostics of hardware.

Sort by: Most Recent

Type

- [ ] Specification (36)
- [ ] White Paper (3)

Project

- [ ] Server (21)
- [ ] Rack & Power (7)
- [ ] Storage (5)
- [ ] Cooling Environments > Advanced Cooling Facilities (1)
- [ ] Data Center Facility (1)
- [ ] Hardware Management (1)
- [ ] Networking (1)

Show more

Contributor

- [x] Microsoft (39)
- [ ] Meta (4)
- [ ] Dell Inc. (3)
- [ ] Facebook (3)
- [ ] Google (3)
- [ ] Broadcom (2)
- [ ] Intel (2)

Show more

Family

INFO, Guidelines for Connection of Liquid Cooled ITE to Data Center Facility Systems

Guidelines for Connection of Liquid Cooled ITE to Data Center Facility Systems

Contributor: Victaulic, Vertiv, J.M. Gross Engineering, Microsoft, Critical Project Services (CPS)
Family: Data Center
Project: Cooling Environments > Advanced Cooling Facilities
Accept Year: 2022

SPEC, Scalable I/O Virtualization Rev1, Intel, Microsoft

Scalable I/O Virtualization (SIOV) Rev1 Specification

Contributor: Intel, Microsoft
Family: Server
Project: Server
Accept Year: 2022

INFO, White Paper, Life Cycle Assessment (LCA) Guidelines for Cloud Providers

Life Cycle Assessment (LCA) Guidelines for Cloud Providers

Contributor: Microsoft, WSP USA
Family: Information
Project: OCP Strategic Initiatives > Sustainability Initiative
Accept Year: 2022

Power Usage Effectiveness (PUE)

- Métrique industrielle qui mesure comment un datacenter consomme et utilise l'énergie électrique
- Ce calcul d'efficacité inclut le refroidissement, l'alimentation des serveurs, des équipements réseaux, de l'éclairage, des systèmes de sécurité…

$$\textbf{power usage effectiveness (PUE)} = \frac{total\ energy\ needed\ for\ facility}{total\ energy\ used\ for\ computing}$$

- Valeur idéale et théorique : 1

1989-2005	2007	2009	2012	2015
Generation 1	**Generation 2**	**Generation 3**	**Generation 4**	**Generation 5**
2.0+ PUE	1.4 – 1.6 PUE	1.2 – 1.5 PUE	1.12 – 1.20 PUE	1.07 – 1.19 PUE
				
Colocation	**Density**	**Containment**	**Modular**	**SW Defined**
Server Capacity 20 year Technology	Rack Density & Deployment Minimized Resource Impact	Containers, PODs Scalability & Sustainability Air & Water Economization Differentiated SLAs	ITPACs & Colocations Reduced Carbon Right-Sized Faster Time-to-Market Outside Air Cooled	Fully Integrated Resilient Software Common Infrastructure Operational Simplicity Flexible & Scalable

$$\text{PUE} = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$


Quincy, WA

GENERATION 3

Containment

1.4 – 1.6 PUE

GENERATION 6

GENERATION vNEXT

2017

Scalable form factor

- ✓ Reduced infrastructure
- ✓ Repeatable form factor
- ✓ Scale to demand
- ✓ Adiabatic cooling
- ✓ High-temperature DC
- ✓ 5x9 availability
- ✓ 1.17 – 1.19 PUE

Datacenter generations

GENERATION 1

Colocation

2.0+ PUE

GENERATION 2

Density

1.5 – 1.8 PUE

GENERATION 3

Containment

1.4 – 1.6 PUE

GENERATION 4

Modular

1.1 – 1.3 PUE

GENERATION 5

Hyper-scale

1.17 – 1.25 PUE

GENERATION 6

Scalable form factor

1.17 – 1.19 PUE

GENERATION 7

GENERATION 8

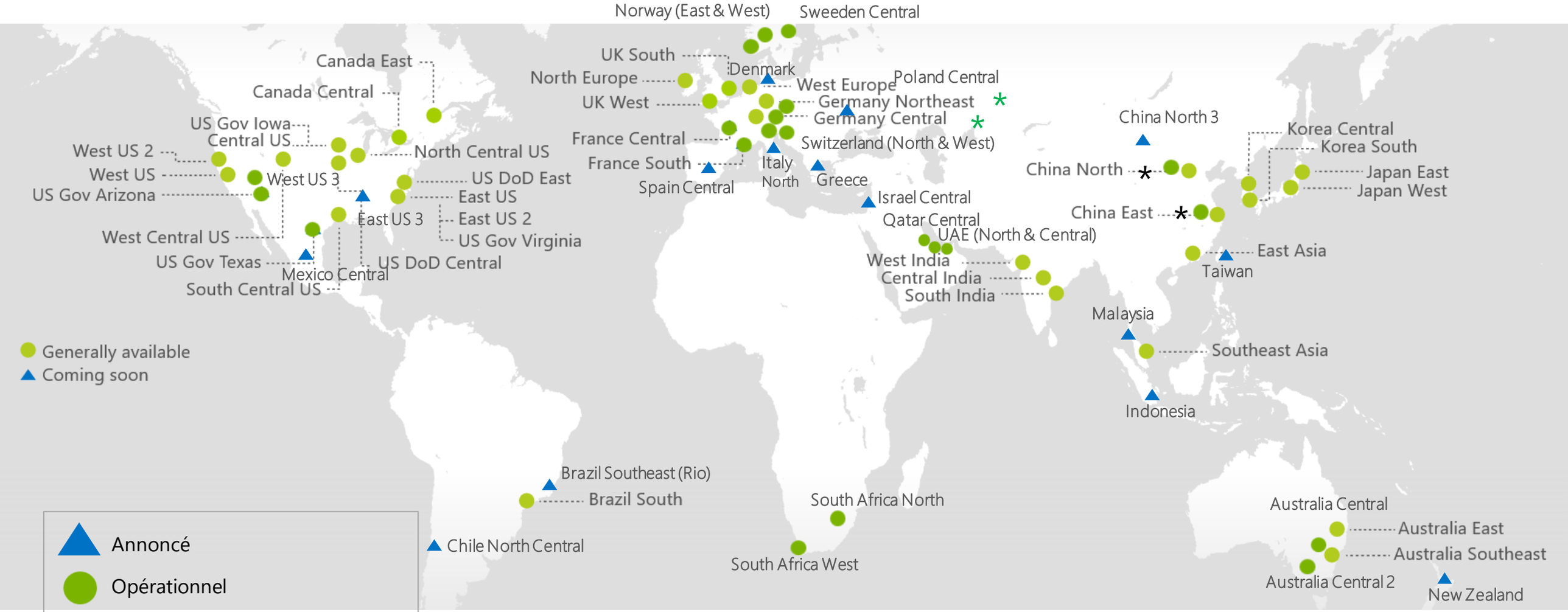
GENERATION vNEXT

2018

Ballard

- ✓ Design execution efficiency
- ✓ Simplification of critical environment
- ✓ Flex capacity enabled
- ✓ Highly competitive
- ✓ Adiabatic cooling
- ✓ High-temperature DC
- ✓ 5x9 availability
- ✓ 1.15 – 1.18 design PUE

Les régions Azure (> 66)



https://azure.microsoft.com/en-us/global-infrastructure/geographies/#overview

How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability

Posted on April 22, 2022

Noelle Walsh, Corporate Vice President, Cloud Operations + Innovation

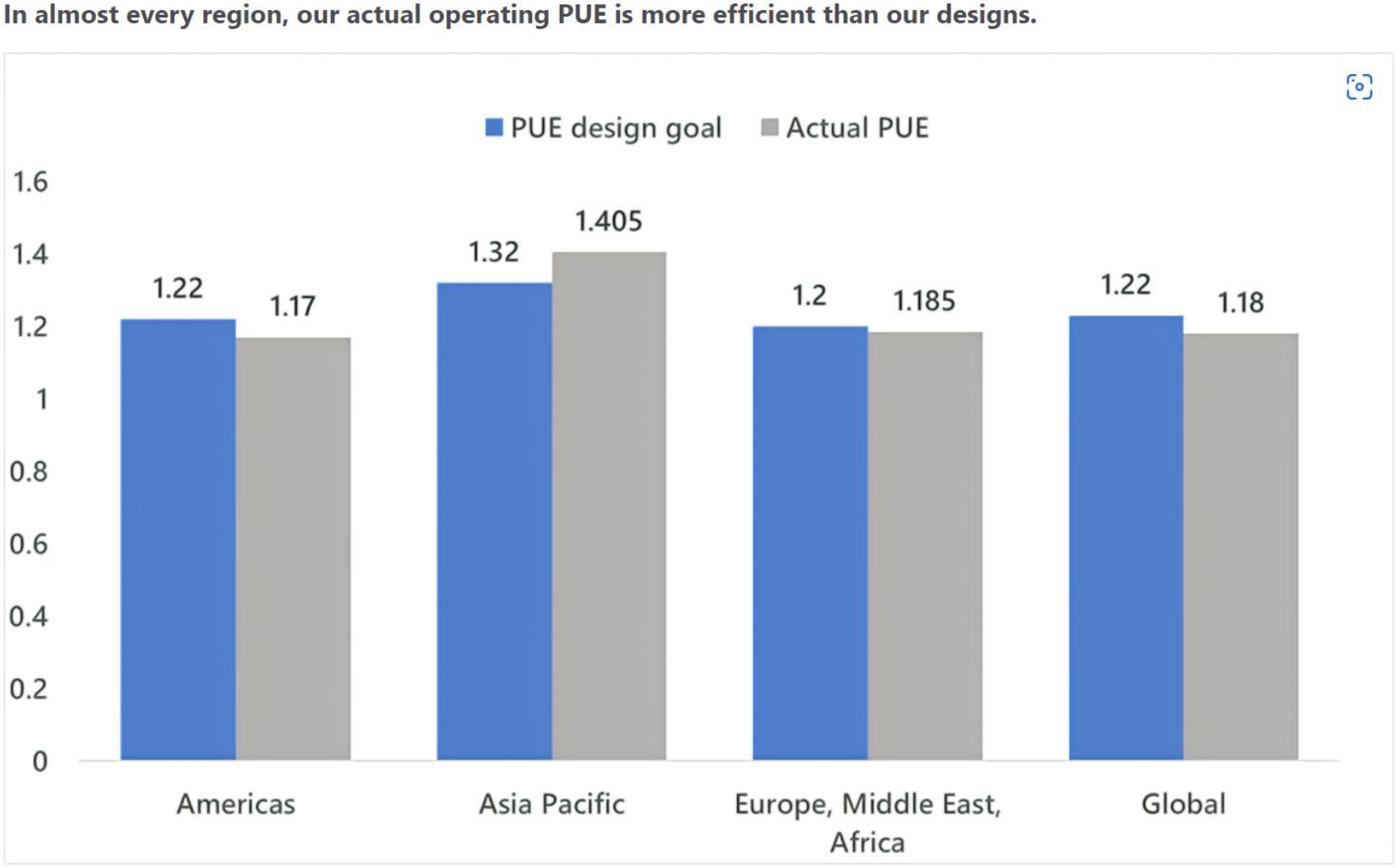
One of the biggest topics of discussion at COP26, the global climate conference held in November 2021, was how a lack of reliable and consistent measurement hampers progress on the path to Net Zero. I have been reflecting on this issue and, on this Earth Day, I would like to provide an update on how we are measuring energy and water use at our datacenters to improve sustainability across the Azure Cloud.

Today, we're sharing an important update on how Microsoft, and our datacenters, are helping to solve our part of this measurement challenge.

While the environmental goals are similar, each industry has unique challenges in measuring its carbon emissions to build its sustainability strategy. It's one of the key reasons we, together with ClimateWorks Foundation and 20 other leading organizations, launched the Carbon Call. It's also why we developed Microsoft Cloud for Sustainability, an Azure-based platform that allows organizations to combine disparate data sources into one place and help provide insights into how to improve their sustainability approaches.

Consommation énergétiques des datacenters de Microsoft

- Nos DC de dernières générations sont conçus pour avoir un PUE de 1.12
- Les chiffres ici sont la moyenne des DC d'une zone géographique



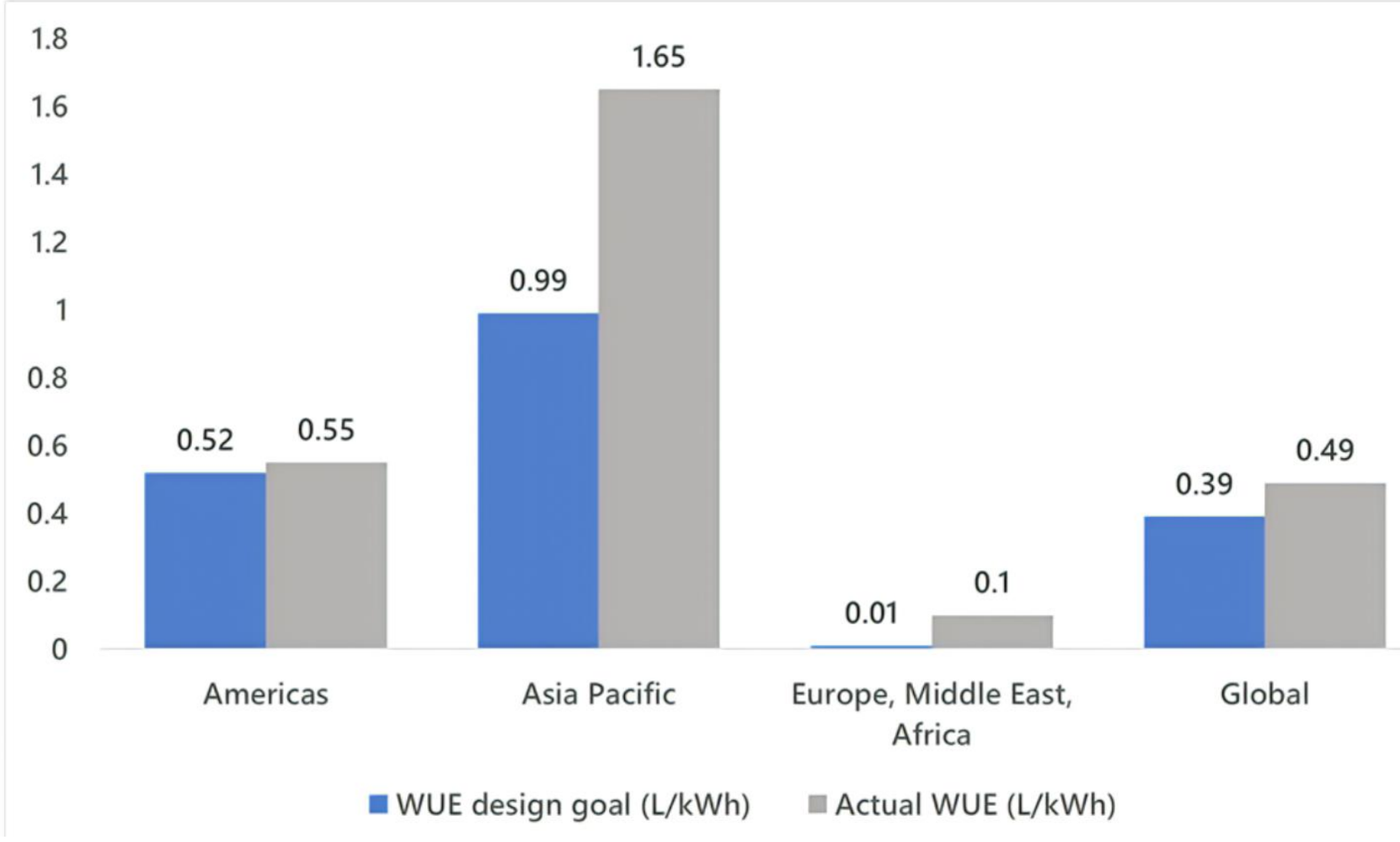
Water Usage Effectiveness (WUE)

- Une autre mesure très importante calculée en divisant le nombre de litres d'eau utilisés pour l'humidification et le refroidissement par le montant total de puissance (en kWh) nécessaire pour opérer les équipements IT du datacenter

$$\text{water usage effectiveness (WUE)} = \frac{annual\ liters\ of\ water\ used\ for\ humidification\ and\ cooling}{total\ annual\ kWh\ used\ to\ power\ IT\ equipment}$$

Consommation en eau des datacenters de Microsoft

- Nos DC récents dans des zones froides (Suède, Finlande) sont évidemment moins consommateurs d'eau



Azure | Global infrastructure

Microsoft datacenter sustainability fact sheets

We're working to ensure that Microsoft datacenters in the regions where we operate have as little impact on the environment as possible. Learn more about our exciting sustainability initiatives through our most recent datacenter fact sheets, organized into the geographies we operate our datacenters in. Please note that not all geographies and datacenter regions may be currently included; we will be providing updates and further announcements as we continue to grow our sustainability practices.

Asia Pacific
Singapore: Southeast Asia

Austria
Austria East

Brazil
Brazil South

Chile
Chile Central

Denmark
Denmark East

Europe
Ireland: North Europe
Netherlands: West Europe

Finland
Finland Central

Greece
Greece Central

India
India South Central

Indonesia
Indonesia Central

Italy
Italy North

Korea
Korea Central

Mexico
Mexico Central

New Zealand
New Zealand North

Poland
Poland Central

Spain
Spain Central

Sweden
Sweden Central

Taiwan
Taiwan North

United States
Arizona: West US 3
California: West US
Georgia: East US 3
Illinois: North Central US
Iowa: Central US
Texas: South Central US
Virginia: East US & East US 2
Washington: West US 2
Wyoming: West Central US

https://aka.ms/SustainabilityFactSheets

Microsoft Azure

Ireland

Sustainability fact sheet

As more people and businesses rely upon technology to stay connected, informed, and productive, digital needs in Ireland and around the globe are growing and that means the need for datacenters is growing too.

The Microsoft Cloud offers customers an energy efficient and carbon neutral alternative to running their own private datacenters. Research shows that Microsoft Cloud services can be up to 93 percent more energy efficient than traditional enterprise datacenters.

We're committed to providing a sustainable Microsoft Cloud, so we wanted to share information about how we take responsibility for our datacenter operations.

For Microsoft's datacenters in the North Europe region, located in Dublin, Ireland, we have included local sustainability investments and datapoints in support of meeting and exceeding our commitments around carbon, water, waste, and ecosystems.

Published September 2022. This document shares information we have as of the publication date, and it includes estimated information and projections. The information is provided as-is and may change without notice.

Carbon

Power usage effectiveness (PUE)

1.22

January 2021 – December 2021

Design PUE for new datacenters **1.12**

Microsoft's commitment

Microsoft pledged to become carbon negative by

2030

And by 2050, removing historical carbon since its founding in 1975

49% renewable energy coverage from wind power

approximate energy procured through June 2021

Microsoft will reduce Scope 1 and 2 emissions to near zero through energy efficiency work and by reaching

100% renewable energy supply by 2025.

In Ireland, we are transitioning from petroleum-based diesel to power our backup generators to a renewable biofuel blend that reduces net carbon emissions.

Microsoft's datacenters in Ireland are ISO50001 certified, and LEED Gold certified.

Microsoft has also committed by 2030 to be:

Diesel free

Reducing our Scope 3 emissions by more than half

100% of electricity consumption, 100% of the time, matched by zero carbon energy purchases

Learn about PUE and WUE

Water

Water usage effectiveness (WUE)

0.03 $\frac{L}{kWh}$

January 2021 – December 2021

In 2020, Microsoft pledged to be water positive for our direct operations by 2030.

Through this commitment, we will replenish the water consumed by datacenter operations in water-stressed regions. We have also committed to reduce water waste in our datacenter operations by 95% by 2024.

Blog / Announcements

Microsoft's newest sustainable datacenter region coming to Arizona in 2021

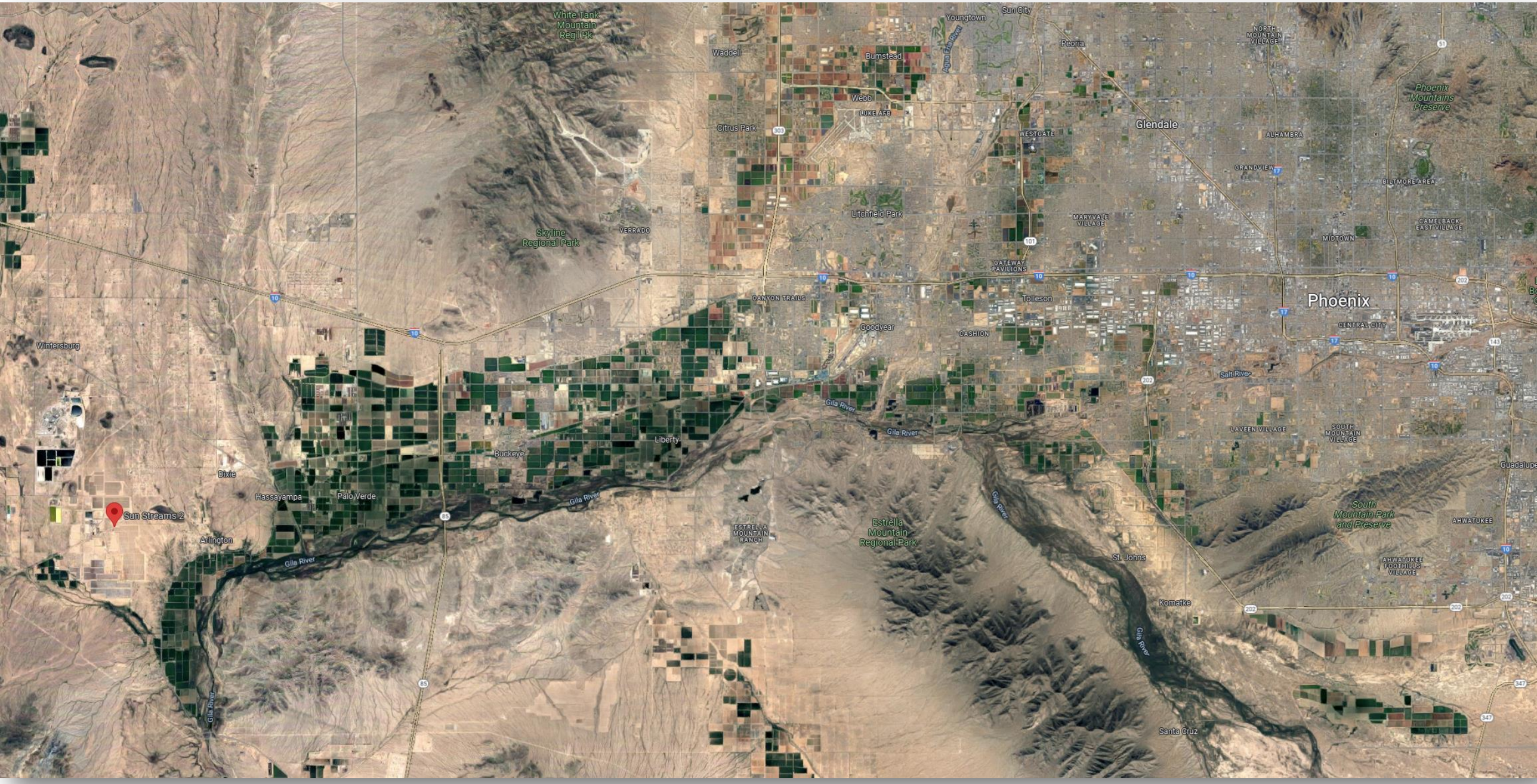
Posted on September 21, 2020

Brian Janous, General Manager, Energy and Sustainability

On our journey to become carbon negative by 2030, Microsoft is continually innovating and advancing the efficiency and sustainability of our cloud infrastructure, with a commitment to use 100 percent renewable energy in all of our datacenters and facilities by 2025. Today, we are taking a significant step toward that goal, revealing plans for our newest sustainable datacenter region in Arizona, which will become our West US 3 region.

Companies are not only digitally transforming their operations and products to become more sustainable—they're also choosing partners with shared goals and values. In developing the new West US 3 region, we have water conservation and replenishment firmly in mind. Today, Microsoft announced an ambitious commitment to be water positive for our direct operations by 2030. We're tackling our water consumption two ways: *reducing* our consumption and *replenishing* water in the regions we operate. Since we announced our plans to invest in solar energy in Arizona to build more sustainable datacenters last year, we have been working with the communities of El Mirage and Goodyear on water conservation, education and sustainability projects to support local priorities and needs.

West US 3

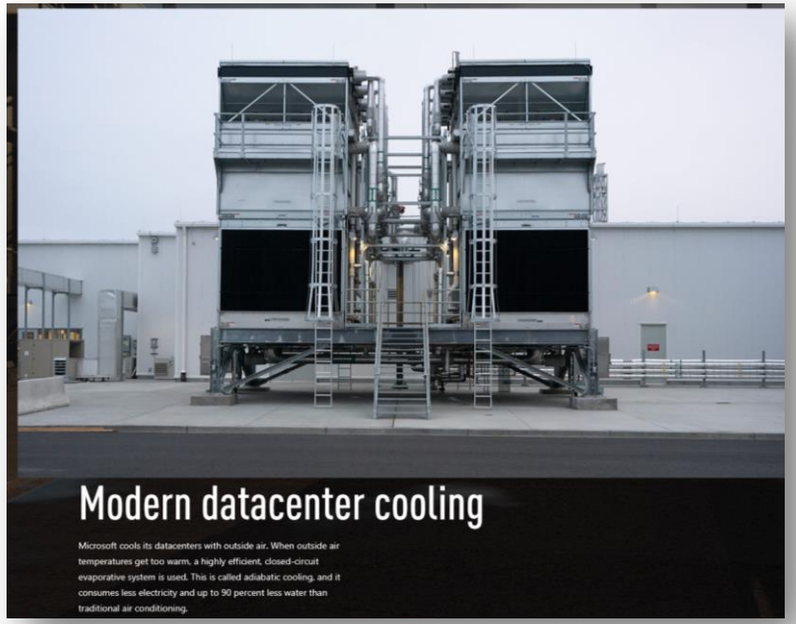

White Tank Mountain Regl Pk
Waddell
Youngtown
Sun City
Peoria
Bumstead
Agua Fria River
NORTH MOUNTAIN VILLAGE
Phoenix Mountains Preserve
Webb
LUKE AFB
Citrus Park
WESTGATE
Glendale
ALHAMBRA
GRANDVIEW
BILTMORE AREA
Litchfield Park
MARYVALE VILLAGE
CAMELBACK EAST VILLAGE
Skyline Regional Park
VERRADO
MIDTOWN
GATEWAY PAVILIONS
CANYON TRAILS
Tolleson
Phoenix
Goodyear
CASHION
CENTRAL CITY
Wintersburg
Salt River
Gila River
LAVEEN VILLAGE
SOUTH MOUNTAIN VILLAGE
Liberty
Buckeye
Guadalupe
Dixie
Hassayampa
Palo Verde
Sun Streams 2
South Mountain Park and Preserve
AHWATUKEE
ESTRELLA MOUNTAIN RANCH
Estrella Mountain Regional Park
Arlington
St. Johns
AHWATUKEE FOOTHILLS VILLAGE
Komatke
Santa Cruz

LONGROAD ENERGY COMPLETES TERM FINANCING OF 200 MWdc SUN STREAMS 2 SOLAR PROJECT

Project declares commercial operations; tax equity and lenders have funded the permanent capital structure

Boston, MA — July 20, 2021 – Longroad Energy, a US-based renewable energy developer, owner and operator, announced today the completion and funding of the term financing for Sun Streams 2, its 200 MWdc solar project located in Maricopa County, Arizona. Longroad owns 100 percent of the project after acquiring it in early 2021 from First Solar, the original developer.

Another primary area of energy use in the datacenter is keeping the servers cool to maintain performance and the lifetime use of the server. Microsoft's Arizona datacenters will use zero water for cooling for more than half the year, leveraging a method called adiabatic cooling, which uses outside air instead of water for cooling when temperatures are below 85 degrees Fahrenheit. When temperatures are above 85 degrees, an evaporative cooling system is used, which operates like "swamp coolers" in residential homes. This system is highly efficient, using less electricity and a fraction of water used by other water-based cooling systems, such as cooling towers. In addition, we're pursuing Leadership in Energy and Environmental Design (LEED) Gold certification, which will help conserve additional resources including energy and water, generate less waste, and support human health. Finally, we are committed to zero-waste certified operations for this new region, which means a minimum of 90 percent of waste will be diverted away from landfills through reduction, reuse, and recycling efforts.



Indirect evaporative cooling (IDEC)

IDEC uses a "fluid-cooler" that takes advantage of water evaporation to cool air flowing through an external air-to-water heat exchanger (radiator) to remove heat from the datacenter and return chilled cooling fluid to it. IDEC is a closed system that prevents polluted air from being introduced into the datacenter and does not require potable water.

Power usage	Low
Water usage	Medium
Cost	Low

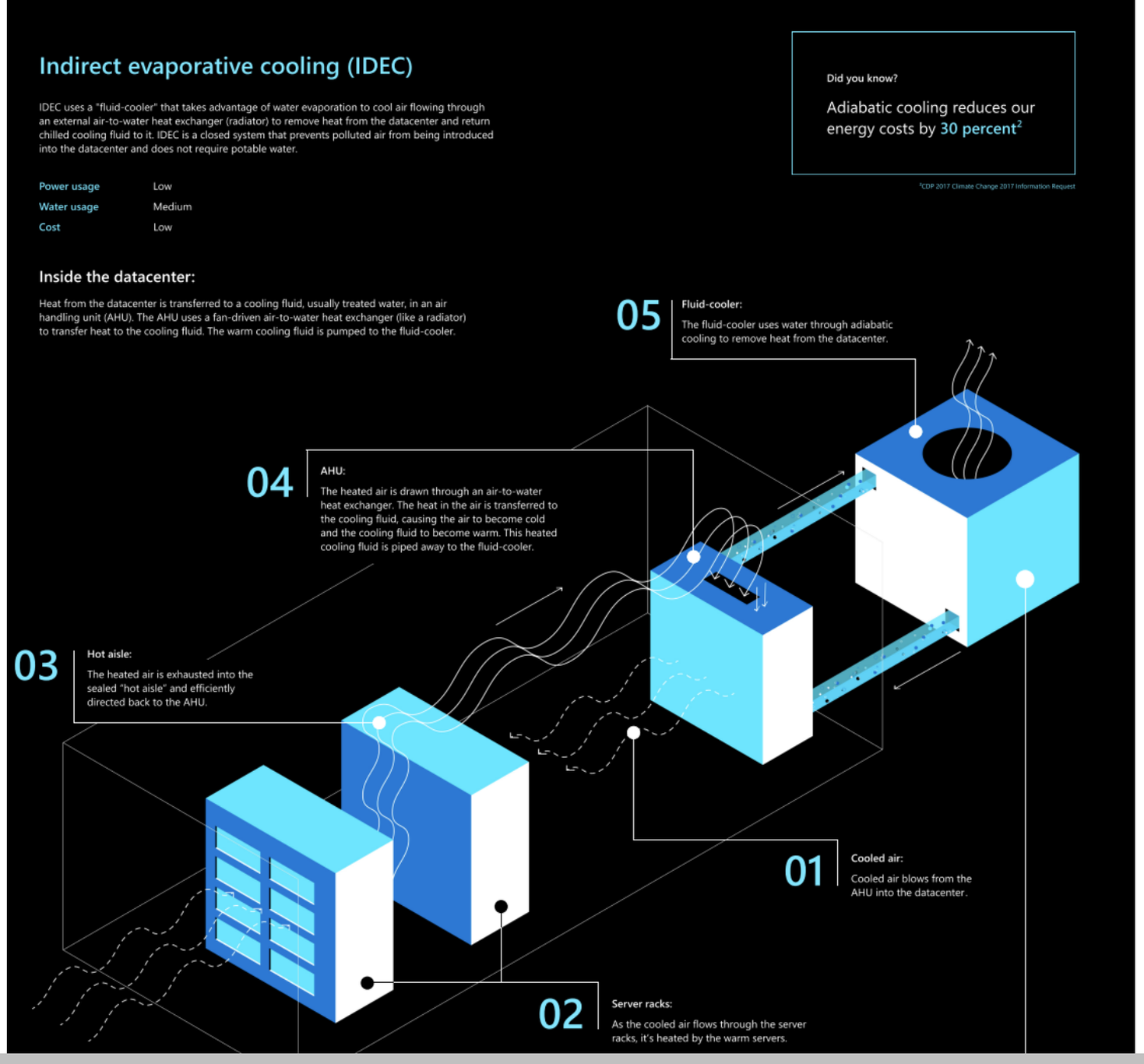
Did you know?

Adiabatic cooling reduces our energy costs by 30 percent[2]

[2]CDP 2017 Climate Change 2017 Information Request

Inside the datacenter:

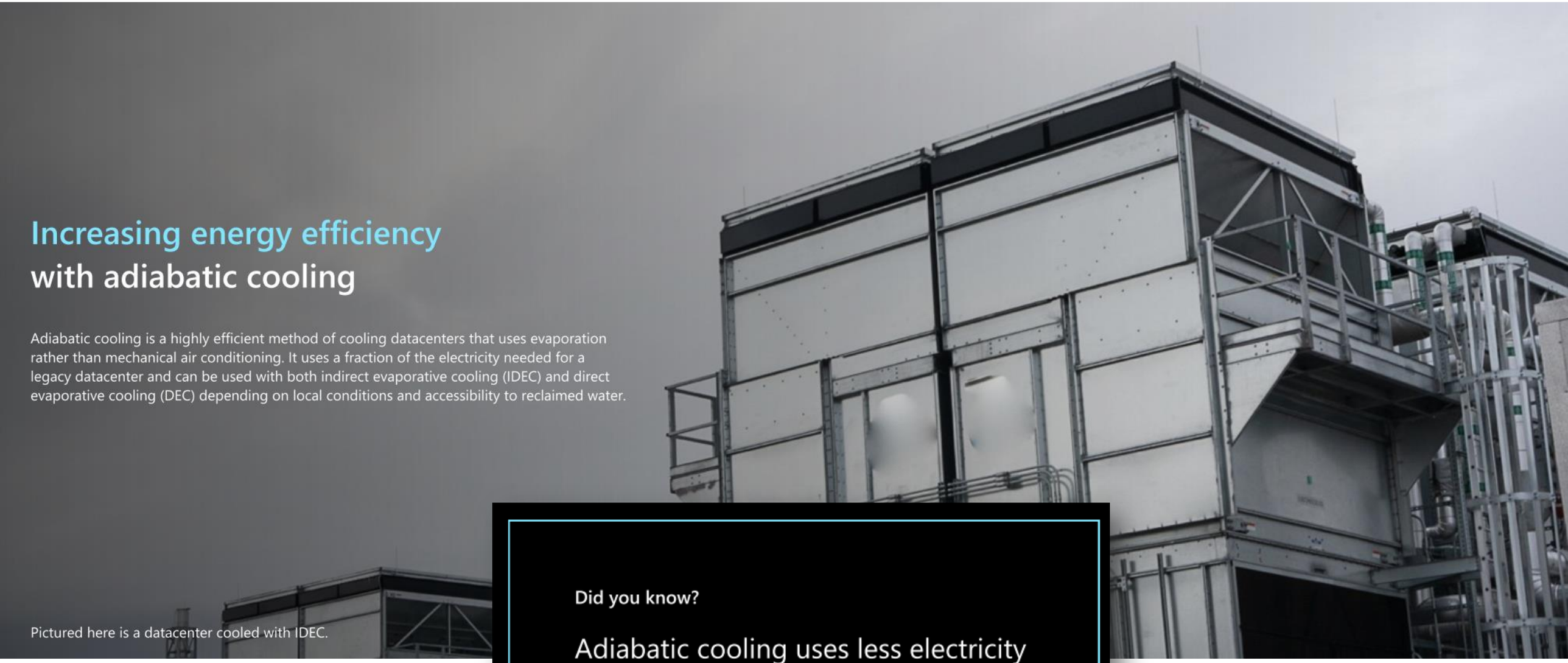
Heat from the datacenter is transferred to a cooling fluid, usually treated water, in an air handling unit (AHU). The AHU uses a fan-driven air-to-water heat exchanger (like a radiator) to transfer heat to the cooling fluid. The warm cooling fluid is pumped to the fluid-cooler.



Source : https://go.microsoft.com/fwlink/p?linkid=2158328&clcid=0x409

Increasing energy efficiency with adiabatic cooling

Adiabatic cooling is a highly efficient method of cooling datacenters that uses evaporation rather than mechanical air conditioning. It uses a fraction of the electricity needed for a legacy datacenter and can be used with both indirect evaporative cooling (IDEC) and direct evaporative cooling (DEC) depending on local conditions and accessibility to reclaimed water.



Pictured here is a datacenter cooled with IDEC.

Did you know?

Adiabatic cooling uses less electricity and up to **90 percent** less water than other water-based cooling systems[3]

Microsoft announces plans to establish a new datacenter region in Denmark to accelerate the country's green, digital transformation

The investment includes a sustainable datacenter region, powered by 100 percent renewable energy, the ability to store data in Denmark, and a plan to bring digital upskilling to 200,000 Danes by 2024

COPENHAGEN, Denmark – December 7, 2020 – Microsoft Corp. today announced the most significant investment in its 30-year

Innovating for a sustainable future

With 100 percent renewable energy, the datacenter region will enable Danish customers to reduce their own carbon footprint associated with computing while accelerating digitization. According to The Footprint Firm**, the total emissions from on-premises servers in Denmark amount to an estimated 355,000 tons of CO2 per year. A report from WSP*** shows that Microsoft Cloud services can be up to 93 percent more energy efficient than traditional enterprise datacenters.

Microsoft is working with Energinet to explore future ways of tracking and matching renewable energy on an hourly basis. Building on Energinet's pioneering work and open-source data approach with Azure, Microsoft intends to support the launch of a new renewable energy supply option, which provides customers more transparency about carbon mitigation from clean energy.

How Microsoft's new datacenter region in Sweden incorporates the company's sustainability commitments

Microsoft opens its sustainable datacenter region in Sweden, creating new opportunities for a cloud-first Sweden

16/11/2021

STOCKHOLM, Nov. 16, 2021 – Today, Microsoft Corp. announced the launch of its newest sustainable datacenter region in Sweden, with presence in Gävle, Sandviken and Staffanstorp. Microsoft's new world-class datacenters in Sweden are open for business with Microsoft Azure and Microsoft 365 available today, delivering advanced data security, privacy and the ability to store data in the country. The new datacenter region brings the best of Microsoft's sustainability investments, powering the datacenter with 100% carbon-free energy and supporting zero-waste operations, underscoring Microsoft's ongoing investment to help create new long-term opportunities across both commercial and public sectors in Sweden.

VATTENFALL

Who we are | What we do | Investors | **Press and media** | Careers

Media bank | Press contacts

PRESS RELEASE • DECENTRALISED SOLUTIONS • 24 NOVEMBER 2020, 12:15 CET • 2 MIN

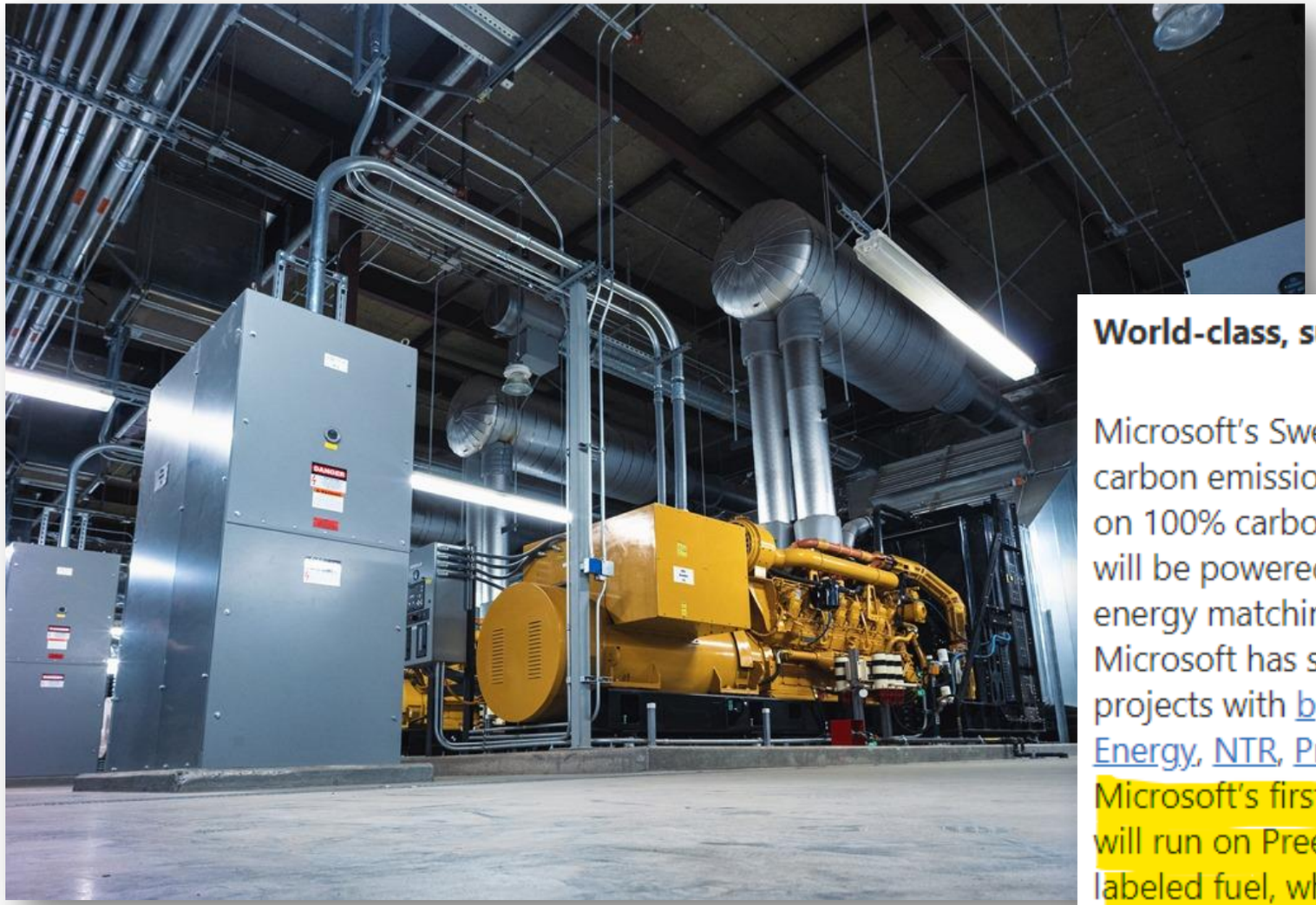
Vattenfall to deliver renewable energy 24/7 to Microsoft's Swedish datacenters

Vattenfall and Microsoft in Sweden piloted 24/7 Matching solution last year and are now intensifying the partnership as energy partners. The 24/7 Matching solution will be used in the new Microsoft datacenters in Sweden to measure renewable energy consumption per hour. Vattenfall will also deliver EPD (Environmental Product Declaration) labelled hydro and wind power 24/7 to the three datacenters in 2021.

Quid des générateurs et autres UPS ?

- Microsoft cherche un remplaçant sans émission de carbone aux générateurs fonctionnant avec des moteurs Diesel
- Objectif : supprimer l'usage de fuel (diesel) pour les générateurs de nos datacenters d'ici 2030
- Ceci tout en conservant l'objectif de supporter la continuité des opérations dans l'éventualité d'une panne de courant globale

Quid des générateurs et autres UPS ?



World-class, sustainable datacenters in Sweden

Microsoft's Sweden datacenter region is committed to cutting carbon emissions, achieving zero waste certification, and running on 100% carbon-free energy. Microsoft's Sweden datacenter region will be powered by 100% carbon-free energy with 24/7 hourly energy matching with partner Vattenfall. To support its operations, Microsoft has signed agreements for new renewable energy projects with bp, Enlight Renewable Energy, European Energy, NTR, Prime Capital and wpd. In addition, Sweden is Microsoft's first datacenter region whose backup generators will run on Preem Evolution Diesel Plus, the world's first Nordic Eco-labeled fuel, which contains at least 50% renewable raw material, and nearly an equivalent reduction in net carbon dioxide emissions compared with standard fossil diesel blends.

"What makes a datacenter a datacenter is that it can operate even though the grid is not. When there's a blackout, the servers stay up. That's the difference between a datacenter and a warehouse chock full of computers."

Sean James, Microsoft's director of datacenter research

Pile à hydrogène

- Microsoft a commence à étudier la technologie fuel cell dès 2013 avec le National Fuel Cell Research Center de l'Université de Californie
- A partir de 2018, cette technologie est considérée comme une option potentielle et future au remplacement des générateurs Diesel

Power Innovations built a 250-kilowatt fuel cell system to help Microsoft explore the potential of using hydrogen fuel cells for backup power generation at datacenters. In a proof of concept, the system powered a row of datacenter servers for 48 consecutive hours. Credit: Power Innovations.

Fuel cell technology

- En 2020, Microsoft engage Power Innovations à Salt Lake City (Utah) pour contruire un système pouvant alimenter 10 racks d'un datacenter

Microsoft used hydrogen stored in tanks on trailers parked outside a lab near Salt Lake City, Utah, to fuel hydrogen fuel cells that powered a row of datacenter servers for 48 consecutive hours. Credit: Power Innovations.

Source : https://news.microsoft.com/innovation-stories/hydrogen-datacenters/

Hydrogen fuel cells could provide emission free backup power at datacenters, Microsoft says

John Roach

Jul 28, 2022

Latham, New York – Hydrogen fuel cells packed into a pair of 40-foot-long shipping containers here ramped up on an overcast day early this June as engineers gathered around laptops displaying data on the state, health and power output of the cells in this first-of-a-kind hydrogen generator.

"This is it, it's running at three megawatts right now," exclaimed Mark Monroe, a principal infrastructure engineer on Microsoft's team for datacenter advanced development.

A round of applause and shouts of joy from engineers at Plug, who built the fuel cell system, pierced the white noise from fans atop the shipping containers and pipes venting steam. The moment was the latest milestone on Microsoft's quest to find a zero-carbon emissions replacement for the backup diesel-powered generators that support continuous operations in the event of power outages and other service disruptions.

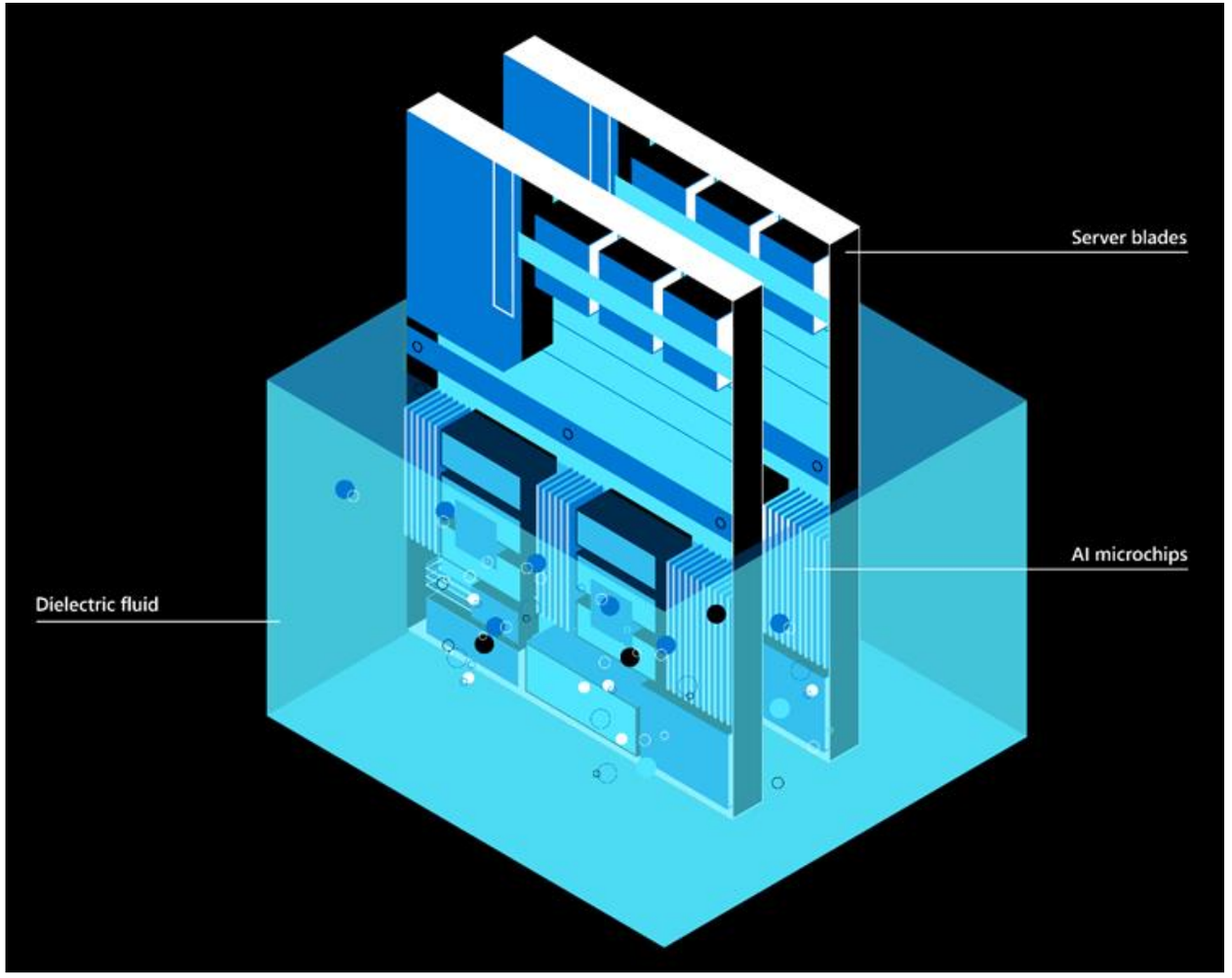
"What we just witnessed was, for the datacenter industry, a moon landing moment," said Sean James, Microsoft's director of datacenter research. "We have a generator that produces no emissions. It's mind-blowing."

Source : https://news.microsoft.com/innovation-stories/hydrogen-fuel-cells-could-provide-emission-free-backup-power-at-datacenters-microsoft-says/

- *"After that successful proof-of-concept demonstration, the team set out to prove the viability of a three-megawatt system, which is of sufficient size to replace a diesel generator at a datacenter*

- *The problem, Monroe noted, was that nobody made PEM fuel cell systems that large – three megawatts is more than 10 times bigger than the system the company tested in Utah. Three megawatts is enough energy to power about 10,000 computer servers or 600 homes "*

Refroidissement des serveurs par immersion



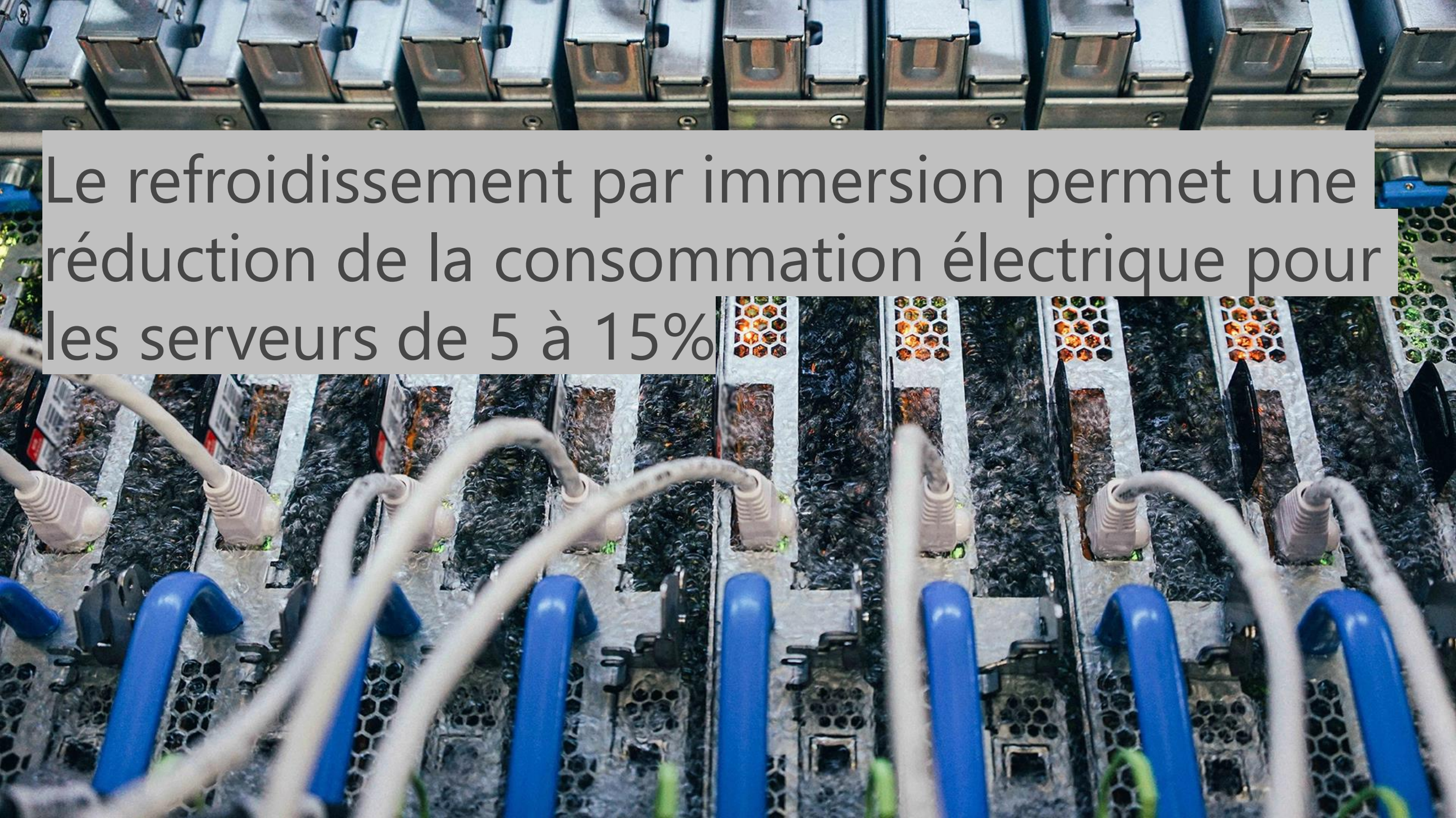
Permettra :

- De baisser la consommation électrique de 5 à 15% et de réduire la consommation en eau (utilisée par les systèmes de refroidissement) des Datacenters
- De densifier les microprocesseurs par serveur et de réduire le nombre de racks et de diminuer la taille des datacenters

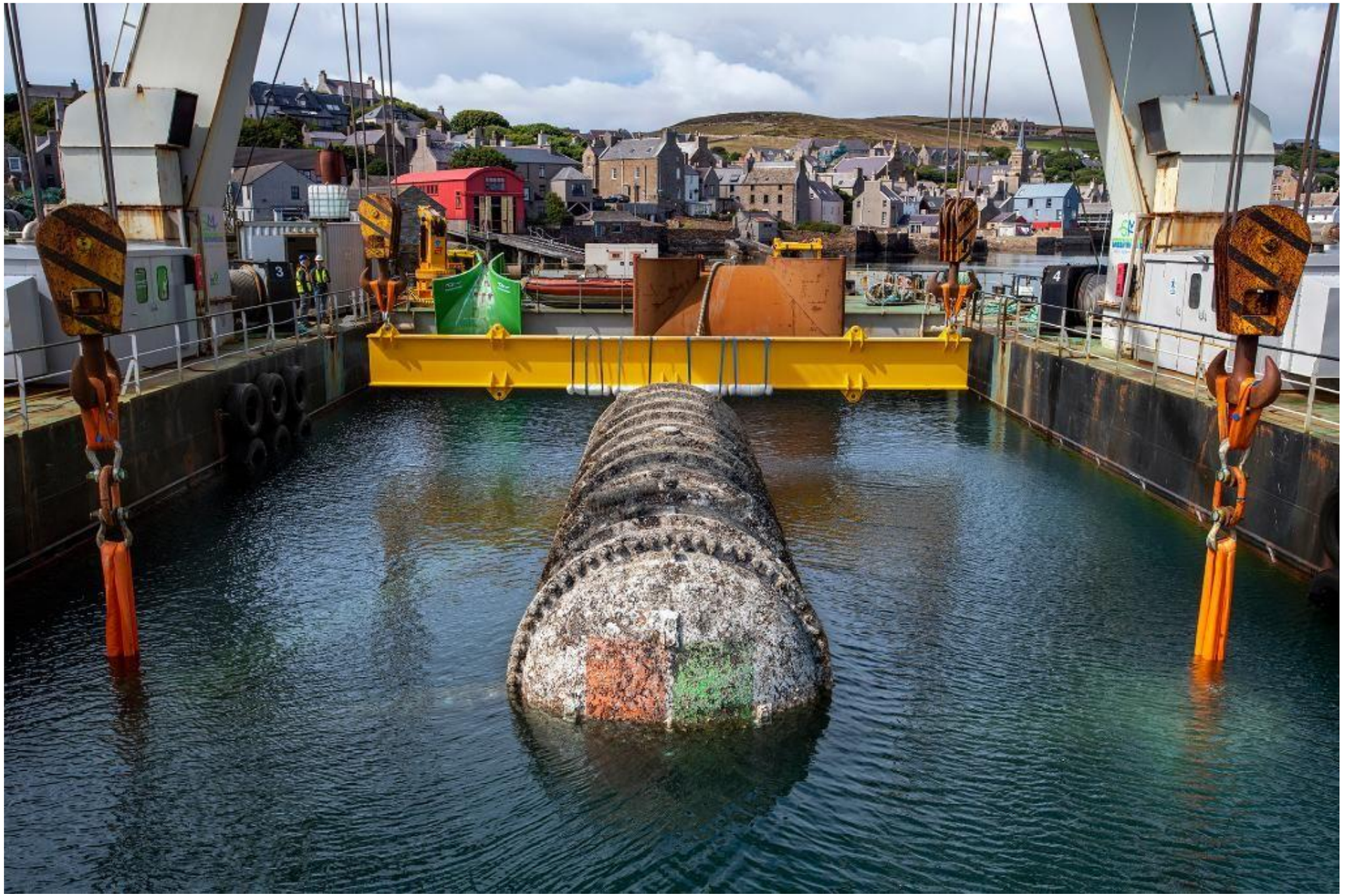
Source : https://download.microsoft.com/download/7/2/8/72830831-5d64-4f5c-9f51-e6e38ab1dd55/Azure_Sustainability_Liquid_immersion_cooling_infographic.pdf


Service Mode
Dashboard
Mode: NA
Pump: On
Liquid Level
Tank Pressure
Condenser Inlet Temp
Flow Rate
Microsoft
Humidity
Liquid Temp
Flow Outlet Status
Flow Inlet Status
Incidents

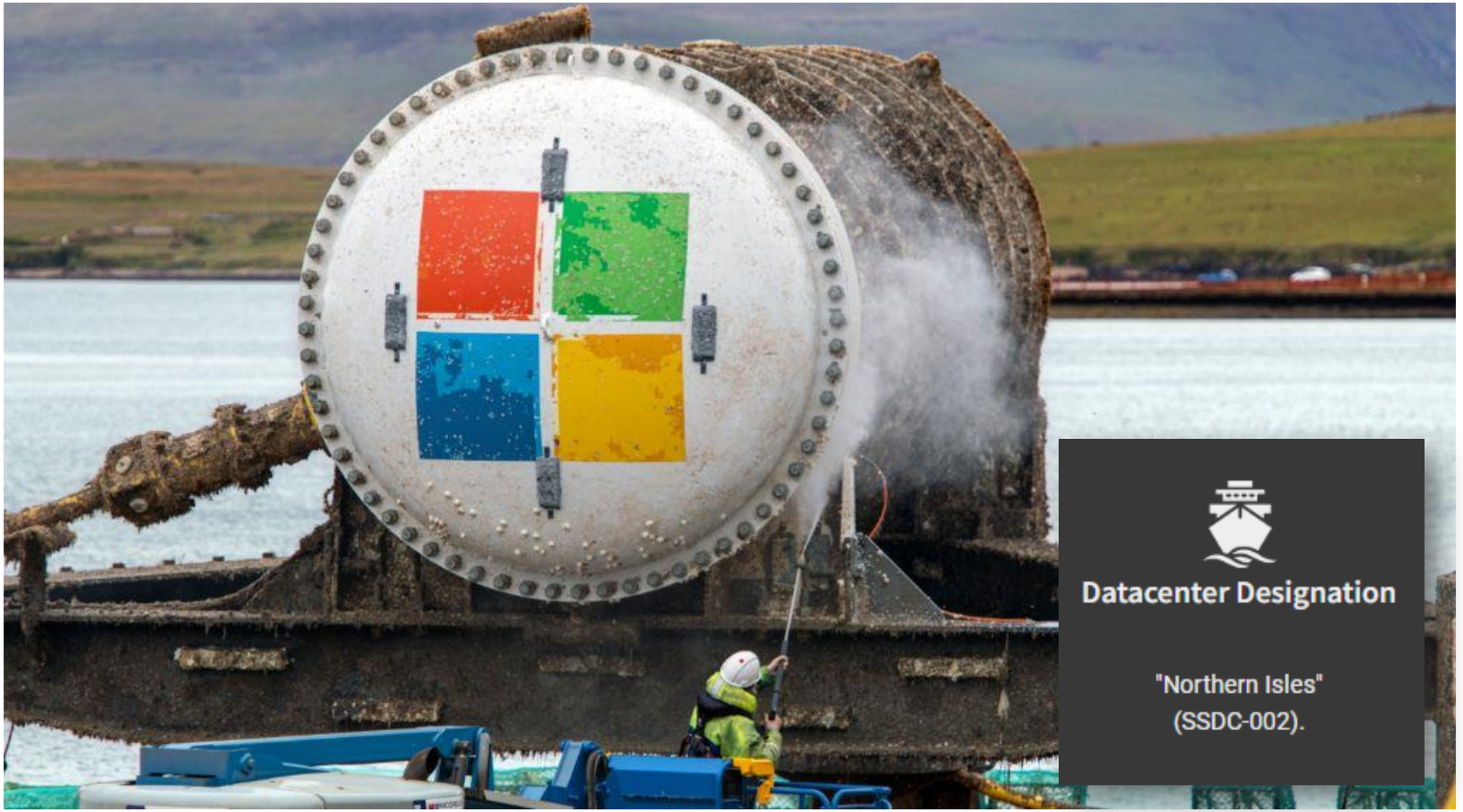
Le refroidissement par immersion permet une réduction de la consommation électrique pour les serveurs de 5 à 15%



Source : https://news.microsoft.com/innovation-stories/datacenter-liquid-cooling/



https://news.microsoft.com/innovation-stories/project-natick-underwater-datacenter/


Datacenter Designation
"Northern Isles"
(SSDC-002).

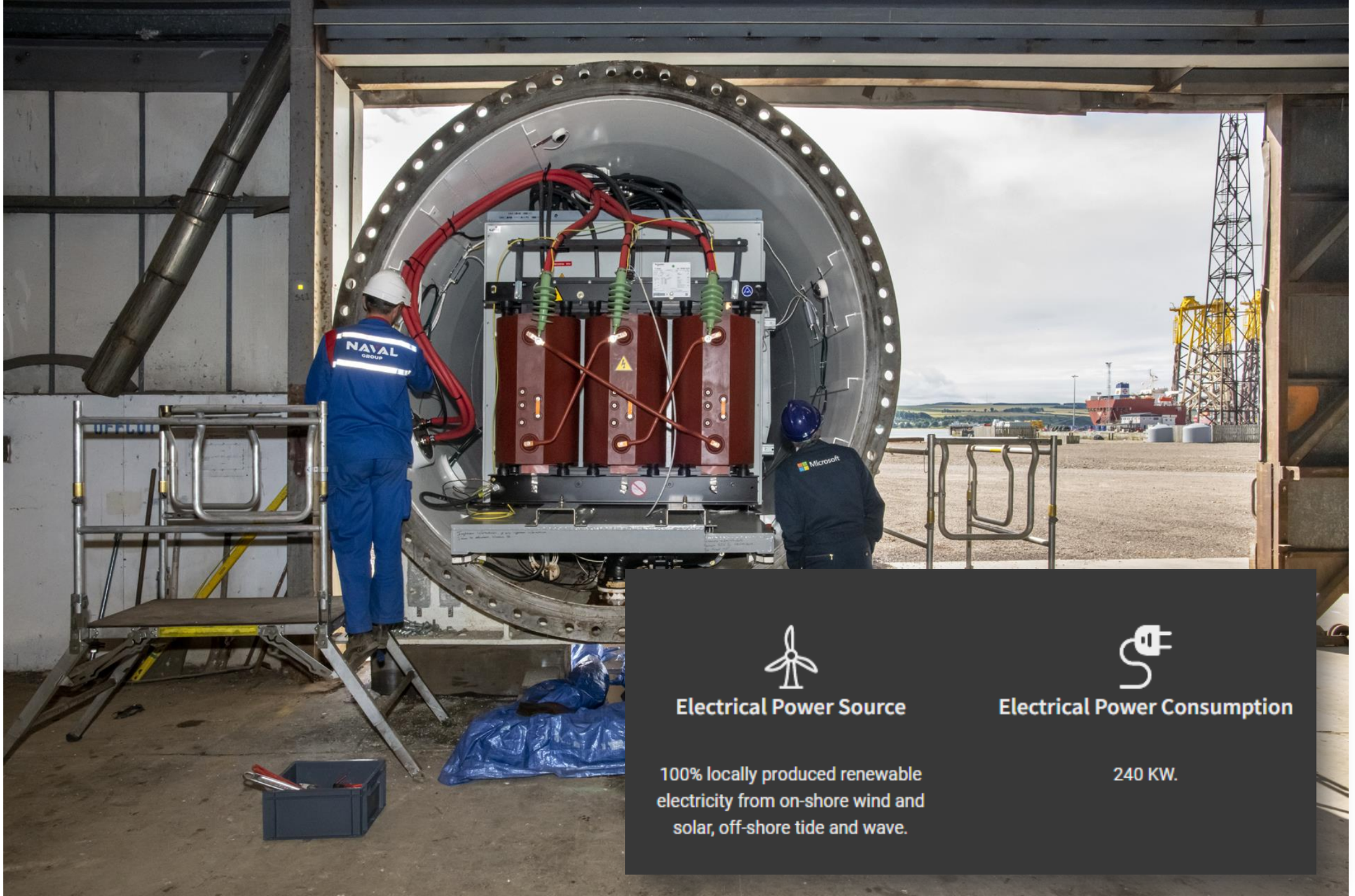
Pressure Vessel Dimensions

12.2m length, 2.8m diameter (3.18m including external components); about the size of a 40' ISO shipping container you might see on a ship, train, or truck.

Subsea Docking Structure Dimensions

14.3m length, 12.7m width.


https://natick.research.microsoft.com/
GLOBAL ENERGY GROUP

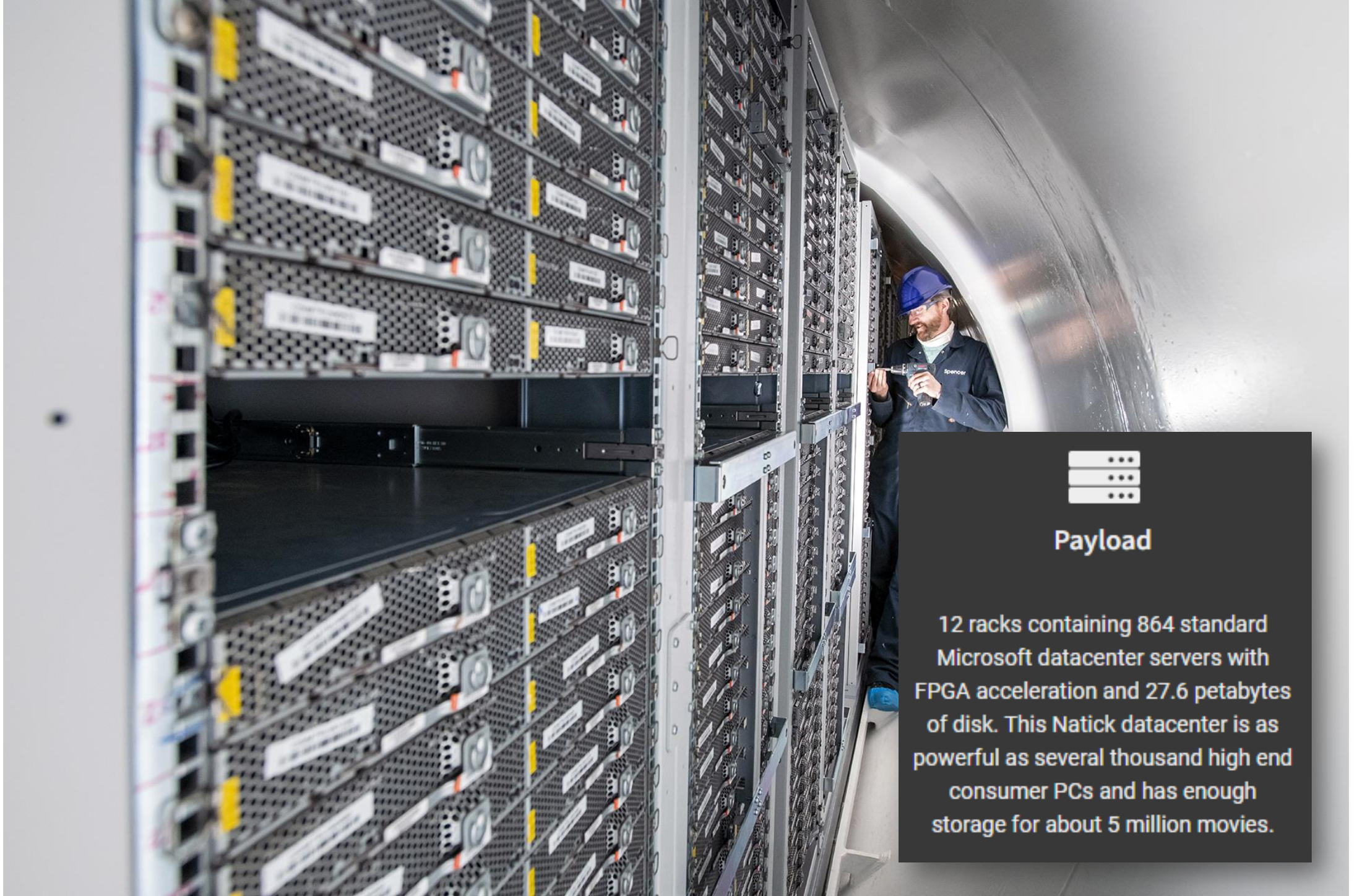


Electrical Power Source

100% locally produced renewable electricity from on-shore wind and solar, off-shore tide and wave.

Electrical Power Consumption

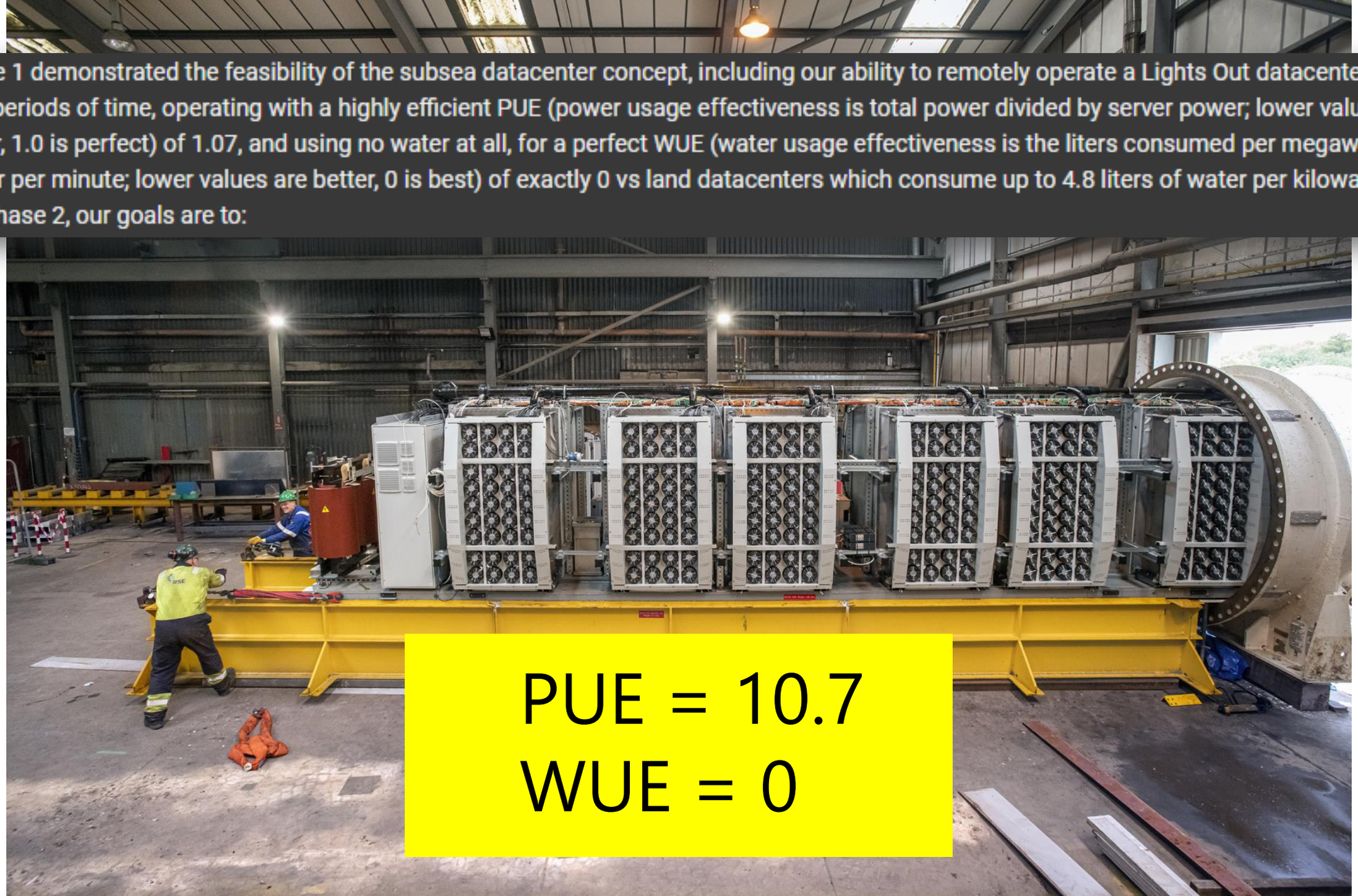
240 KW.



Payload

12 racks containing 864 standard Microsoft datacenter servers with FPGA acceleration and 27.6 petabytes of disk. This Natick datacenter is as powerful as several thousand high end consumer PCs and has enough storage for about 5 million movies.

Phase 1 demonstrated the feasibility of the subsea datacenter concept, including our ability to remotely operate a Lights Out datacenter* for long periods of time, operating with a highly efficient PUE (power usage effectiveness is total power divided by server power; lower values are better, 1.0 is perfect) of 1.07, and using no water at all, for a perfect WUE (water usage effectiveness is the liters consumed per megawatt of power per minute; lower values are better, 0 is best) of exactly 0 vs land datacenters which consume up to 4.8 liters of water per kilowatt-hour. For Phase 2, our goals are to:




1.5 YEARS IN
Project Natick
3 test groups + control
• Nitrogen + Normal temp
• Nitrogen + Cold
• Nitrogen + Constant temp
1/8 failure rate of control
Pulled up last month and being studied currently

Figure 3: The life cycle phases used to define the boundary of energy consumption and carbon emissions considered in the analysis.

Circular Datacenters

Learn how Microsoft Circular Centers are scaling cloud supply chain sustainability

Posted on March 15, 2022

Rani Borkar, Corporate Vice President, Azure Hardware Systems and Infrastructure

Aiming at delivering the most sustainable, scalable, and reliable cloud for Azure customers, continued innovation in cloud hardware is a constant priority for Microsoft. This extends beyond server architecture and rack design to include intelligent provisioning, deployment, and ultimately, decommissioning of cloud computing hardware in datacenters.

As we look to deliver upon Microsoft's commitments towards a net-zero carbon future, our cloud supply chain has integrated a zero-waste philosophy into every stage of the datacenter hardware lifecycle.

Source : https://azure.microsoft.com/en-us/blog/learn-how-microsoft-circular-centers-are-scaling-cloud-supply-chain-sustainability/

Circular Datacenters

In January of 2022, the Microsoft Cloud supply chain achieved significant milestones toward its goal of reusing 90 percent of its cloud computing hardware assets by 2025. We launched two additional Circular Centers, which process decommissioned cloud servers and hardware, sort, and intelligently channel the components and equipment to optimize reuse or repurpose.

Our pilot Circular Center opened in Amsterdam in 2020, and the new centers that went live this year are located at our datacenter campuses in Dublin, Ireland, and Boydton, Virginia. We plan to expand the program at Microsoft datacenters in Quincy, Washington; Chicago, Illinois; Singapore and additional sites over the next few years in Des Moines, Iowa; San Antonio, Texas; Cheyenne, Wyoming; Sydney, Australia; Sweden, and more.

So far, the Amsterdam Circular Center has achieved 83 percent reuse and 17 percent recycle of critical parts while contributing to the goal of reducing carbon emissions by 145,000 metric tons CO2 equivalent. This innovative approach

Source : https://azure.microsoft.com/en-us/blog/learn-how-microsoft-circular-centers-are-scaling-cloud-supply-chain-sustainability/
https://customers.microsoft.com/en-us/story/1431789627332547010-microsoft-circular-centers

Microsoft is pursuing Leadership in Energy and Environmental Design (LEED) Gold certification for the region, which will help conserve additional resources like energy and water, generate less waste and support human health. In addition, the new datacenter region will be home to Microsoft's third European Circular Center, which repurposes servers and hardware in our datacenter. These centers are the first of their kind in the industry. The Swedish Circular Center will have the ability to support 12,000 servers per month. Supporting the company's water positive by 2030 commitment, the datacenters' servers will be cooled with only outside air 100% of the year, and rainwater will be captured at the datacenter, used primarily to provide humidification and to support onsite datacenter facilities.

Source : https://news.microsoft.com/europe/2021/11/16/microsoft-opens-its-sustainable-datacenter-region-in-sweden-creating-new-opportunities-for-a-cloud-first-sweden/

L'autre visage du FinOps

GreenOps

Cloud efficiency

Cloud efficiency = resource cost + carbon emissions + energy consumed

Cost

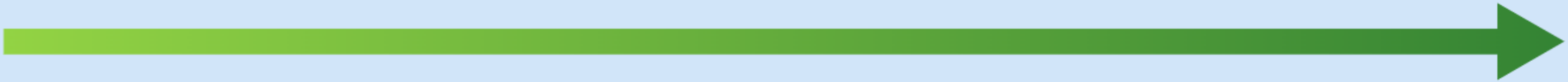
- Optimize resources
- Architecture apps for lower costs

Carbon

- Optimize emissions
- Green architectures

Energy

- Optimize energy used
- Modernize apps



Source : Azure Well Architected Framework

Choisir des régions moins impactantes en termes d'émissions

- Régions avec des PUE faibles (energy efficiency)
- Régions favorisant la consommation d'énergies renouvelables
- Régions avec mix énergétique acceptable en termes d'émission de carbone (carbon efficiency)

Optimisation de l’impact environnemental d’un SI

Optimisation de la valeur (et du coût) d’un SI

GreenOps ≠ FinOps

Optimisation de l’usage des ressources dans un SI

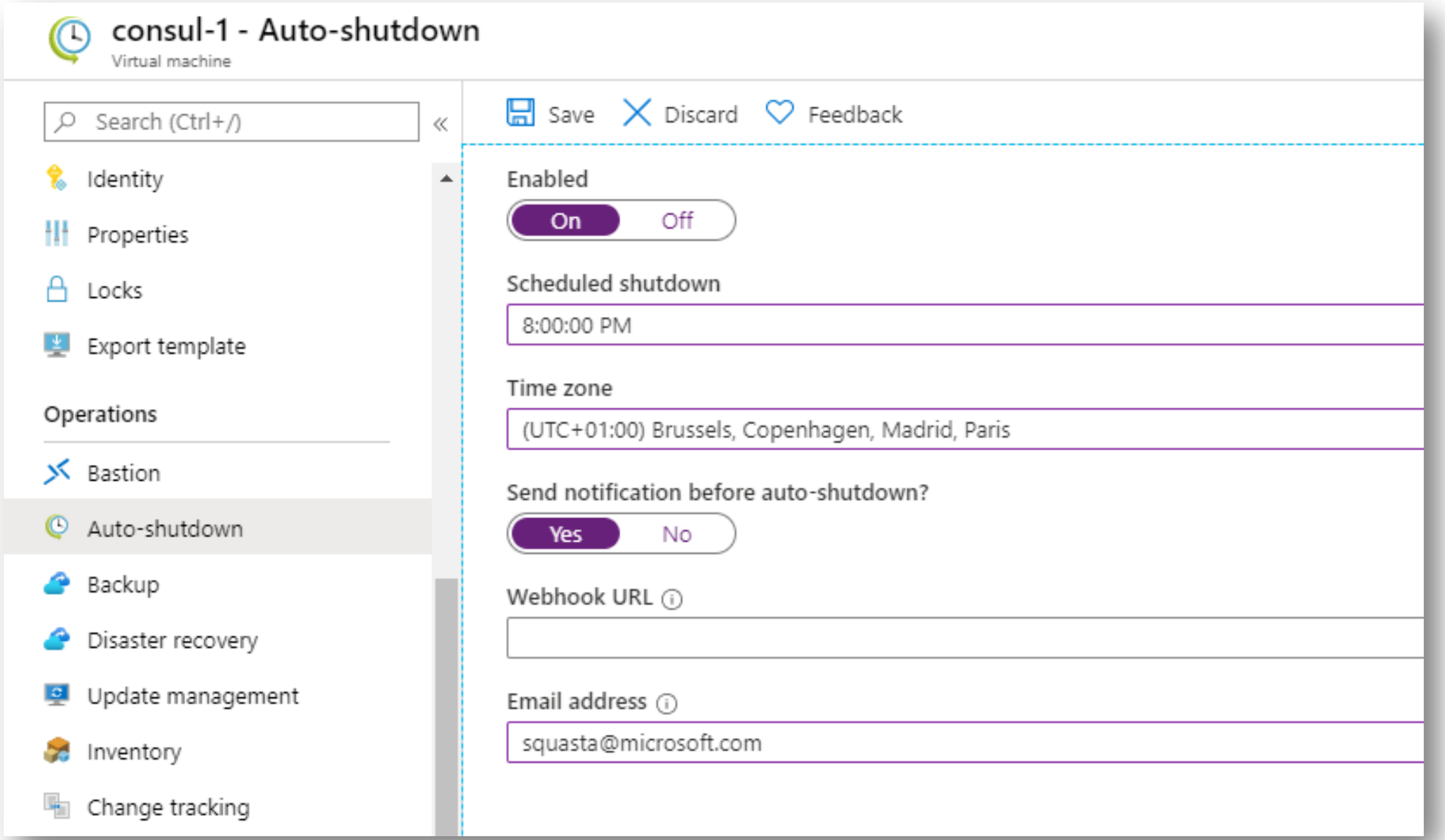
Optimisation de l'usage des ressources dans un SI – principe général

- Arrêter tout ce qui n'a pas besoin de tourner 7/7 H24
- Supprimer les ressources non utilisées
- Nettoyer ou limiter les données inutiles
- Choisir les options les moins énergivores

Arrêt automatique des VMs

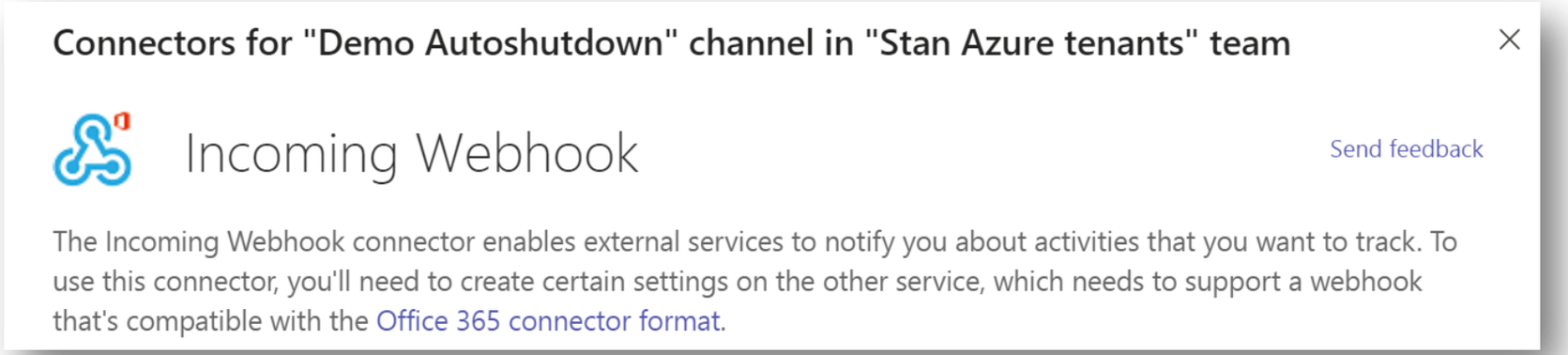
- Pour les environnements non critiques (Test, Dev..), planifier l'arrêt automatique des machines virtuelles et limiter leur usage aux heures ouvrées
- Economies et réduction de l'impact environnemental réalisables pouvant atteindre plus de 60% (versus un usage 24/7)

Auto-shutdown des VM

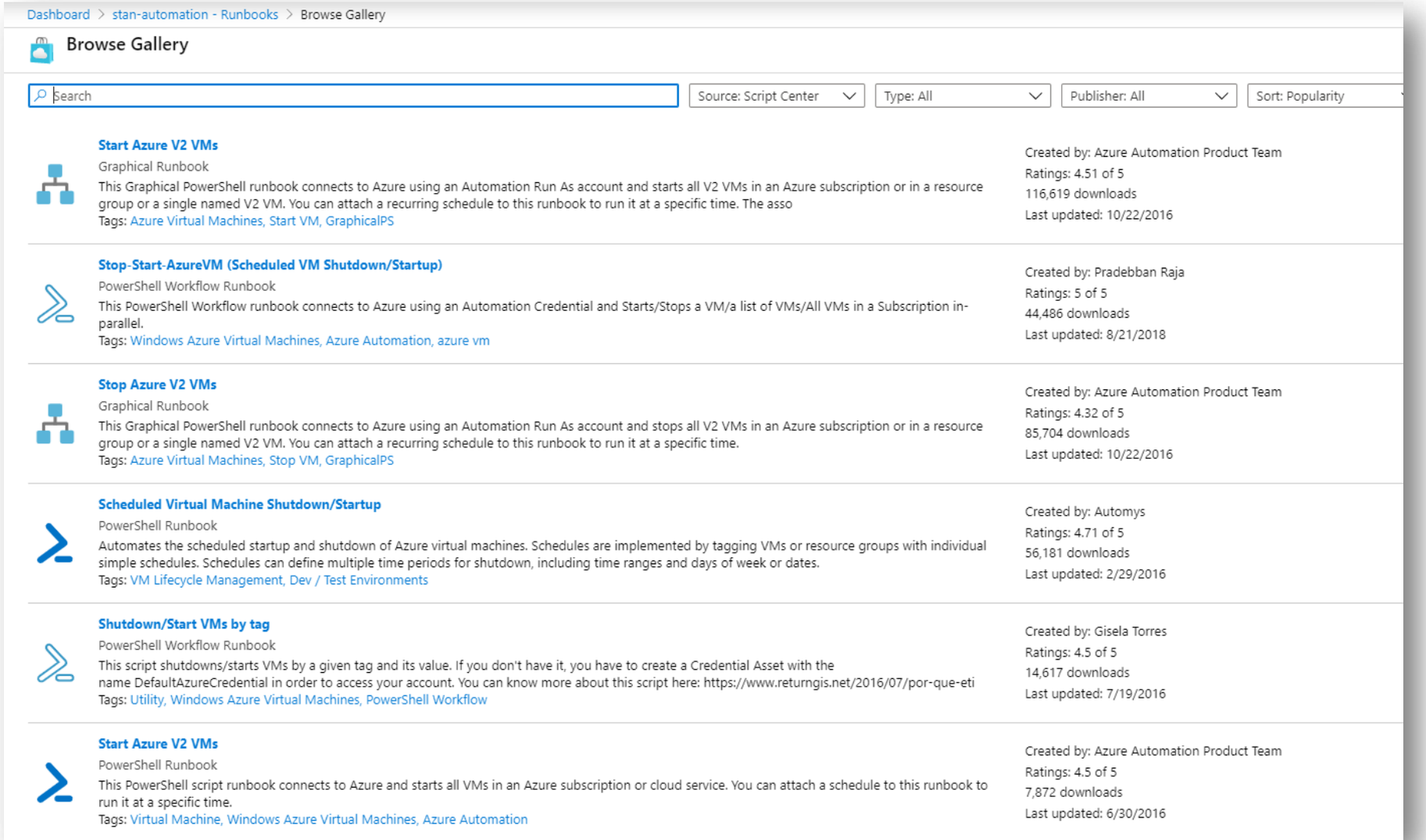


Notification dans Teams

- Via utilisation d'un Webhook

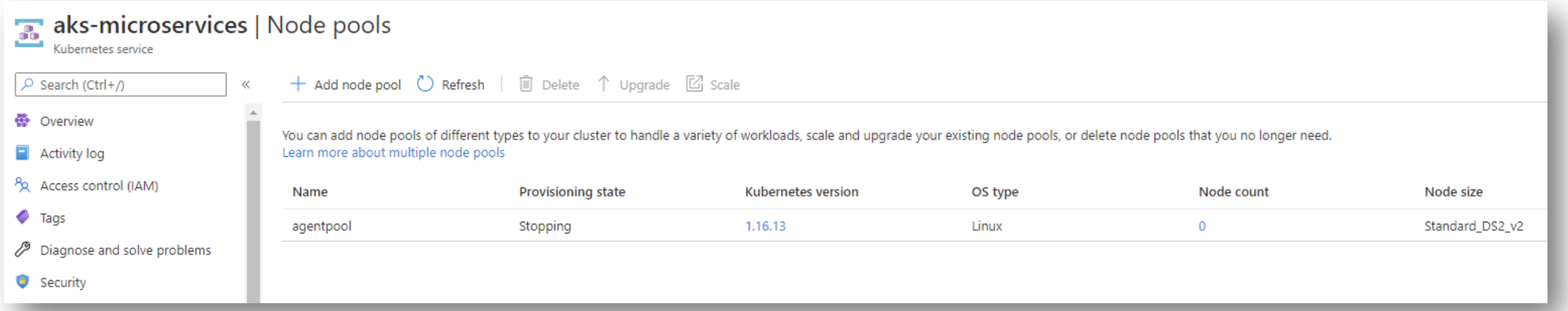


Démarrage et arrêt automatisés avec un Runbook Azure Automation



Arrêt de clusters Kubernetes

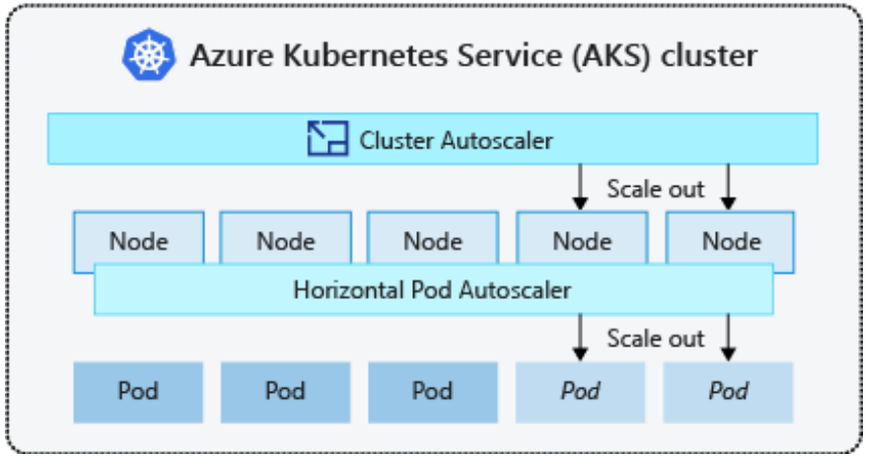
- Possibilité d'arrêter / démarrer un cluster Azure Kubernetes Service
  - Économie proportionnelle au nombre de machines worker nodes du cluster



- Nécessite un cluster en Virtual Machine Scale Set
- Possibilité aussi de scaler à 0 des node pools

Adapter la puissance au besoin pour les clusters Azure Kubernetes Service

- Utiliser les mécaniques d'autoscaling natifs à Kubernetes
  - Horizontal Pod Autoscaling
  - KEDA → event driven, possibilité de scaler à 0
- Combiner le scaling applicatif avec le scaling de l'infrastructure
  - AKS nodepool autoscale
  - AKS Virtual Kubelet



Arrêter les workloads instanciés dans Kubernetes

- Arrêter volontairement des ressources Kubernetes (deployment, cronjobs) pendant les périodes de non utilisation
  - A faire plutôt sur des environnements de type Dev & Test

- Un projet intéressant : kube-green
  - https://github.com/kube-green/kube-green
  - Projet encore jeune, mais à surveiller

kube-green

A Project by Davide Bianchi

Arrêt de base de données MySQL managée (PaaS)

NOW AVAILABLE

General availability: Stop/start functionality for Azure Database for MySQL

Published date: February 03, 2021

General availability support of database stop/start functionality for Azure Database for MySQL is now available. With this functionality, you can stop the database server when not in use (during non-work hours) and start it when it is back online. This could be helpful when using Azure Database for MySQL for dev/test scenarios to save on costs by turning off the server compute when not in use.

Learn more about this announcement in the stop/start documentation.

Arrêt de base de données PostgreSQL

- Stop/Start an Azure Database for PostgreSQL - Flexible Server
- https://www.sqlshack.com/azure-database-for-postgresql-cost-optimization/

Utiliser le bon stockage

- En fonction de la fréquence d'accès aux données, placer les données dans des tiers adaptés :
  - Accès fréquent, besoin de latence faible : tiers Premium, Hot
  - Accès occasionnel : tiers Cool
  - Accès rare : tiers Archive
- Plus le tiers sera froid, plus le coût du stockage au Go sera faible
- Automatiser le déplacement vers des tiers froids ou des suppressions avec des Livecycle Management Policy

demostanlivecycle - Lifecycle Management

Storage account

Search (Ctrl+/)

Blob service

- Containers
- Custom domain
- Data Protection
- Azure CDN
- Add Azure Search
- Lifecycle Management

File service

- File shares

Table service

- Tables

Queue service

- Queues

Monitoring

- Insights (preview)

Add rule | Enable | Disable | Refresh | Delete

List view | Code view

Lifecycle management offers a rich, rule-based policy for general purpose v2 and blob storage accounts. Use the policy to transition your data to the expire at the end of the data's lifecycle. Learn more

Name	Status
regleStan	Enabled

Edit a rule

Action set | Filter set | Review + save

Each rule definition includes an action set and a filter set. The action set applies the tier or delete actions to the filtered set of objects. The filter set limits rule actions to a certain set of objects within a container or objects names.

Rule name * regleStan

Status ○ Disabled ◉ Enabled

Blobs

☑ Move blob to cool storage

Days after last modification 1

☑ Move blob to archive storage

Days after last modification 3

☑ Delete blob

Days after last modification 2555

Review + save | < Previous | Next : Filter set >

Edit a rule

Action set | Filter set | Review + save

Each rule definition includes an action set and a filter set. The action set applies the tier or delete actions to the filtered set of objects. The filter set limits rule actions to a certain set of objects within a container or objects names.

Supported blob types Block Blob

Prefix match

Apply a rule to a container or a subset of virtual folders with the use of up to 10 prefixes as filters. By default, a rule will apply to the entire storage account.

Browse | Delete

Path
mesfichiers/

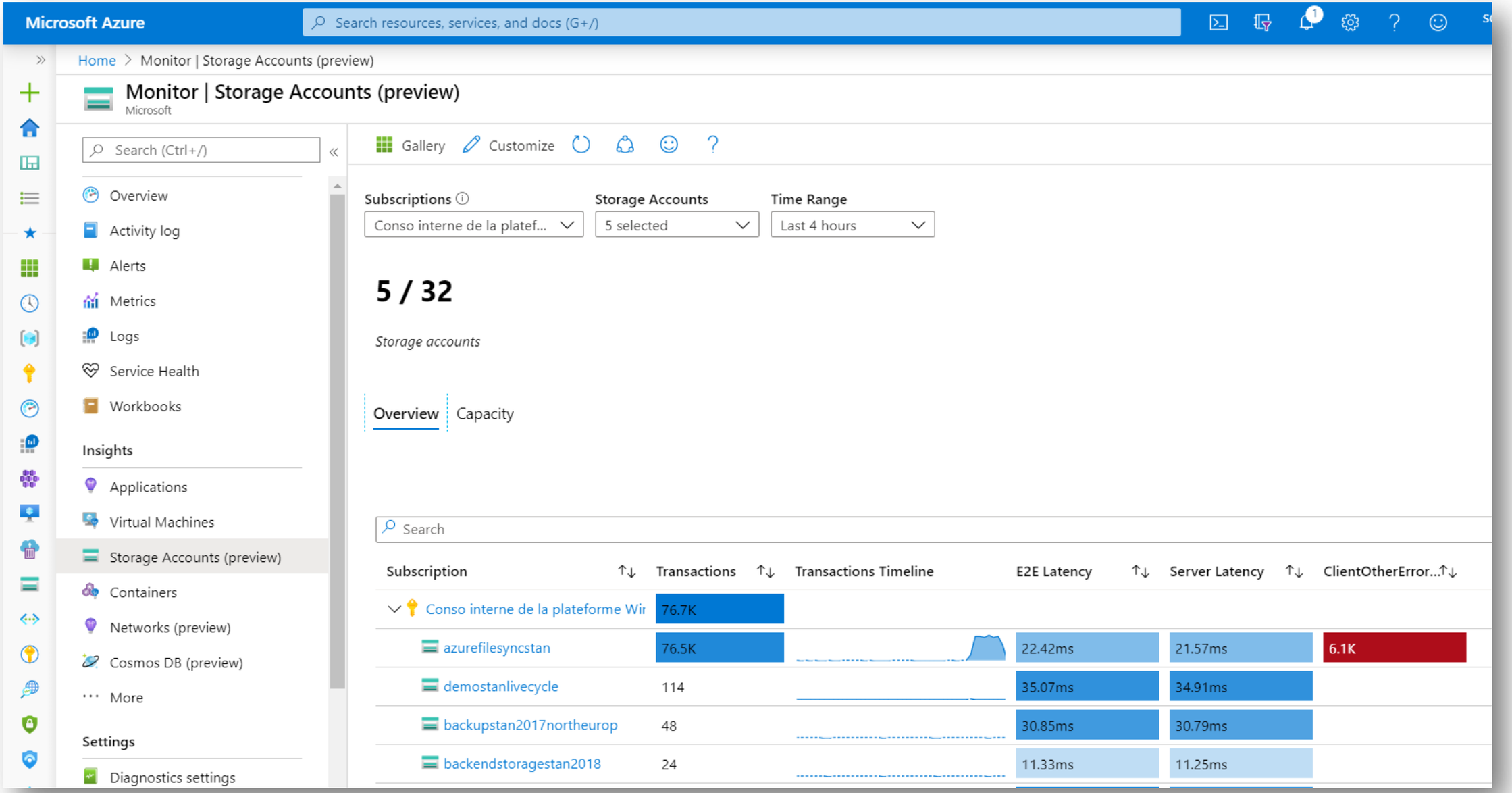
Container/virtualfolder

Review + save | < Previous | Next : Review + save >

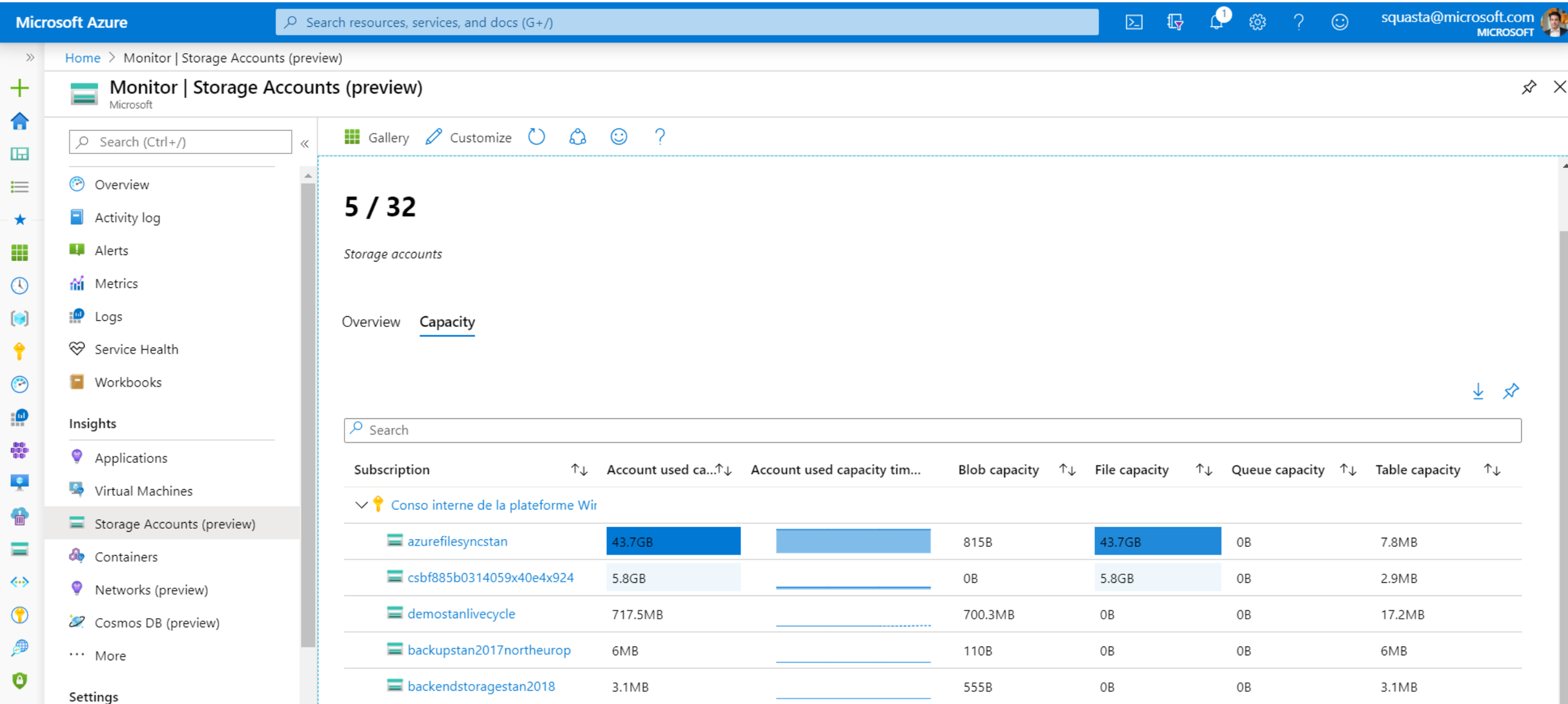
Compte de stockage Azure & optimisation

- Choisir le bon type de réplication : LRS, ZRS, GRS, RA-GRS, GZRS, RA-GZRS
  - Les versions géo-répliquées coûtent deux fois plus chers que les réplications locales sur une seule région Azure (6 copies versus 3)
- Choisir le bon tiers de performances
  - Plus c'est performant et avec une faible latence plus c'est cher et consommateur de ressources
  - Pour la partie stockage Blob, utiliser les stratégies de Livecycle Management pour déplacer les données vers des tiers plus froids (Cool Storage & Archive Storage) et moins chers ou pour supprimer les données au bout d'un certain temps

Surveiller les comptes de stockage



Surveiller les comptes de stockage



Log Analytics Workspace & optimisation

- L'ingestion et le stockage des logs dans un workspace Log Analytics peut devenir coûteux si mal suivi
- La facturation et l'usage se décompose en 2 parties :
  - L'ingestion (et le traitement) des journaux et métriques
  - Le stockage des données (gratuit dans les 90 premiers jours)
- La surveillance de la volumétrie ingérée et des coûts associés est disponible directement dans la console Azure Monitor Logs

Surveiller les coûts d'ingestion des logs

Dashboard > Log Analytics workspaces > WorkspaceBry2017 - Usage and estimated costs

WorkspaceBry2017 - Usage and estimated costs
Log Analytics workspace

Search (Ctrl+/)

- Advanced settings

General

- Quick Start
- Workspace summary
- View Designer
- Workbooks (Preview)
- Logs
- Solutions
- Saved searches
- Pricing tier
- Usage and estimated costs
- Properties
- Service Map

Workspace Data Sources

- Virtual machines
- Storage accounts logs

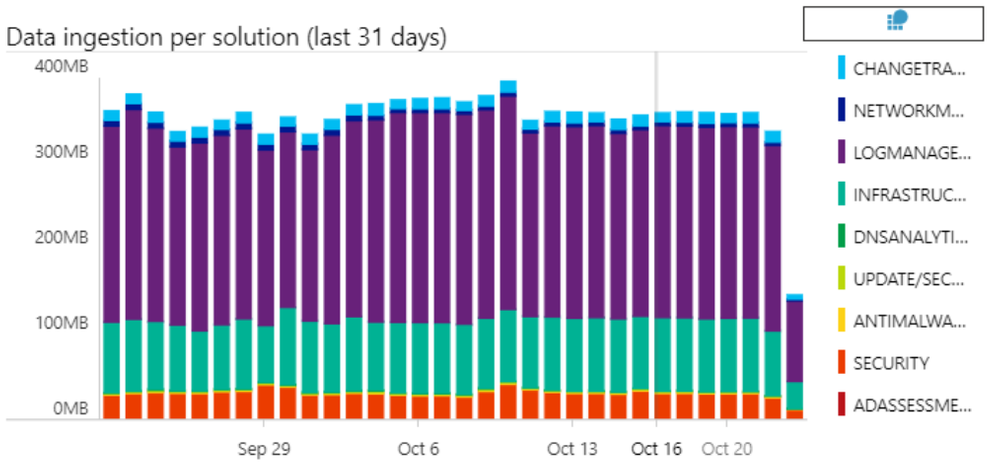
Usage details | Daily cap | Data Retention | Help

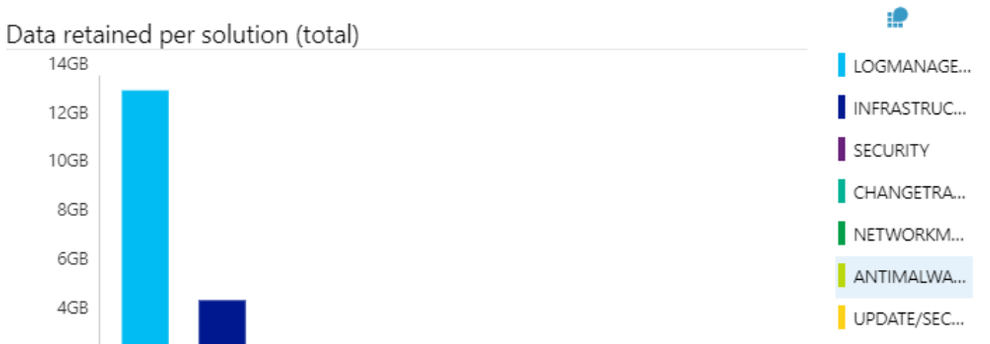
Pricing tier: Per GB (2018)

The table below shows estimated monthly costs* for this Log Analytics resource based on the last month's usage.

Item type	Price	Monthly usage (last 31 days)	Estimated monthly cost
Log data ingestion	€2.52	11.28 GB	€28.44
Log data retention	€0.11	0.00 GB	€0.00
			€28.44

* Estimates do not include taxes which may be applied to this subscription. It doesn't reflect Security nodes charges and included data volume in this blade.



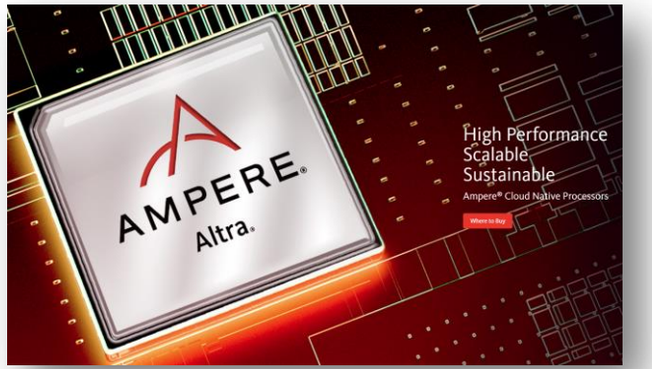


Optimisation des coûts de logs dans Azure

- N'activer les logs pour des services que si les informations sont collectées dans un but de supervision ou d'analyse
  - Journaliser et stocker pour rien n'a pas de sens
- Configurer les agents de collectes ou les paramétrages des diagnostiques des services PaaS
  - Ne journaliser que ce qui est nécessaire

Azure VM avec processeurs Arm64

- Nouveaux gabarits de VM pour usage généraliste ou optimisé mémoire avec des processeurs ARM Ampere Altra
  - Série Dpsv5 pour general-purpose
  - Série Epsv5 pour memory-optimized



- Meilleur rapport prix/puissance comparativement aux VM Azure Linux équivalentes sur base x64
- Consommation réduite par rapport aux CPU x64

Azure VM avec processeurs Arm64

Series	vCPUs	Memory (GiBs)	Local Disk (GiBs)	Max Data Disks	Max NICs
Dpsv5-series	2 – 64	8 – 208	n/a	4 – 32	2 – 8
Dpdsv5-series	2 – 64	8 – 208	75 – 2,400	4 – 32	2 – 8
Dplsv5-series	2 – 64	4 – 128	n/a	4 – 32	2 – 8
Dpldsv5-series	2 – 64	4 – 128	75 – 2,400	4 – 32	2 – 8
Epsv5-series	2 – 32	16 – 208	n/a	4 – 32	2 – 8
Epdsv5-series	2 – 32	16 – 208	75 – 2,400	4 – 32	2 – 8

Microsoft Azure (Preview) | Report a bug | Search resources, services, and docs (G+/)

Home > TDF-Juin2022 > Create a resource >

Marketplace

...

- Get Started
- Service Providers

Management

- Private Marketplace
- Private Offer Management

My Marketplace

- Favorites
- Recently created
- Private products

Categories

- Internet of Things (15)

arm64

Pricing : **All**

Operating System

Azure benefit eligible only

Showing 1 to 18 of 18 results for '**arm64**'. Clear search

Microsoft Windows 11 Preview arm64

Microsoft

Virtual Machine

Microsoft Windows 11 Preview arm64

Create

Ubuntu Server 20.04 LTS ARM64 Preview

Canonical

Virtual Machine

Linux For The Cloud

Create

Ubuntu Server 18.04 LTS ARM64 preview

Canonical

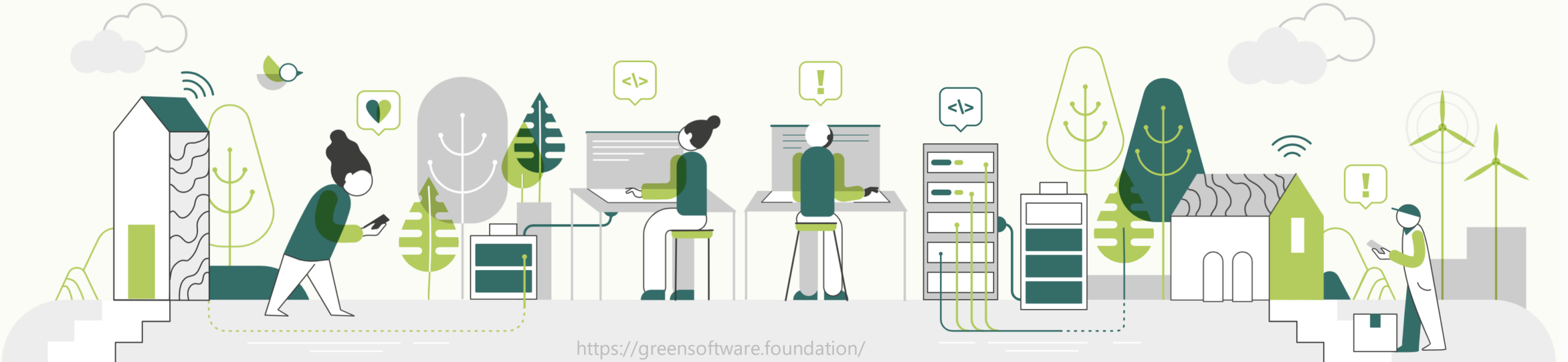
Virtual Machine

Linux For The Cloud

Create

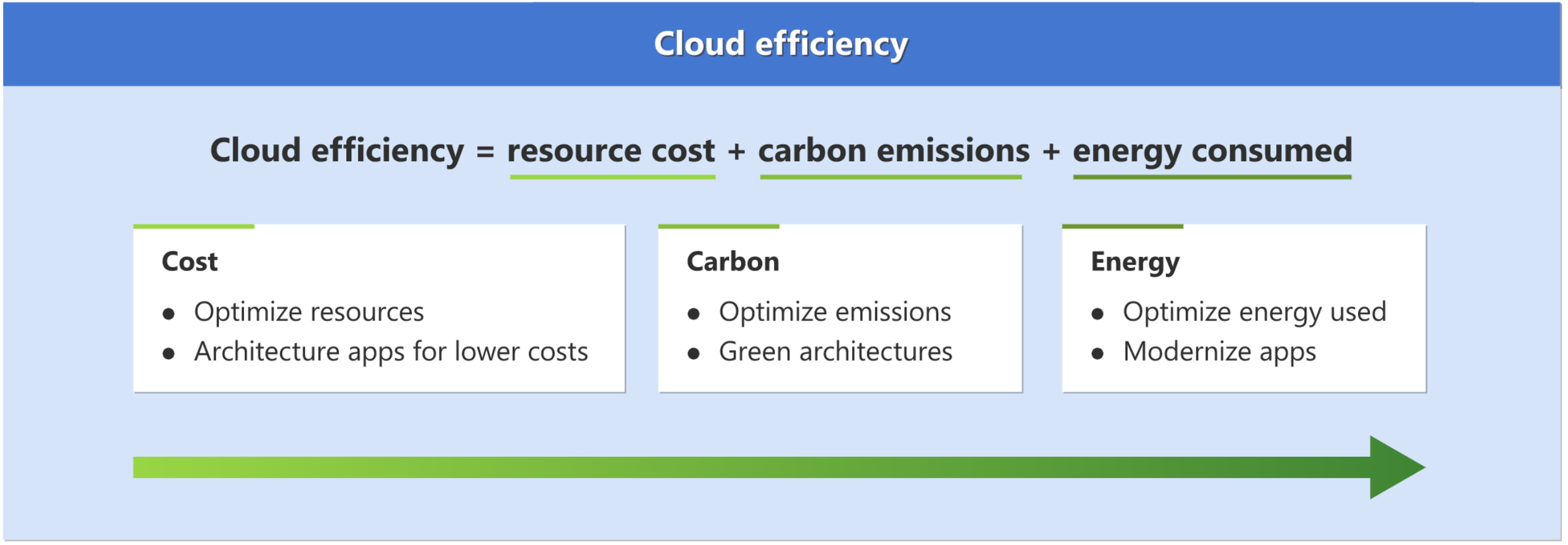
OS supportés : Canonical Ubuntu Linux, CentOS et Windows 11 Professional and Enterprise Edition on Arm

Le développement de logiciels éco-responsables



https://greensoftware.foundation/

Efficacité du ‘Cloud’



Je ne sais pas par où commencer ….

- L'éco-responsabilité consiste à limiter la consommation d'énergie, de ressources, et restreindre le gaspillage.

- Le cloud est un modèle où l'on est **facturé à la consommation** de ressources (calcul, stockage, bande passante) au prorata-temporis

- Quand je réduis ma facture …
  → C'est parce que je consomme moins de ressources
  → Donc je deviens un peu plus éco-responsable ☺

➔ Entamer une démarche de type 'FinOps' est donc pertinente.

Mais ca ne fait pas tout !

Les principes de l'ingénierie logicielle Eco-Responsable

Principles.Green

8 principes :

- Minimiser l'empreinte carbone (ou d'équivalent carbone)
- Efficacité électrique
- Utiliser l'énergie la moins carbonée
  - Localisation de l'exécution de vos applications
- Exploiter le matériel de manière efficace et pérenne :
  - Limiter les prérequis matériels de vos applications
  - Supporter des matériels anciens
- Réduire la quantité de données transitant sur le réseau
- Mesurer, Optimiser de bout en bout
- Orienter la demande vers l'offre plutôt que d'adapter l'offre à la demande

https://principles.green/fr-fr/
https://learn.microsoft.com/en-us/training/modules/sustainable-software-engineering-overview

Intensité carbone : petit rappel de physique

Intensité Carbone

= Quantité de gaz à Effet de Serre émis exprimé en équivalent carbone

= Energie * ($eqCO^2$ par unité d'énergie)

= Puissance * Durée * (eqCo2 par unité d'énergie)
exprimé en $eqCO^2$/kWh

➔ Pour baisser l'intensité carbone, on peut donc intervenir sur

- La puissance consommée (valeur instantanée)
- La durée de fonctionnement d'un système (temps)
- Le mix énergétique : l'endroit ou le moment où fonctionne le système

C’est quoi une application

Réseau
Volume de données
Représentation des données
Nb d’appel
Distance parcourue

Interface utilisateur

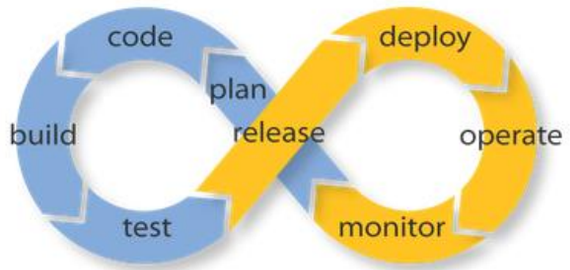
Services fonctionnels

Données

Apparence & Ergonomie
Traitements locaux
Stockage local

Traitements
Choix des services gérés
Utilisation de service géré
Montée en charge
Le code, le code, le code
Architecture applicative

Stockage
Volume, format de données
Choix du Tiers
Traitements
Performance et optimisation des requêtes
Volume de données parcouru



Réseau et communication

- Etude du « Shift Project » de 2019 pour évaluer l'impact environnemental du transfert de données
  - Energie utilisée (kWh) = volume de données en MBs * 0,0023
  - Intensité carbone (Kg eqCO^2) = (volume de données en MBs * 0,0023) * **0,519**
    Sur la base d'une évaluation moyenne mondiale de 0,519 kgCO^2eq/kWh

- Sur réseau mobile, l'énergie consommée est doublée, donc l'impact carbone aussi.

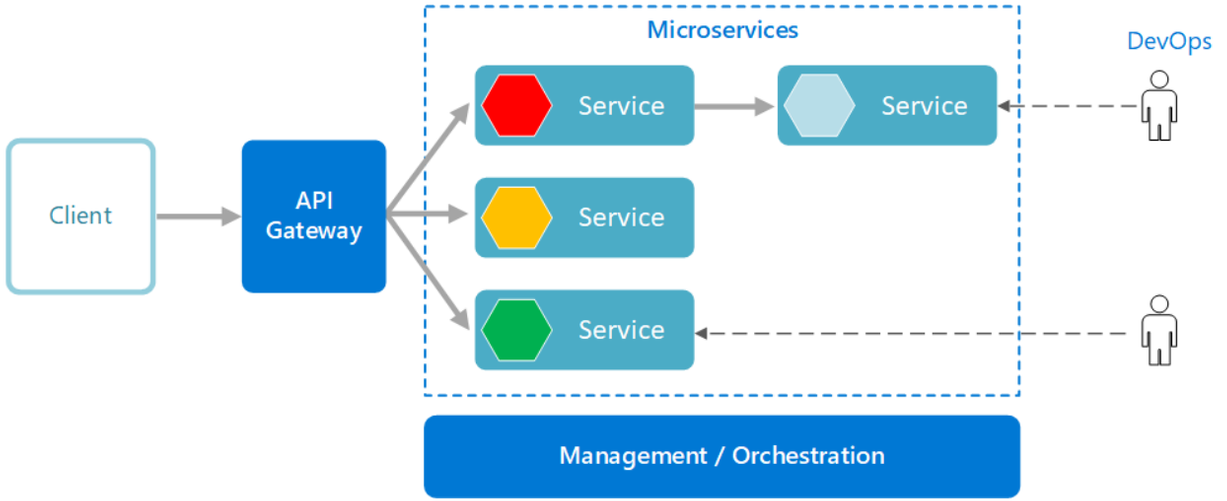
https://theshiftproject.org/en/article/unsustainable-use-online-video/
https://theshiftproject.org/wp-content/uploads/2020/06/Video-Materials-1byte-Model-applied-to-video-traffic-updated-2020.xlsx

Architecture applicative

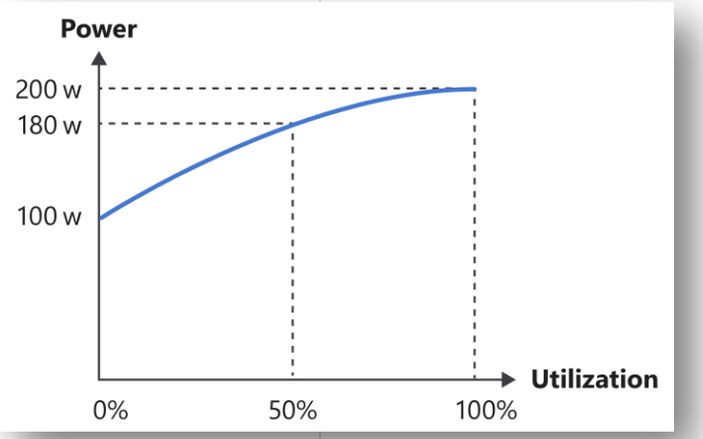
- Certains choix d'architecture vont permettre d'optimiser la consommation de ressources :
  - Microservices Stateless : permet une forte montée en charge mais surtout une élasticité de la plateforme :
  - Event Driven : piloté par des messages ou des évènements
  - Serverless : allocation unitaire à la demande de ressources pour le temps et le besoin d'un traitement

➔ Architecture Cloud Native & Cloud Patterns

Micro-Service

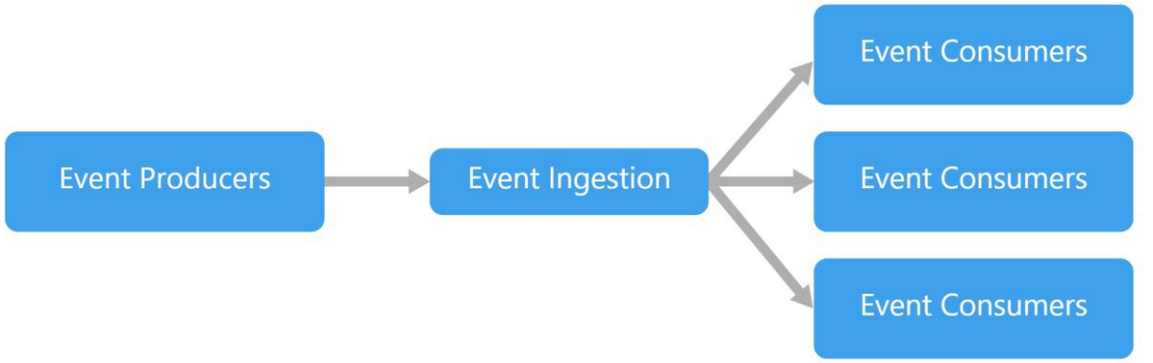


- Densité de déploiement pour une meilleure utilisation des ressources
- Montée en charge



https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/microservices

Event-Driven



- Consommation de ressources au plus juste
- Pas de gaspillage lié à de l'attente active

https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/event-driven

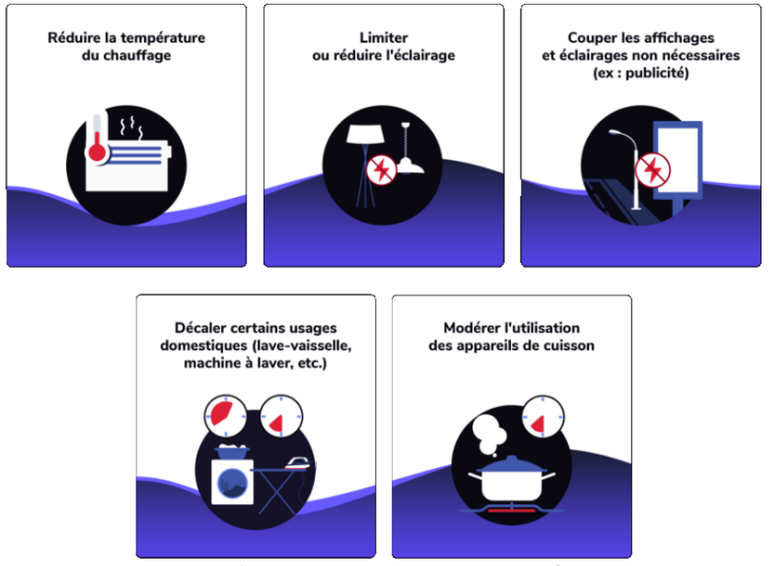
Modèle commercial SaaS

Création d'offres commerciales 'éco-responsable' dans un contexte « **S**oftware **a**s **a** **S**ervice » :

- SLA assoupli lors de période de tension énergétique
- Règles de déclenchement spécifiques pour les traitements planifiés
- …

➔Eviter les périodes de tensions énergétiques pour réduire l'impact environnemental des applications

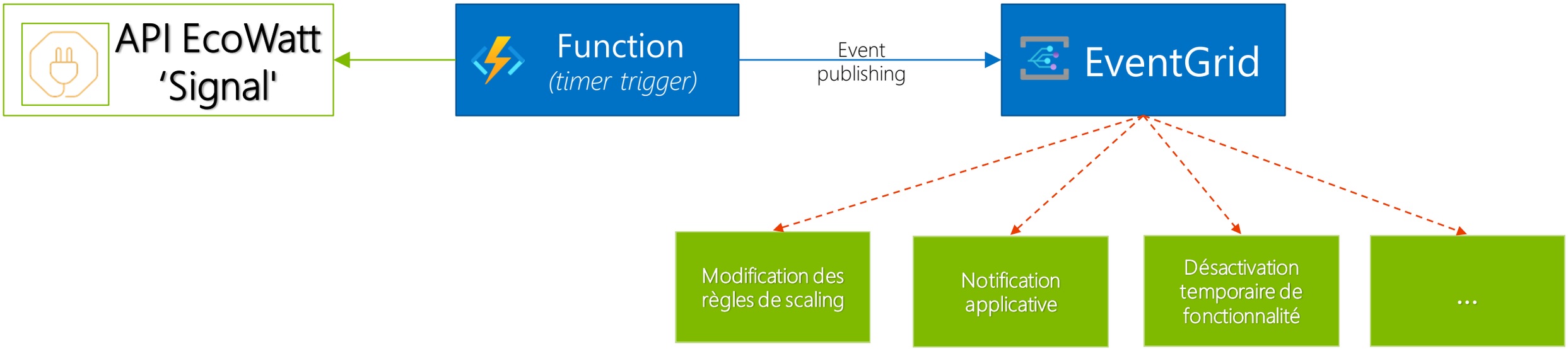
- APIs Ecowatt de RTE (alerte , qualité de l'électricité, consommation française temp réel, historique, mix énergétique, …)
  https://data.rte-france.com/catalog/-/api/consumption/Ecowatt/v4.0



https://www.monecowatt.fr/

Réagir dynamiquement

- Notification intelligente au sein d'un système applicatif

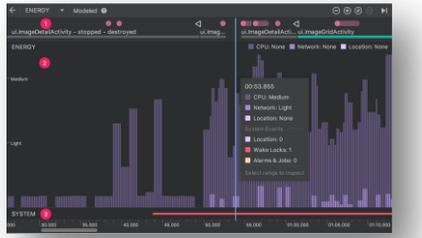


Application : UI

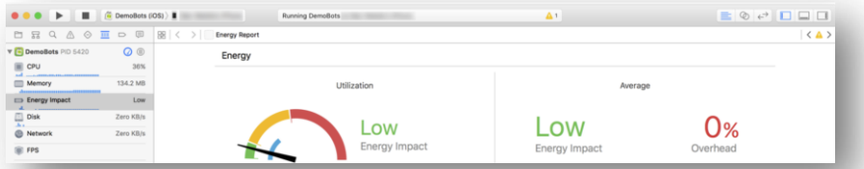
- Pas de fioriture inutile (ou option)
- Supporter les fonctionnalités d'économie d'énergie des systèmes d'exploitation ou des navigateurs (mise en veille, endormissement, …)
- Limiter les 'améliorations' de type « saisie assistées », ou pré-chargement de données
- Ergonomie des applications (navigation)
- Médias : Format, compression et chargement différé

Profiling énergétique : Application native

- Energy profile avec AndroidStudio : https://developer.android.com/studio/profile/energy-profiler



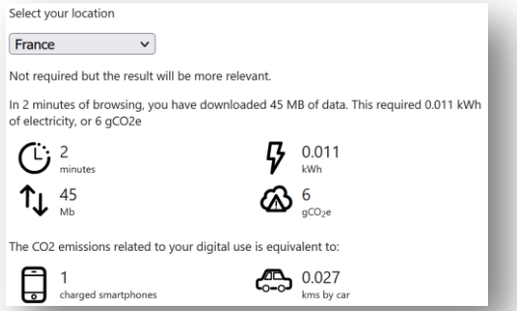
- Apple Xcode : https://developer.apple.com/library/archive/documentation/Performance/Conceptual/EnergyGuide-iOS/MonitorEnergyWithXcode.html

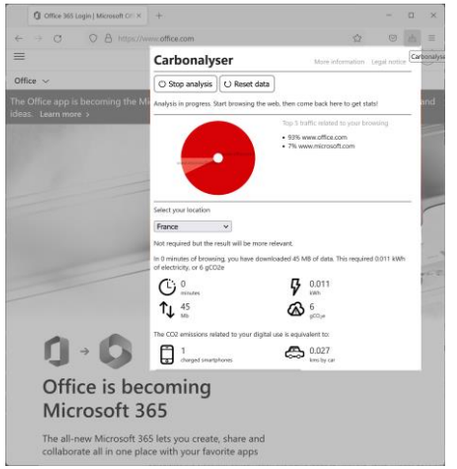


- Windows EcoQoS https://devblogs.microsoft.com/sustainable-software/introducing-ecoqos/
- Windows Energy Estimation Engine (WE3) https://devblogs.microsoft.com/sustainable-software/measuring-your-application-power-and-carbon-impact-part-1/

Profiling énergétique : Web

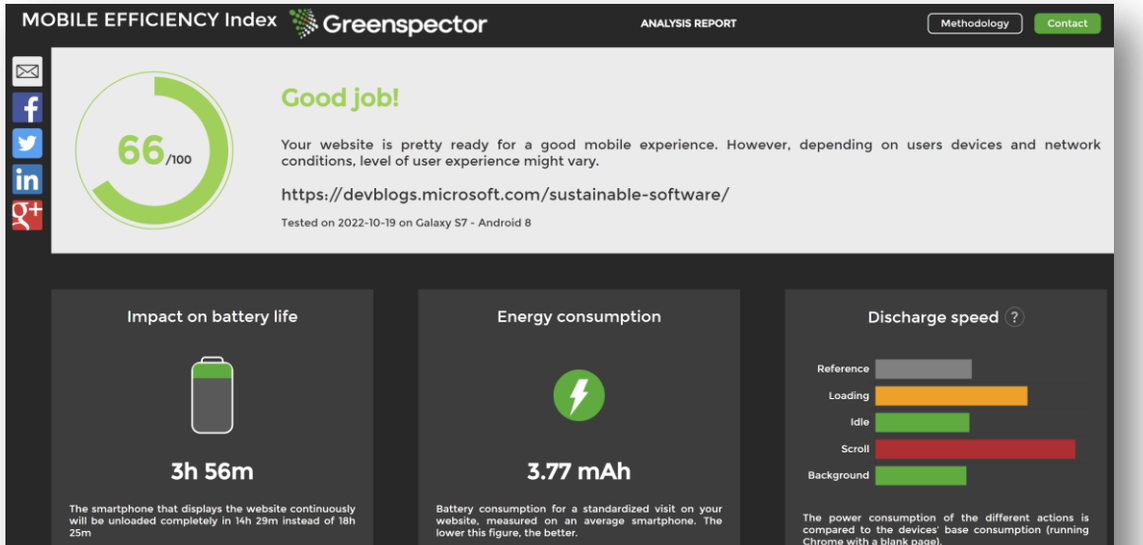
- Extension pour les navigateurs Web :
  - « **Carbonalyser** » du Shift Project
    https://theshiftproject.org/en/carbonalyser-browser-extension/





  - *Lighthouse* (OSS, extension Chrome, CLI, webui) :
    https://developer.chrome.com/docs/lighthouse/overview/

- Site Web :
  - GreenSpector : https://mobile-efficiency-index.com/en/



Services

- Choix du service d'Hébergement
  - Kubernetes Service, App Services, Container App, Container Instances, Static WebApp, …
- Exposer des apis optimisées
  - Filtrage des données inutiles/souhaitée, navigation paginée,…
- Optimisation des communications
  - Compression
  - HTTP/SOAP/XML vs. HTTP/REST/JSON vs. GRPC/Protobuf
- Mise en cache de courte de durée des données

Le langage de développement

- Apprendre et comprendre le fonctionnement de la technologie choisie
  - Ne soyez pas un simple assembleur de paquets (npm, nuget, …)
- Adopter une programmation asynchrone qui ne bloque pas inutilement des ressources par des entrées/sorties en attente
- Intégrer l'efficacité énergétique dans le choix d'un langage ou d'une technologie de développement

Efficacité énergétique des langages

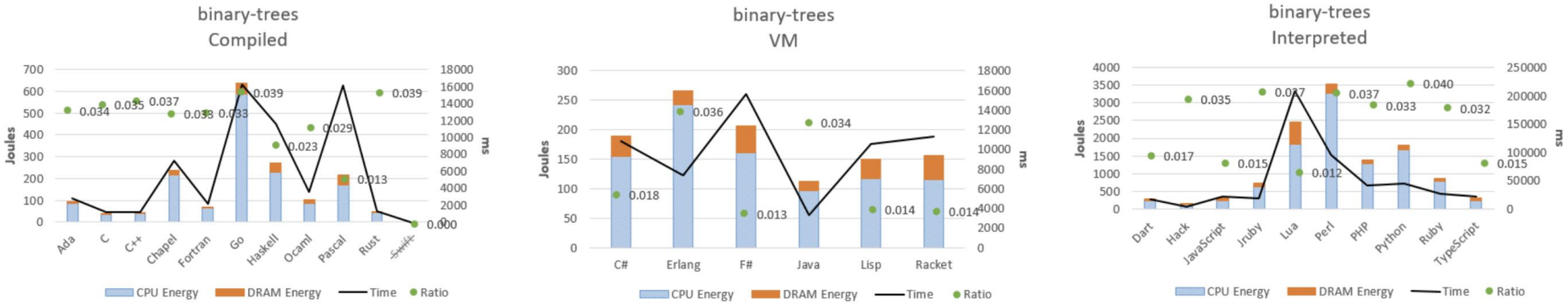


Figure 1. Energy and time graphical data for binary-trees

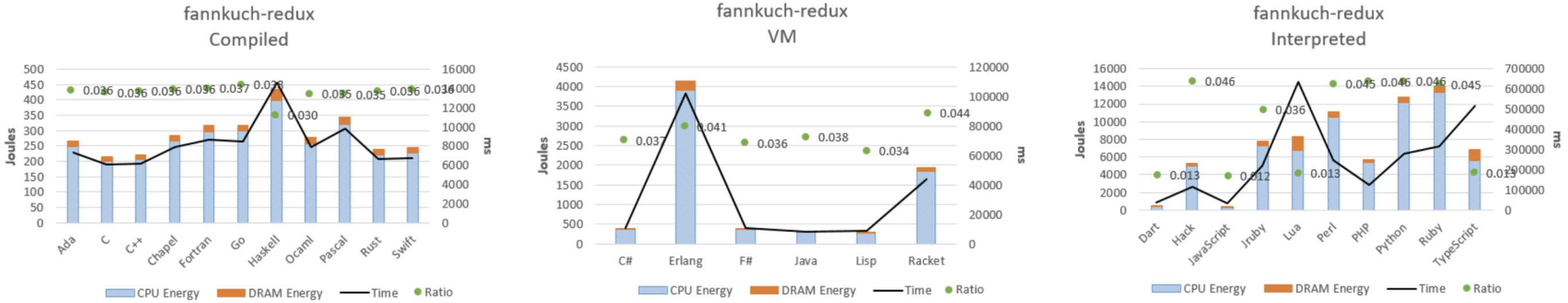


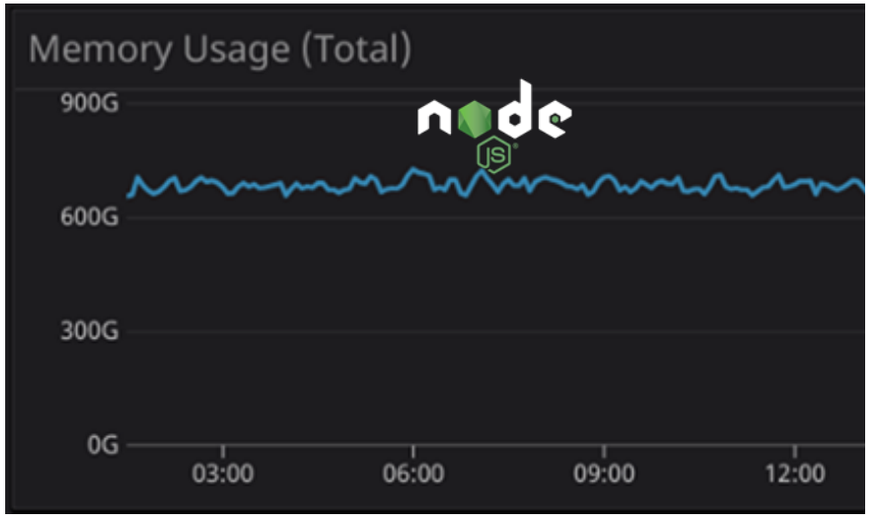
Figure 2. Energy and time graphical data for fannkuch-redux

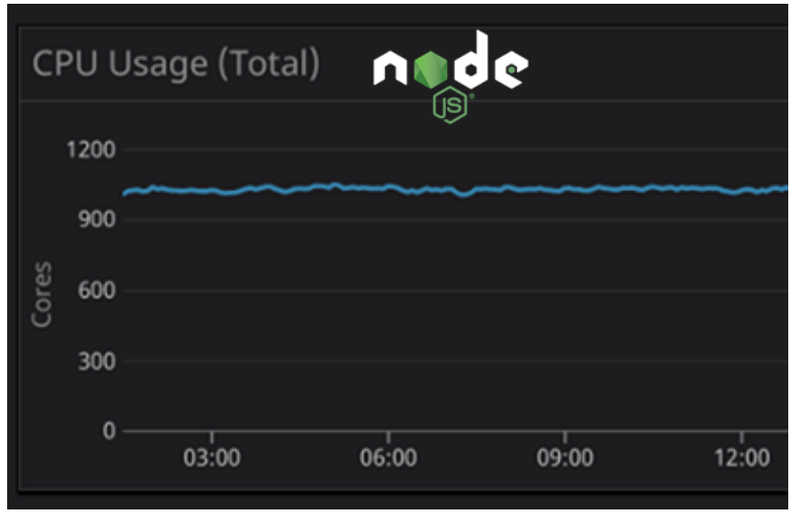
	Energy
(c) C	1.00
(c) Rust	1.03
(c) C++	1.34
(c) Ada	1.70
(v) Java	1.98
(c) Pascal	2.14
(c) Chapel	2.18
(v) Lisp	2.27
(c) Ocaml	2.40
(c) Fortran	2.52
(c) Swift	2.79
(c) Haskell	3.10
(v) C#	3.14
(c) Go	3.23
(i) Dart	3.83
(v) F#	4.13
(i) JavaScript	4.45
(v) Racket	7.91
(i) TypeScript	21.50
(i) Hack	24.02
(i) PHP	29.30
(v) Erlang	42.23
(i) Lua	45.98
(i) Jruby	46.54
(i) Ruby	69.91
(i) Python	75.88
(i) Perl	79.58

« Energy Efficiency accross Programming Languages », Pereira, Couto, Ribeiro, Cunha, Fernandes & Saraiva – oct 2017
https://greenlab.di.uminho.pt/wp-content/uploads/2017/10/sleFinal.pdf

Changer de langage : NodeJS → Rust

- Migration d'une partie de la solution de cyber Security **Tenable** de Javascript vers Rust

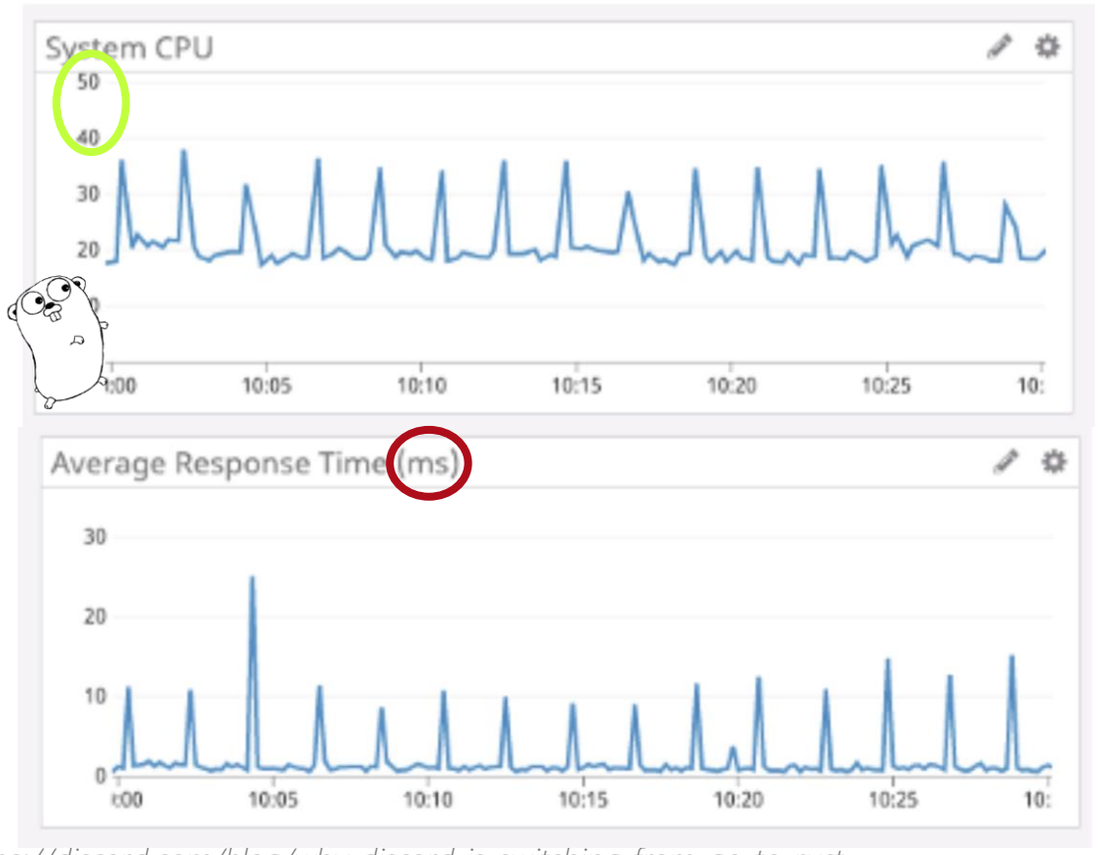


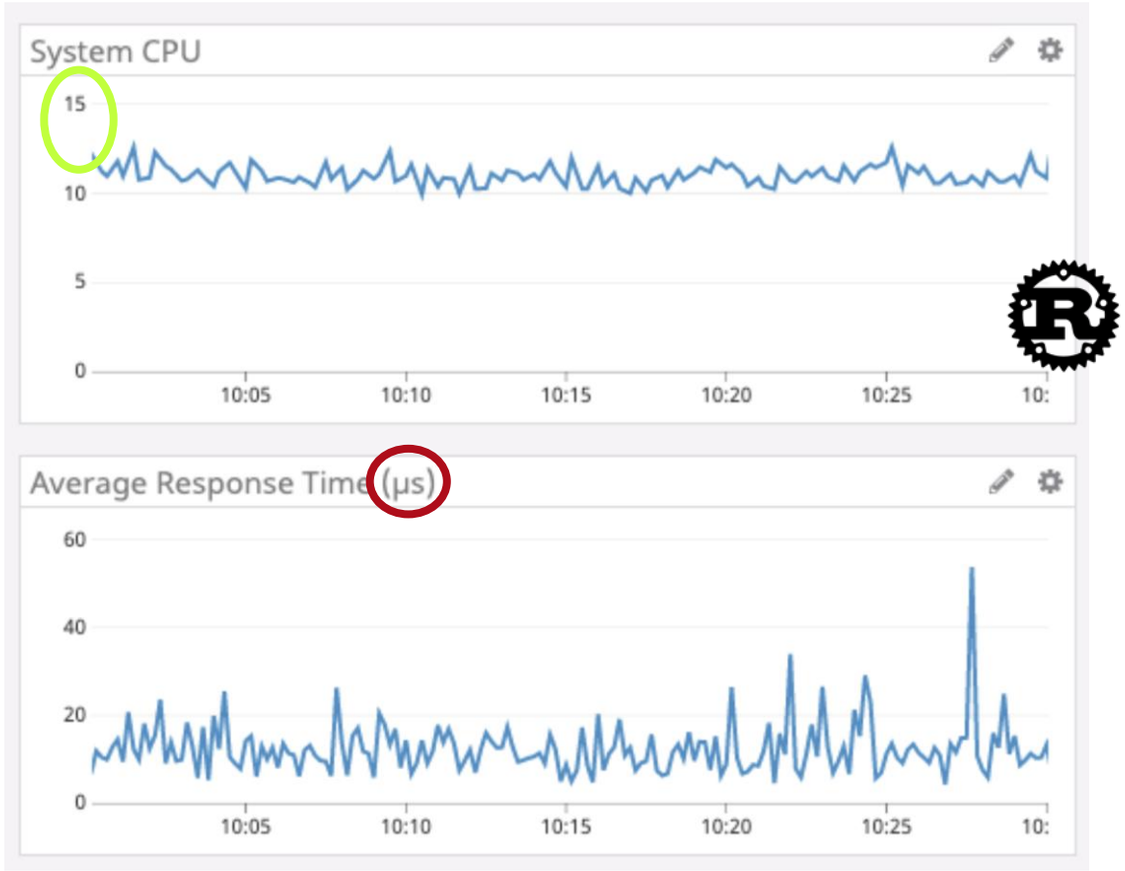


https://aws.amazon.com/blogs/opensource/sustainability-with-rust/

Changer de langage : GoLang → Rust

- **Discord** à choisi de basculer de GoLang vers Rust





https://discord.com/blog/why-discord-is-switching-from-go-to-rust

Performance du code

- Ne pas résoudre – systématiquement - un souci de performance par l'ajout de puissance de calcul
- Utilisez les outils de monitoring des performances (consommation mémoire, cpu, energy) pour identifier les traitements à optimiser
  - Analysez le problème, identifier l'origine, proposer une solution intégrant le moindre impact environnemental

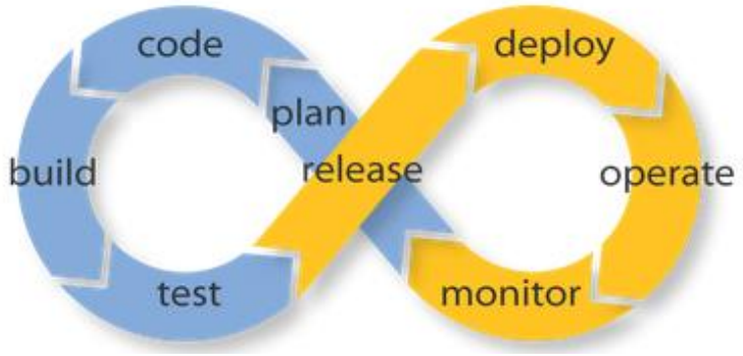
Choix de la base de données

- Privilégier les bases de données élastiques, voire serverless
  - Découplage de l'allocation de temps CPU vs. Capacité de stockage
- Quand cela est possible, dénormaliser les données pour alléger les accès et les faire correspondre aux scenarios d'utilisation
  - Modélisation 'parfaite' et 3e forme normale (3NF) vs. Pragmatisme et orientation usage
  - Sans schéma imposé
  - Base de données non-relationnelle ('NoSQL')
  - Attention aux impacts par rebond

Persistance des données

- Limiter le volume des données stockées et échangées
- Envisager dès la conception l'utilisation de différents tiers de stockage
  - Mettre en place du 'sharding' pour iloter les données par rapport à leur fréquence d'usage/obsolescence
  - Azure :
    - Storage : Tiers HOT – COOL – ARCHIVE
    - Disk Standard, SSD, SSD Premium
    - Database SKU : Basic, Premium, …

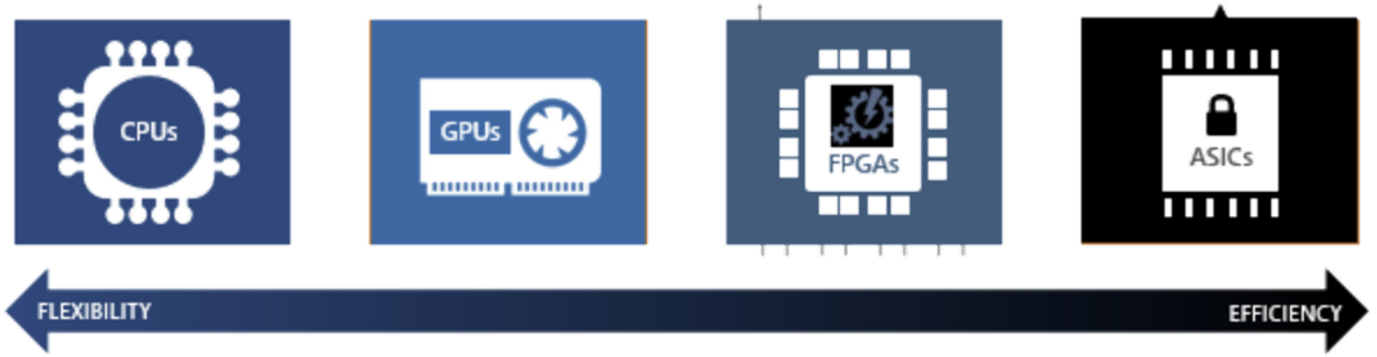
Impact des Processus de développement



- Privilégier des ressources de type (VM) Burst pour les phases de dev/test
- Impact de vos process DevOps
  - Allocation d'agent d'exécution dédié, à la demande ou mutualisé,
  - Fréquence et horaire d'exécution des pipelines d'intégration continue
  - Fréquence et durée des tests de charge/performance
  - Utilisation d'environnement provisionné à la demande plutôt qu'en attente
  - Sélectionner les périodes de temps « Low Carbon »

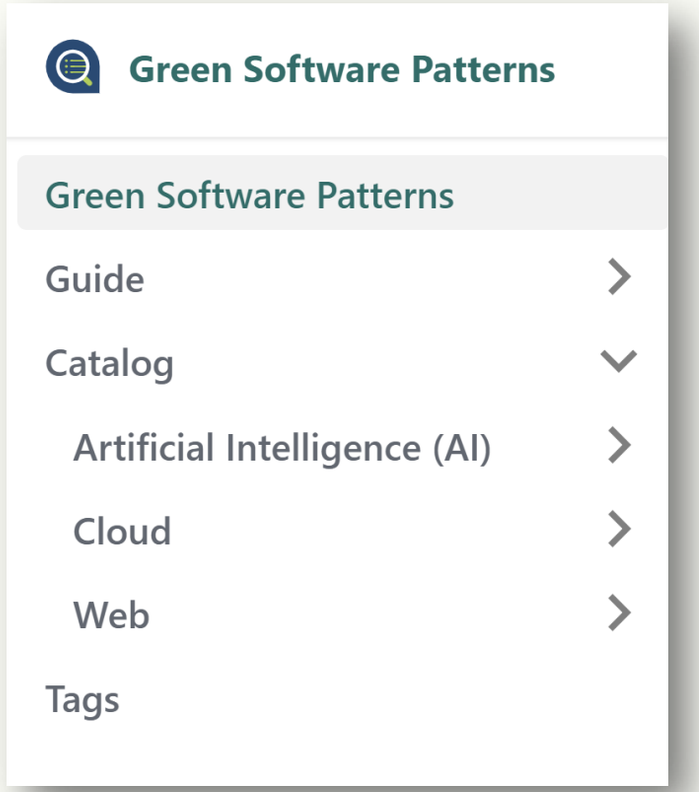
Optimisation par du hardware

Utiliser du matériel spécialisé et optimisé pour certaines taches



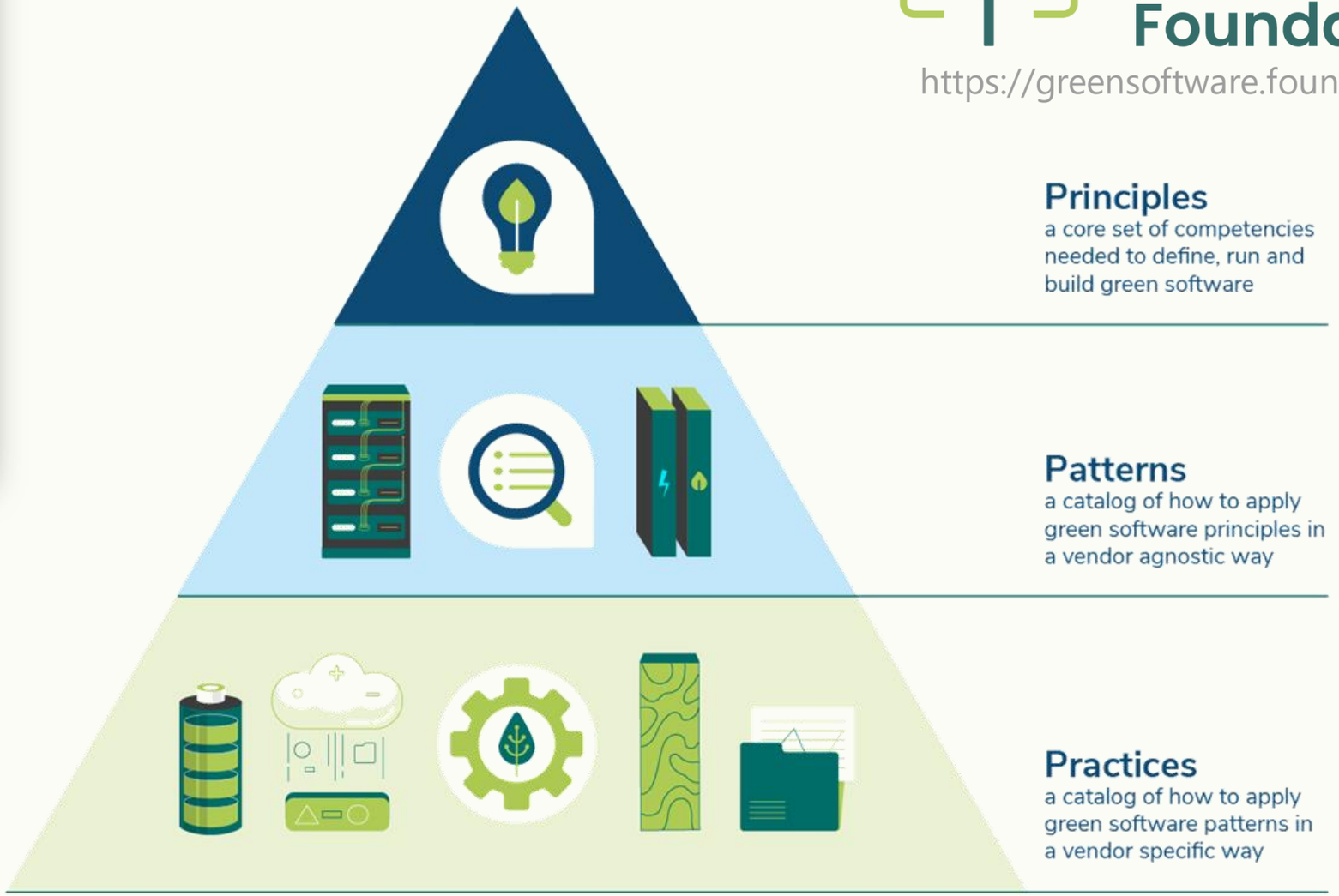
- Calcul parallèle massif, rendu 3D : GPU Nvidia
- Algorithme optimisé pour du Silicium (FPGA) :
  - VM NP-series avec FPGA Xilinx U250 (de 1 à 4)
    https://www.xilinx.com/products/boards-and-kits/alveo/cloud-solutions/microsoft-azure.html

  - *Azure Machine Learning Hardware Accelerated Models* sur FPGA Intel Arria 10
    https://learn.microsoft.com/en-us/azure/machine-learning/v1/how-to-deploy-fpga-web-service
    https://learn.microsoft.com/en-us/events/build-2018/brk3202



Green Software Foundation

https://greensoftware.foundation/



Principles

a core set of competencies needed to define, run and build green software

Patterns

a catalog of how to apply green software principles in a vendor agnostic way

Practices

a catalog of how to apply green software patterns in a vendor specific way

Green Software Foundation

greensoftware.org

Green AI

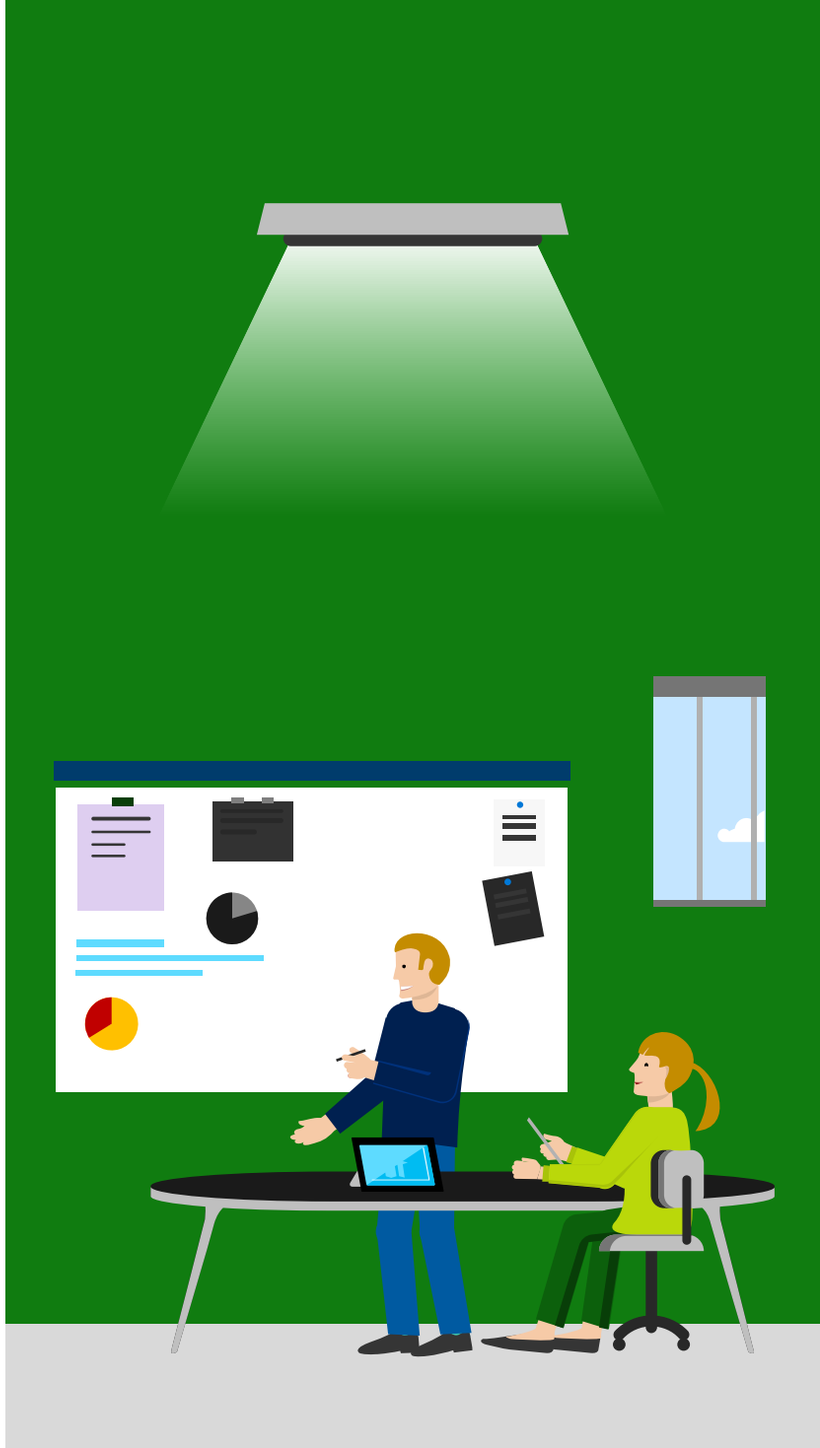
How to reduce the carbon footprint of ML workflows?

Agenda

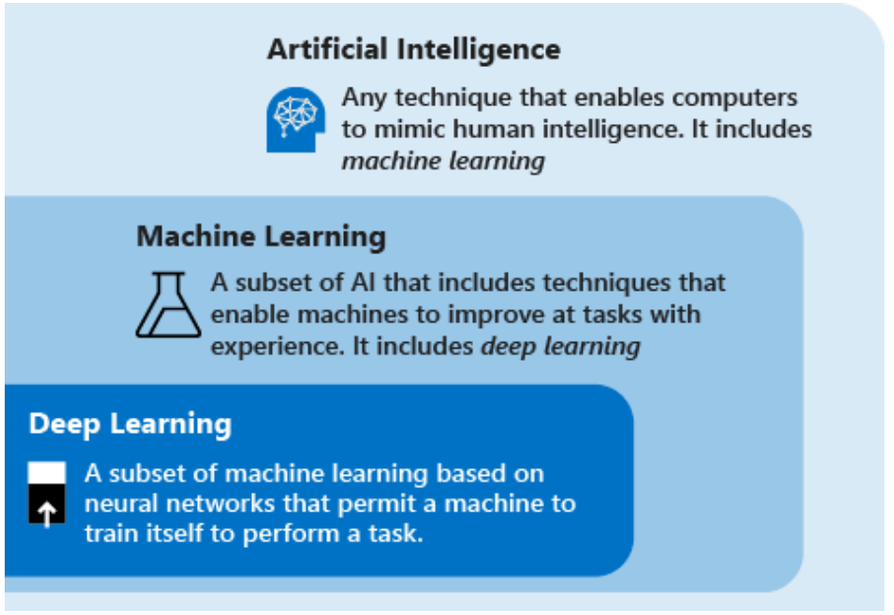
1. Why Green AI?

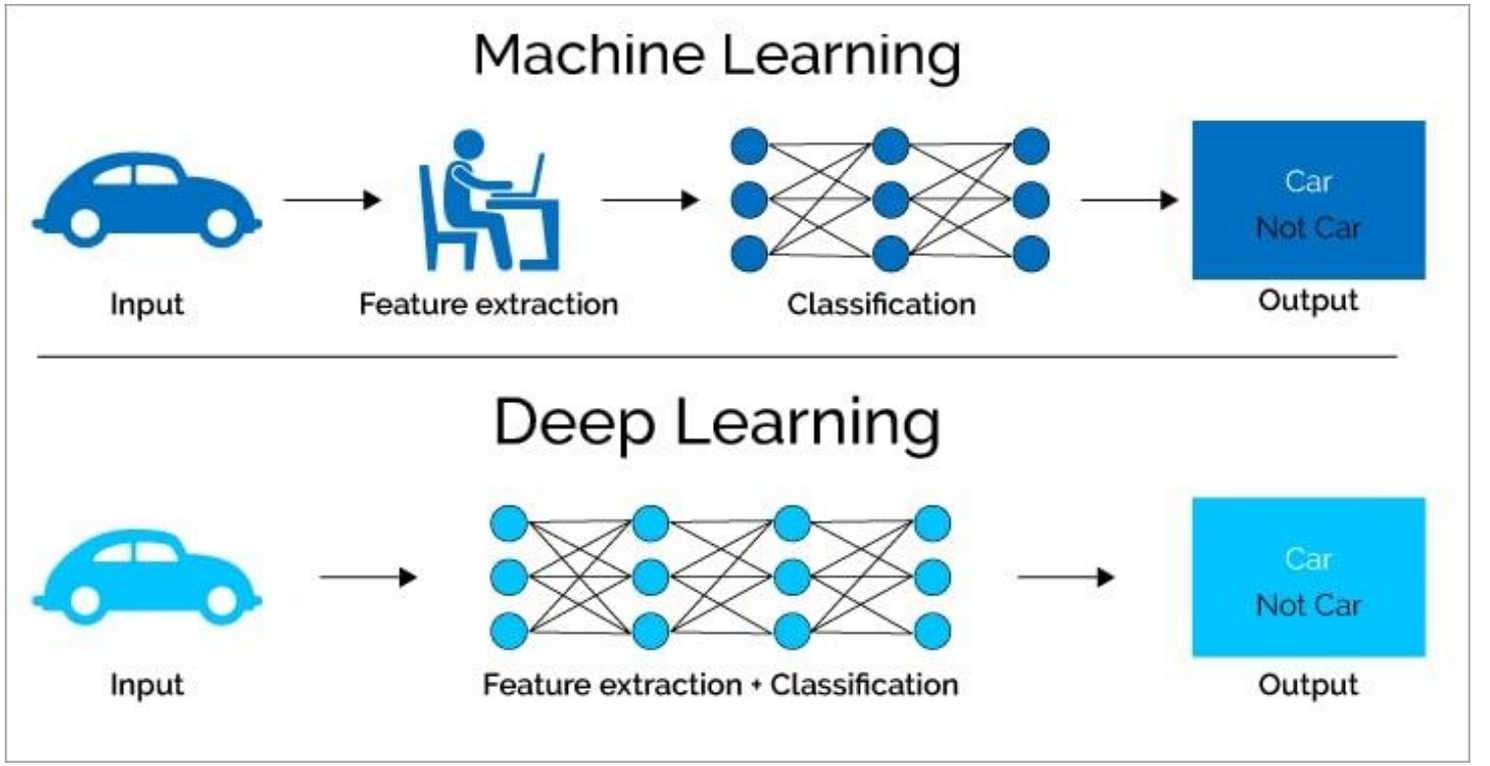
2. What is Green AI?

Understand – Measure - Reduce



Quick Overview

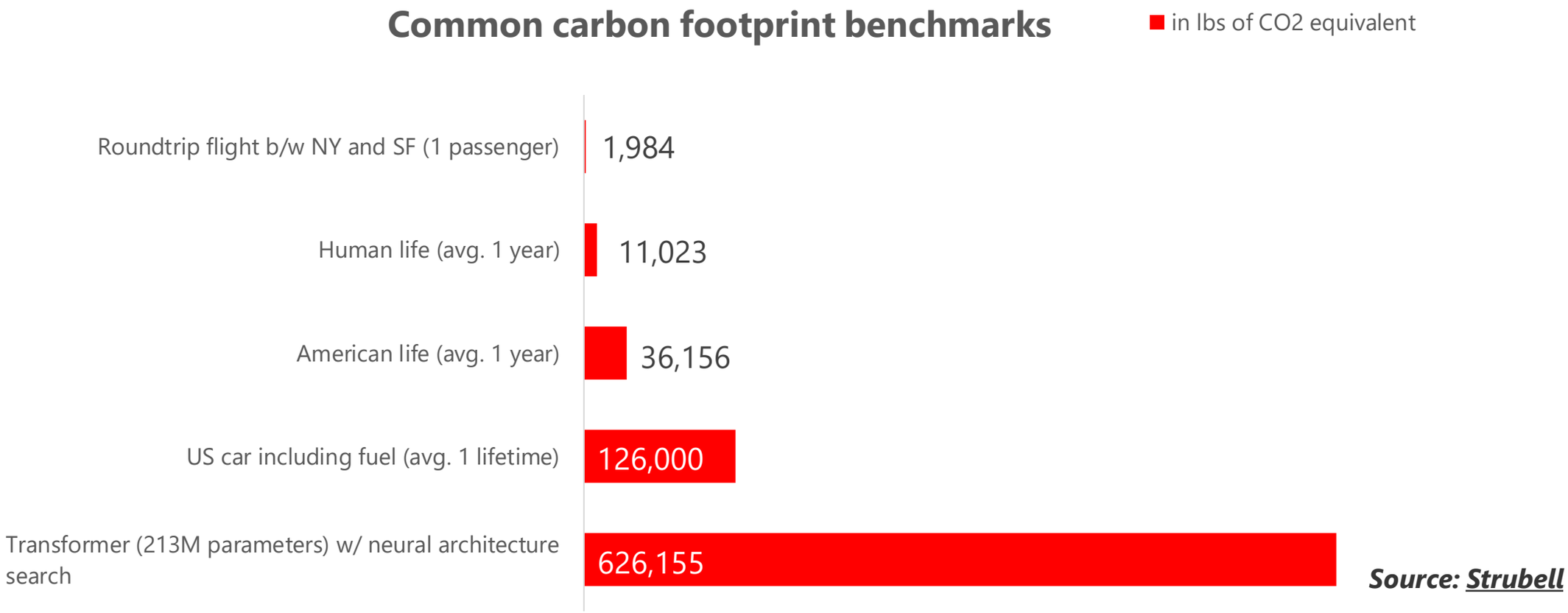




Why Green AI?

The Staggering Cost of AI

- Computational costs have increased **300.000X** from 2012 to 2018
- Only **11%** of firms are seeing a 'significant' ROI on their AI workloads(wired)
- GPT3 training emits as much carbon as **3X** round-trip transcontinental flights (SF<>NYC)

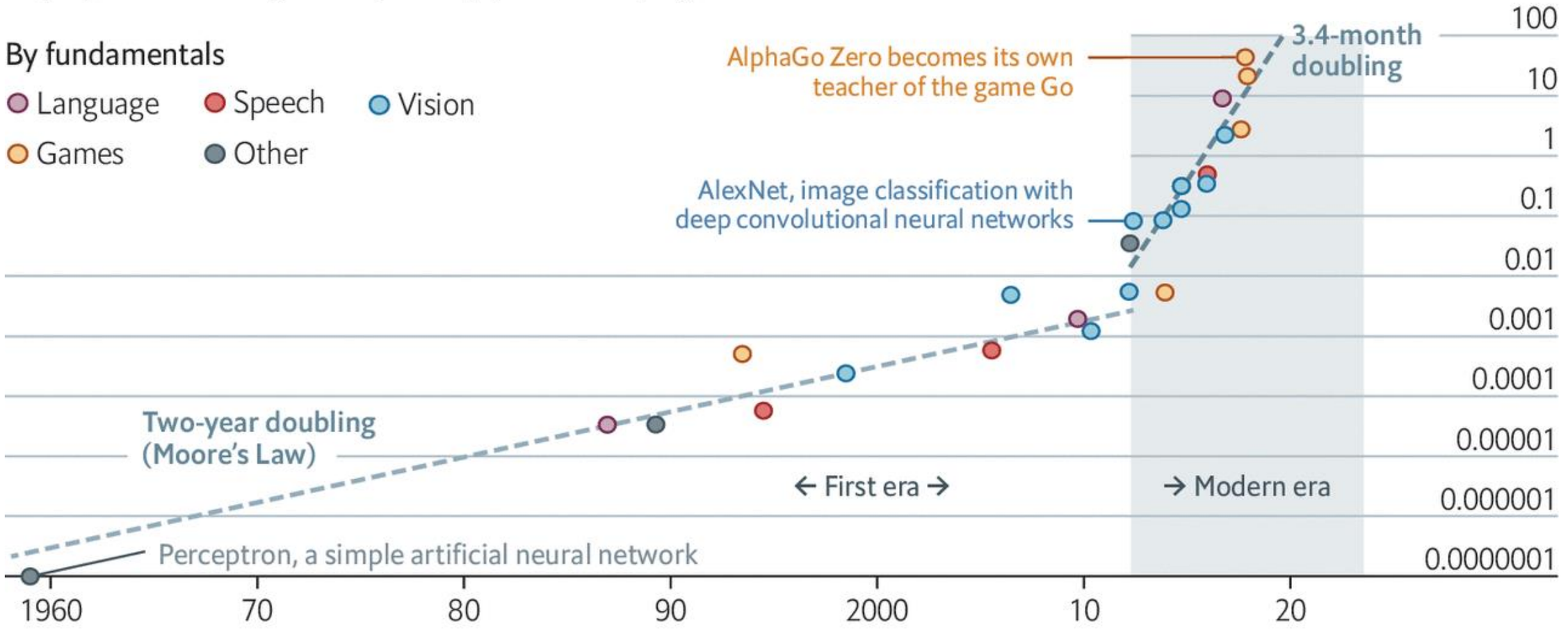


Source: Strubell

Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale

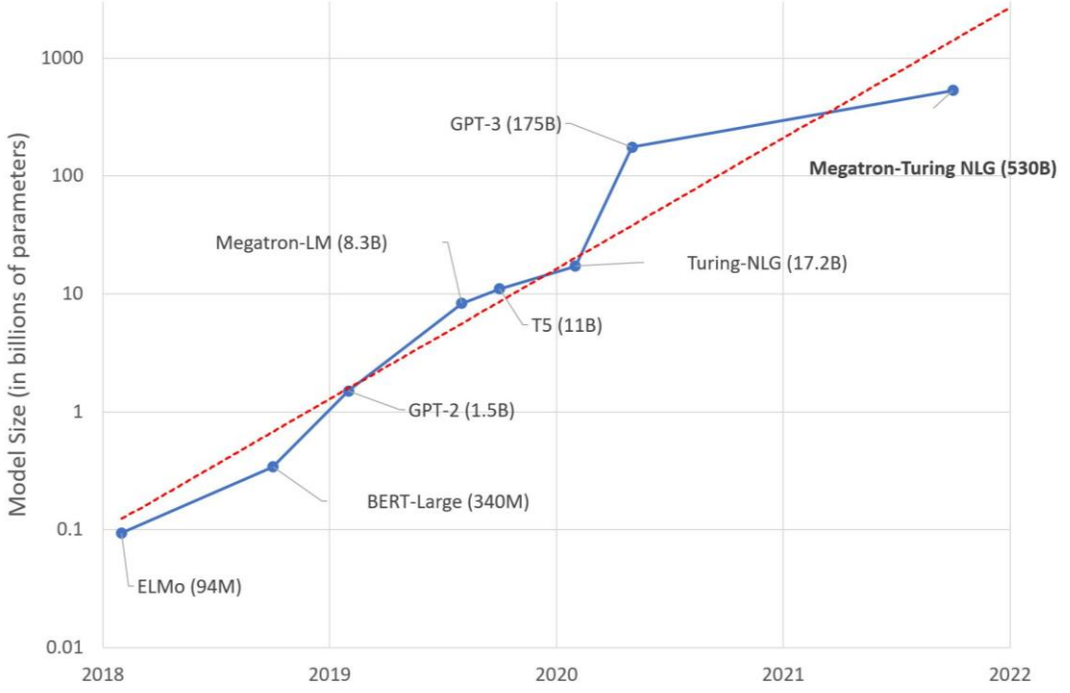


Source: OpenAI

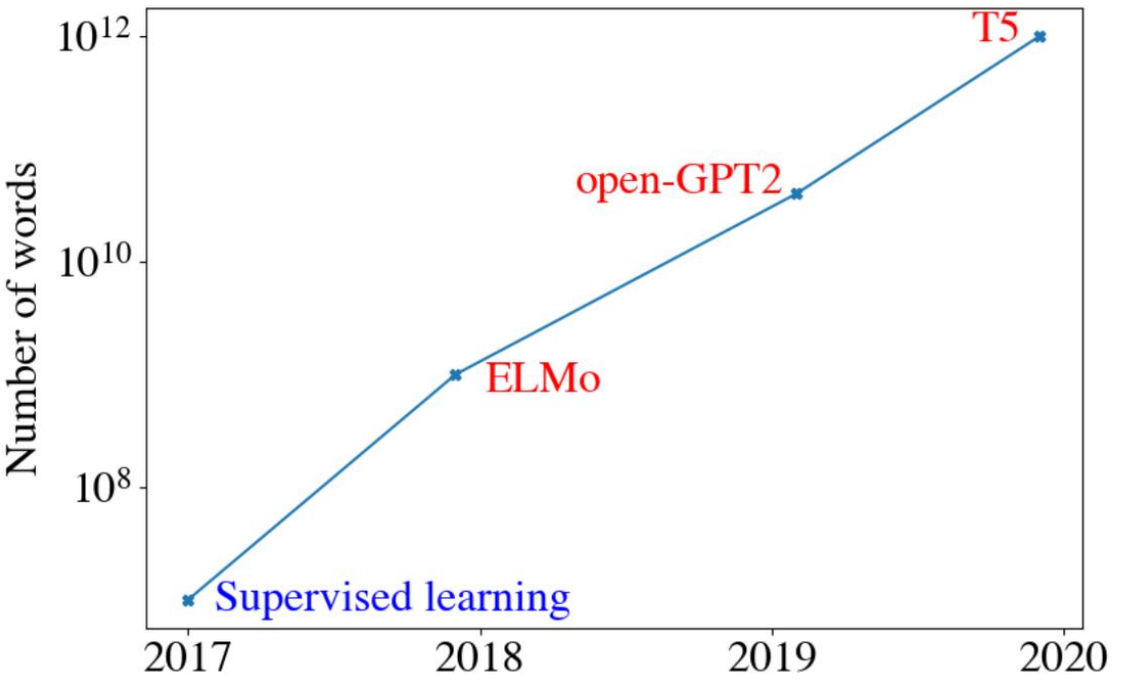
*1 petaflop=10^{15} calculations

The Economist

Bigger Models – Larger Datasets

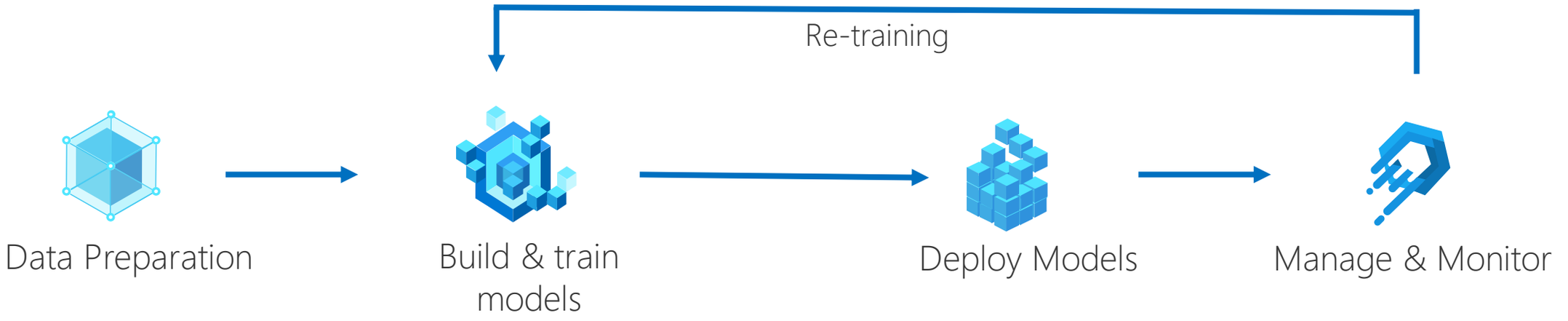


Parameter counts of several recently released pretrained language models



More Data 100000x in 3 Years!

AI and Power



Hardware and power needed:

DL training requires specialized GPU hardware
GPUs are power-hungry (often 250-350W)
Inferences may use GPUs, FPGAs or CPUs (typical CPU ~135W, some up to 280W)

Problems with Big Models

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

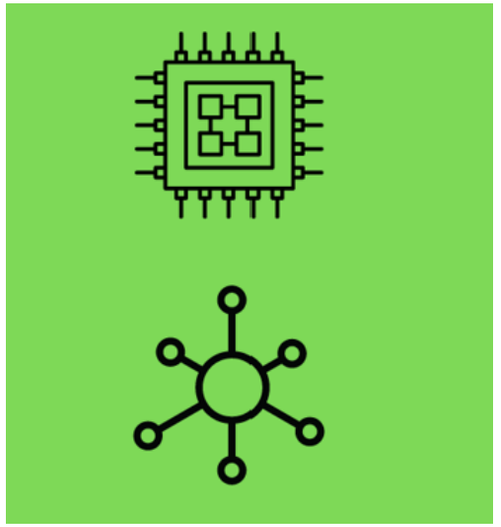
Inclusiveness

Adoption

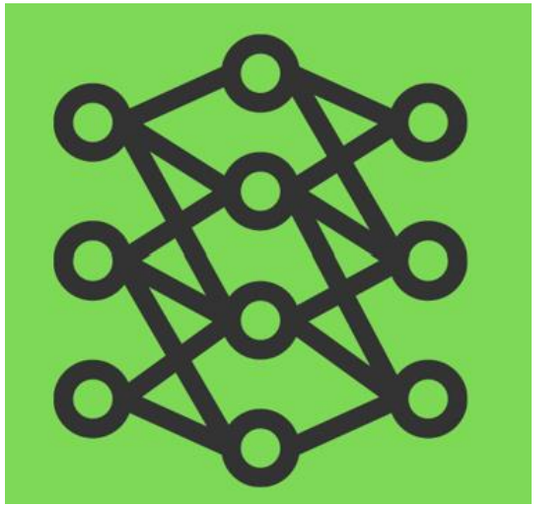
Environment

What is Green AI?

What is sustainable AI?



Alternate deployment strategies



Elevate smaller models

Carbon-efficiency and Carbon awareness

[The Imperative for Sustainable AI Systems (thegradient.pub)]

Understand – Emission scopes

Greenhouse Gas Emissions (GHG) measurement

Scope 1

Direct emissions **created by your activities,** like consumption of gas, fuel oil or even leaks of refrigerants, present in the cooling and air conditioning circuits of data centers in particular

Scope 2

Indirect emissions from the production of electricity or heat you use to power buildings or processes

Scope 3

Indirect value chain emissions from all other activities in which you're engaged.
- Manufacturing, delivery and end of life of IT equipment related to the training and production of AI and edge equipment on which AI is deployed
- Purchases of technical and IT services and services dedicated to AI projects (software license, outsourcing, etc.)
- Use of the products / services targeted by the AI project

Understand - Considerations

All 3 scopes of GHG emissions

Entire life cycle of AI, from ideation and design to inference.

Impact of all the infrastructures and services associated with the AI project.

Include the Green AI approach within a more global Green IT approach.

Carbon is not only environmental impact of AI

[Green AI - impact-ai]

Peter Drucker “You can’t manage what you can’t measure”

Measure – ML lifecycle cost metrics framework

- **Training: ~12%** of models makes it to production
- **Inference: 80-90%** of carbon cost (NVIDIA)

Cost Metric	Training	Inference
Dollars	Jobs/pipelines	Operational Cost
Runtime	Core-seconds by SKU	Core-seconds by SKU
Energetic	GPU energy	GPU energy
Utilization	GPU Utilization (%) GPU Memory Utilization (%)	GPU Utilization (%) GPU Memory Utilization (%)

Operational Lifecycle Analysis Monitoring

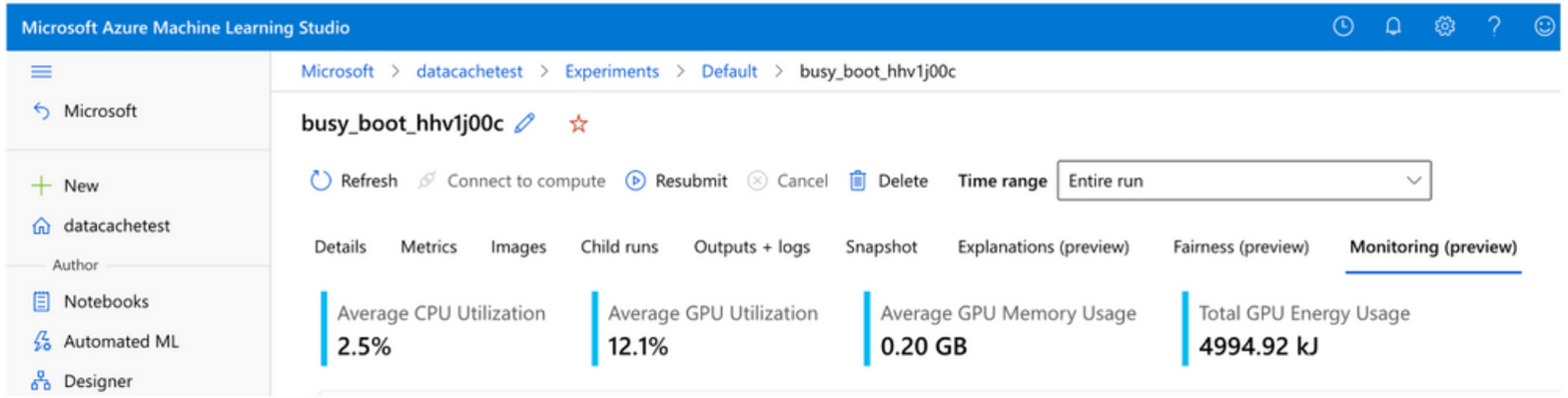
- **Monitoring Capabilities:** training/inference for cost ($, energy, carbon)
- **Tools:** Cost/benefit tradeoffs to optimize ROI

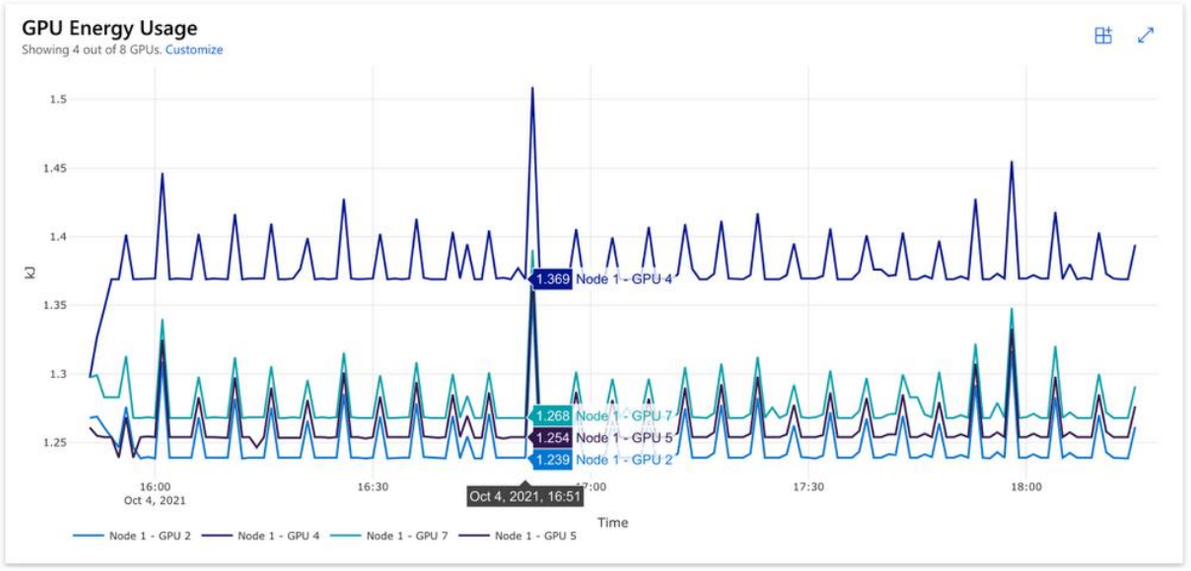
Measure – Training energy in Azure ML

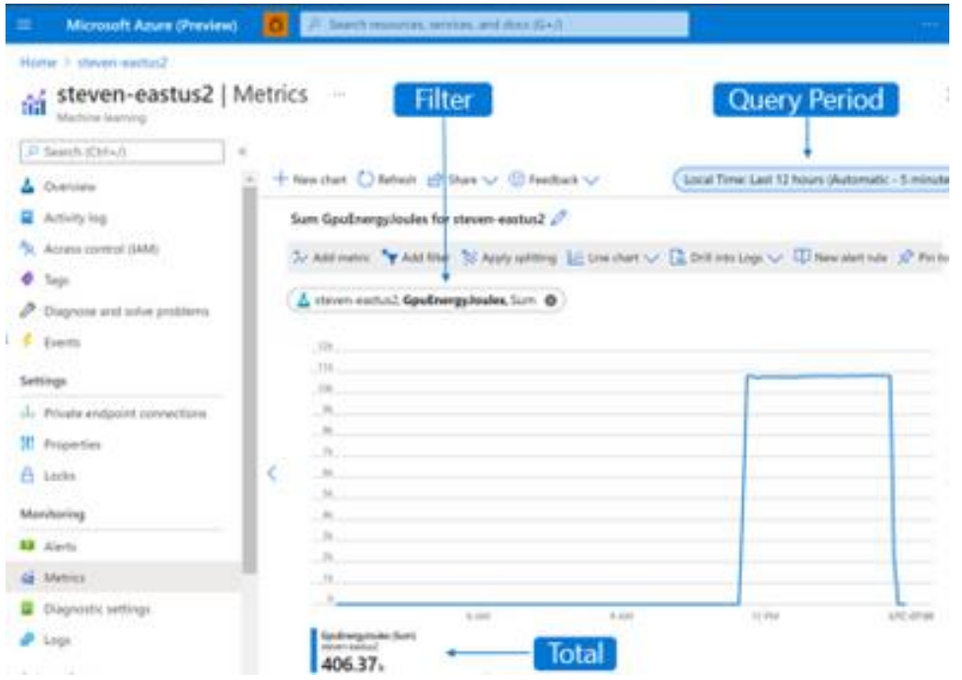
Opt-in metrics now let users sort to find the most expensive jobs & pipelines

- **Energy:** GPU energy consumed per job/pipeline (also avail in Azure Monitor)
- **Utilization**: GPU utilization, memory

Measure – Energy in Azure ML







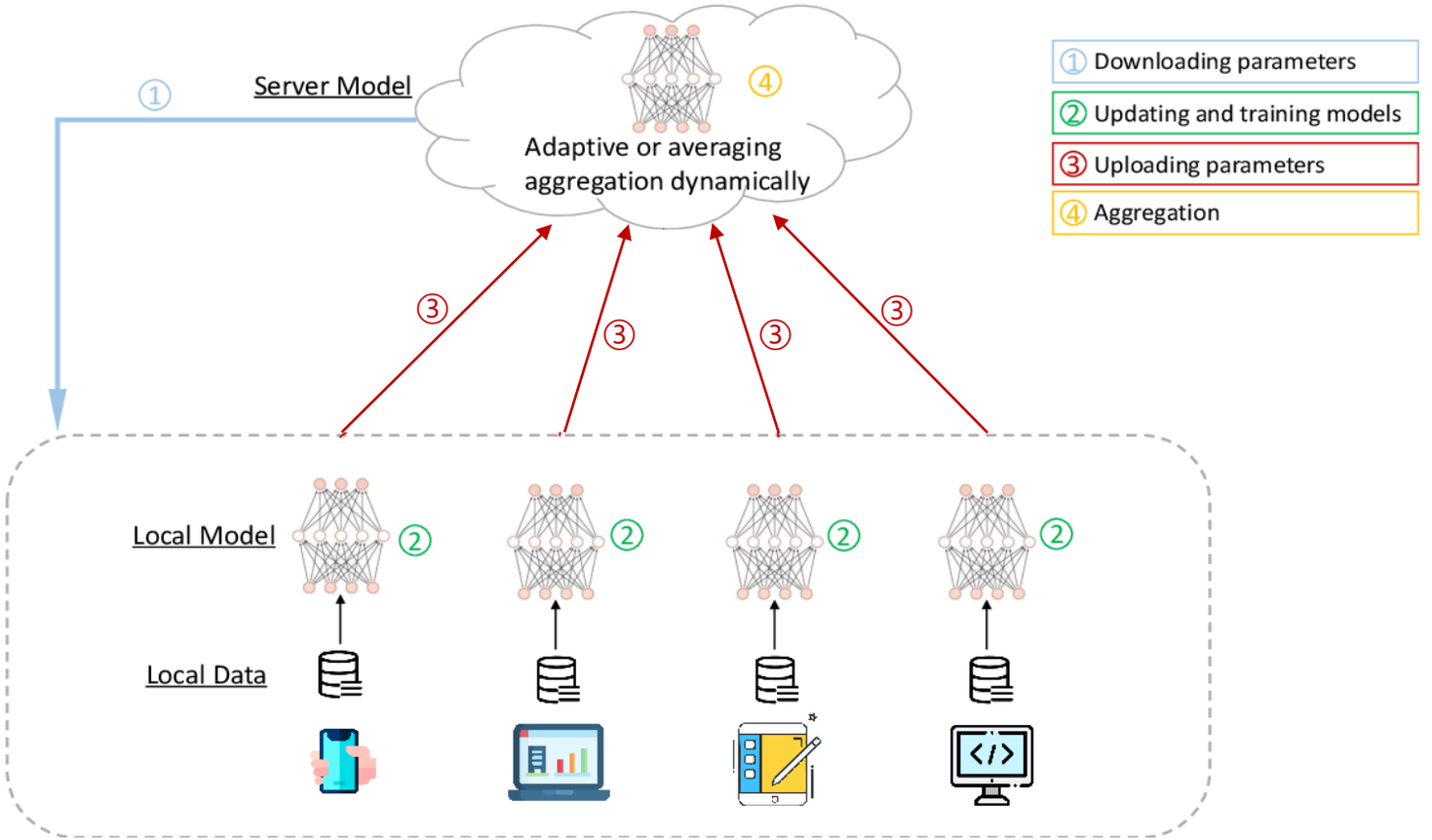
Reduce – Attenuate hardware impact

1. Use specialized hardware like ASICs/ FPGA to accelerate the run times of these jobs
2. Obtain higher utilization rates on existing hardware
3. Optimizing the use of existing hardware like general-purpose CPUs → reduce the demand for manufacturing new hardware

Scenarios & configurations on Azure	Supported DNN models	Regional support
+ Image classification and recognition scenarios + TensorFlow deployment (requires Tensorflow 1.x) + Intel FPGA hardware	- ResNet 50 - ResNet 152 - DenseNet-121 - VGG-16 - SSD-VGG	- East US - Southeast Asia - West Europe - West US 2

Reduce – Federated learning

Collaborative machine learning without centralized training data



https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Reduce – Federated learning

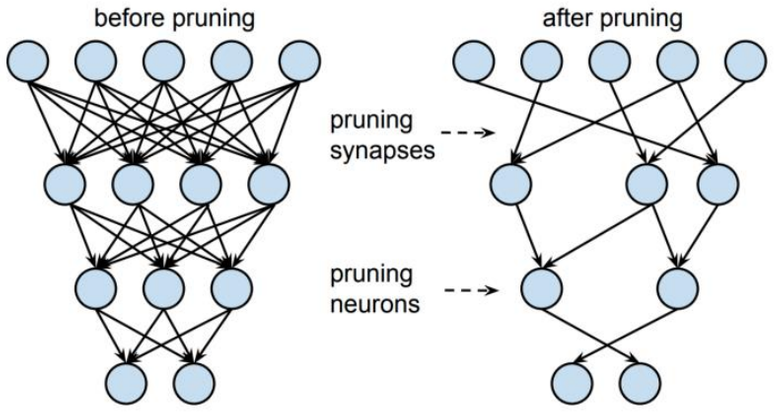
Country/CO2(g)	V100 $PUE = 1.67$	K80 $PUE = 1.67$	V100 $PUE = 1.11$	K80 $PUE = 1.11$	FL IID	FL non-IID
USA	1.6	5.2	1.1	3.5	0.5	1.0
China	2.9	9.2	1.9	6.2	0.9	1.7
France	0.2	0.8	0.2	0.5	0.1	0.1

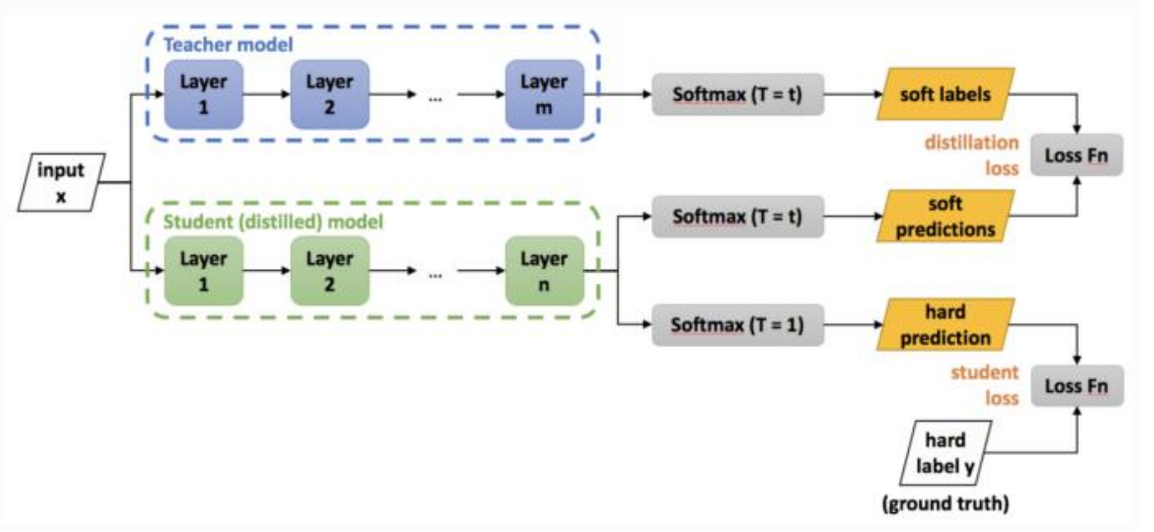
Table 4: CO_2 emissions (expressed in grams, *i.e.* **lower is better**) for centralized training and FL on Fashion-MNIST. Emissions are calculated once the top-1 accuracy on the test set reaches 90%. The number of epoch reported on the FL column relates to the number of local epoch done per client. "IID" and "non-IID" terms are employed to distinguish between clients that have an evenly distributed set of samples containing all the classes (IID) and clients that have more samples of certain classes (non-IID).

Can federated learning save the planet?

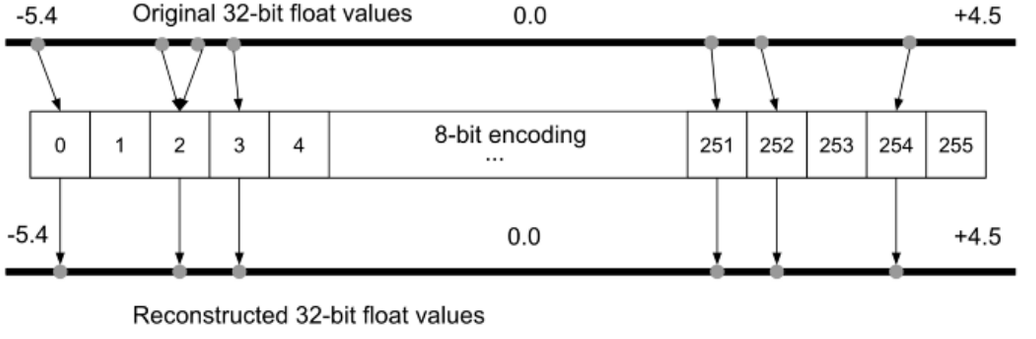
Reduce – Elevating smaller models

Pruning



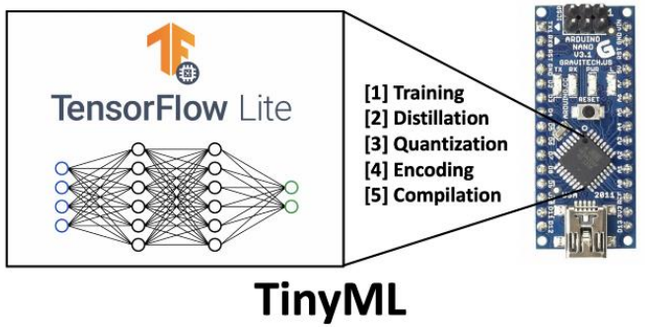


Knowledge distillation Source: ArXiv.



Quantization & Factorization

Reduce – Tiny ML



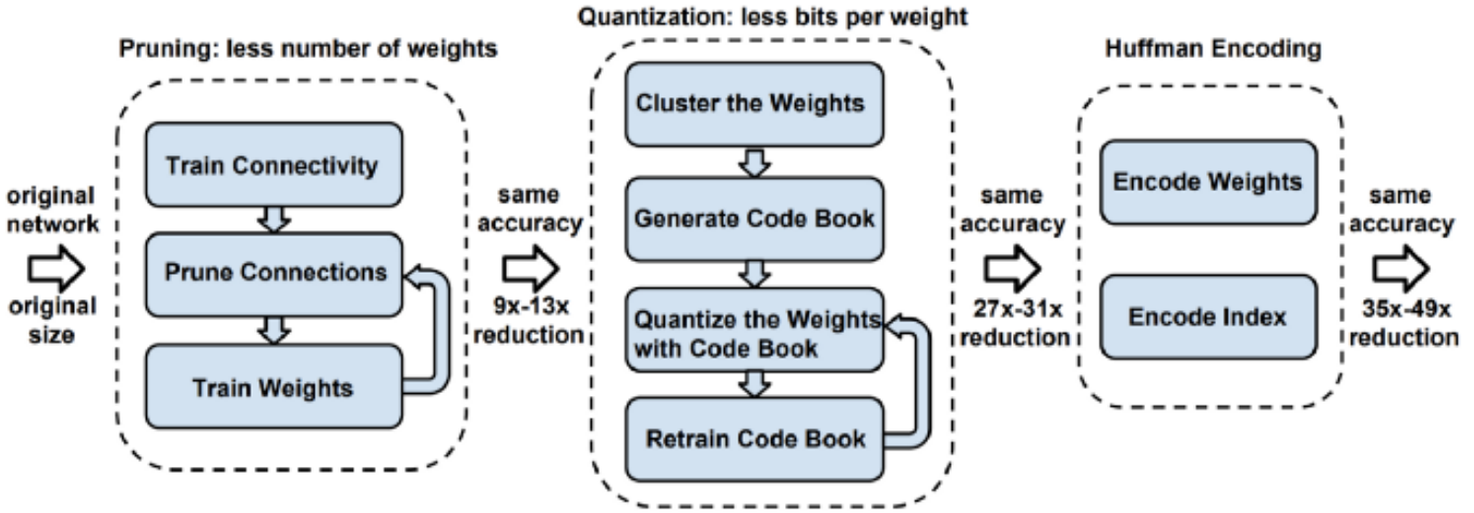


Diagram of the deep compression process. Source: ArXiv.

Reduce - Efficient model training

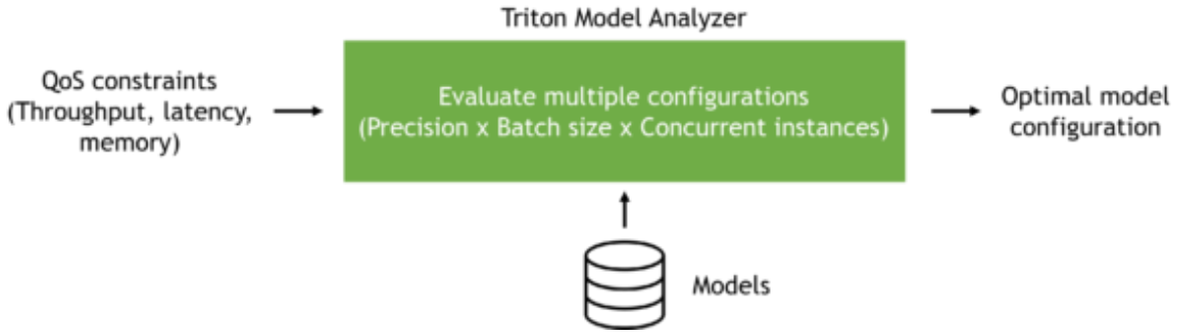
- Use feature stores

- Training refined into two stages for LLM:
  - Pre-training of a general model
  - Fine-tuning to produce accurate outcomes on a specific task

- Neural architecture search (NAS) and Hyperparameter Optimization (HPO) can also be used to satisfy different objective functions, such as computational efficiency or cost.

- muTransfer, that can transfer training hyperparameters across model sizes
  Enables equivalent accuracy levels while using at least an order of magnitude (~10x) less compute, with no limit to the efficiency gain as the target model size grows.

Reduce – Efficient inferencing



Overview of NVIDIA Triton Model Analyzer

Online Result Summary

Model: bert-large

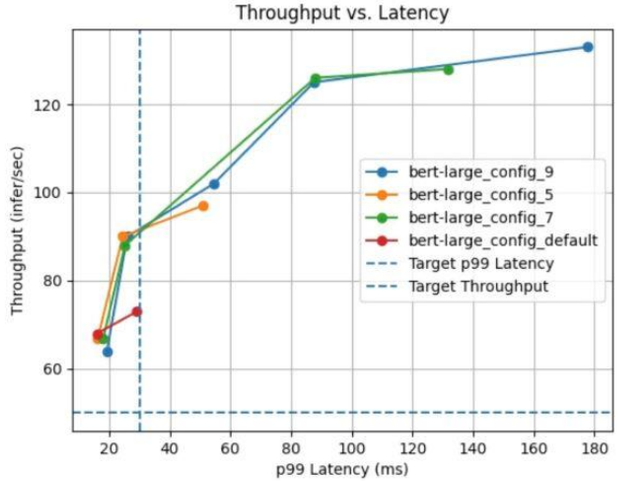
GPU(s): A100-SXM4-40GB

Total Available GPU Memory: 39.6 GB

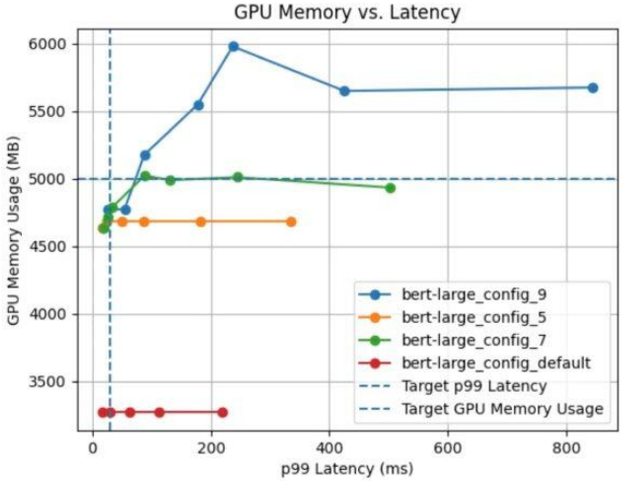
Constraint targets: Min Throughput : 50 infer/sec, Max p99 Latency : 30 ms, Max GPU Memory Usage : 5000 MB

In 161 measurement(s), config bert-large_config_9 (2/GPU model instance(s) with max batch size of 16 and dynamic batching enabled) on platform pytorch_libtorch delivers maximum throughput under the given constraints on GPU(s) A100-SXM4-40GB.

Curves corresponding to the 3 best model configuration(s) out of a total of 23 are shown in the plots.



Throughput vs. Latency curves for 3 best configurations.



GPU Memory vs. Latency curves for 3 best configurations.

Think Green

- Red AI
  - Big models, large datasets
  - Inclusiveness, adoption, environment

- Green AI
  - Enhance reporting of computational budget
  - Promote efficiency as a core evaluation for AI

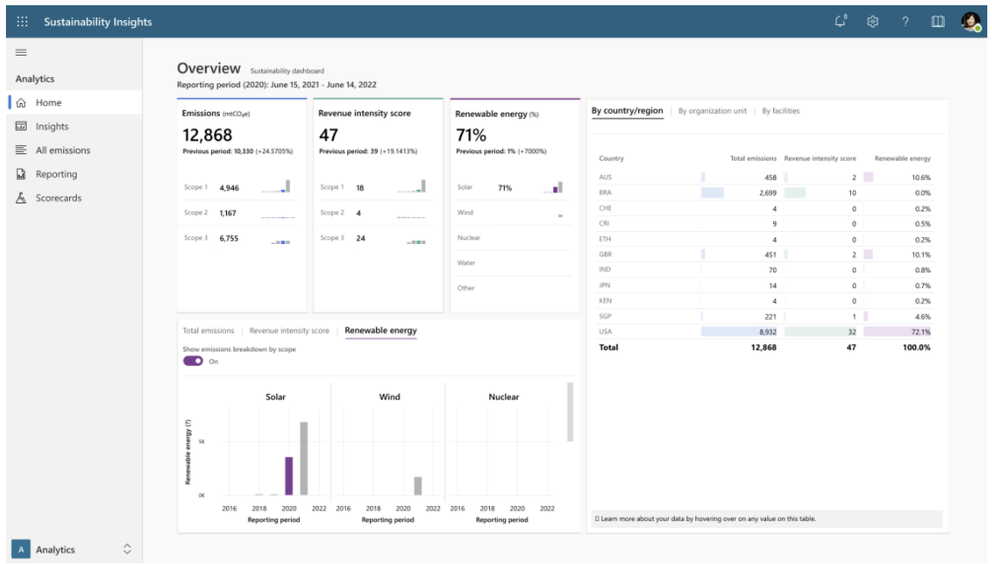

QUESTIONS
ANSWERS

Merci de votre participation !

Ressources utiles

- Green Software Foundation
  https://greensoftware.foundation/
- Sustainable Software devblog https://devblogs.microsoft.com/sustainable-software/
  https://devblogs.microsoft.com/sustainable-software/measuring-the-carbon-impact-of-web-browsing/
- Principles of Green Software Engineering
  https://principles.green/fr-fr
- The principles of sustainable software engineering
  https://learn.microsoft.com/en-us/training/modules/sustainable-software-engineering-overview
- Well Architected Framework §Sustainability
  https://review.learn.microsoft.com/en-us/azure/architecture/framework/sustainability/
- https://review.learn.microsoft.com/en-us/azure/architecture/framework/sustainability/sustainability-application-design
- The Shift Project
  https://theshiftproject.org/en/article/unsustainable-use-online-video/
- Green Software Lab :
  https://github.com/greensoftwarelab
- Github repos avec le tag « energy consumption » :
  https://github.com/topics/energy-consumption

Try Microsoft Sustainability Manager free for 30 days



Experience Microsoft Sustainability Manager - a Microsoft Cloud for Sustainability solution that unifies data to help you monitor and manage your environmental impact.

Microsoft Sustainability Manager | Trial

Let's get started

Enter your email to start your free trial *

By signing up, you agree to the Microsoft Online Subscription Agreement and Supplemental Terms governing this trial and Microsoft Privacy Statement. Preview will be provisioned in English only.

Start your free trial

Calculate your cloud footprint

- https://www.microsoft.com/en-us/sustainability/emissions-impact-dashboard?activetab=pivot_2:primaryr12

Emissions Impact Dashboard

by Microsoft Corporation

Power BI apps

★ 2.0 (18 AppSource ratings)

Pricing Free | Get it now | ☆ Save to my list

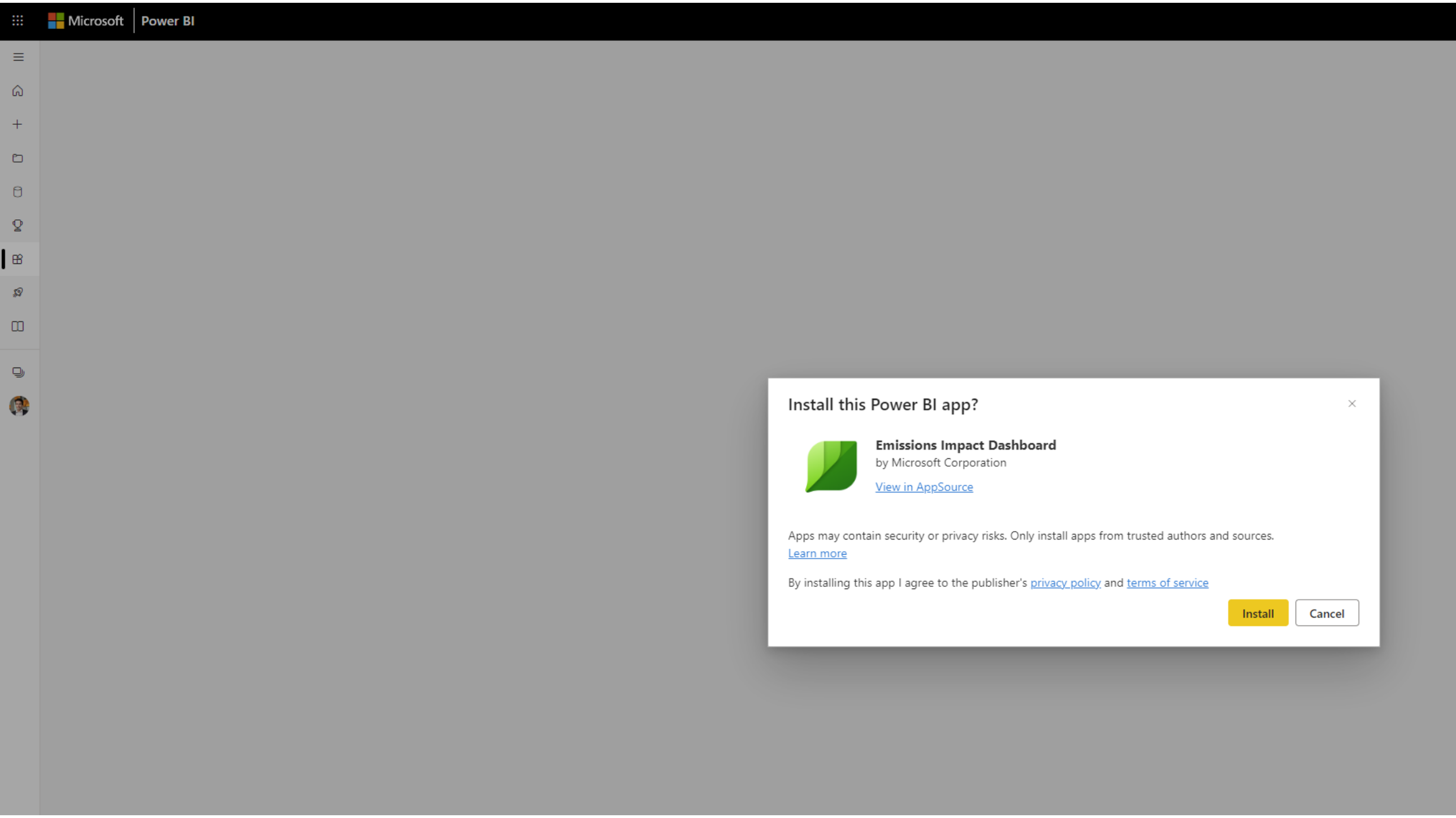
Emissions Impact Dashboard

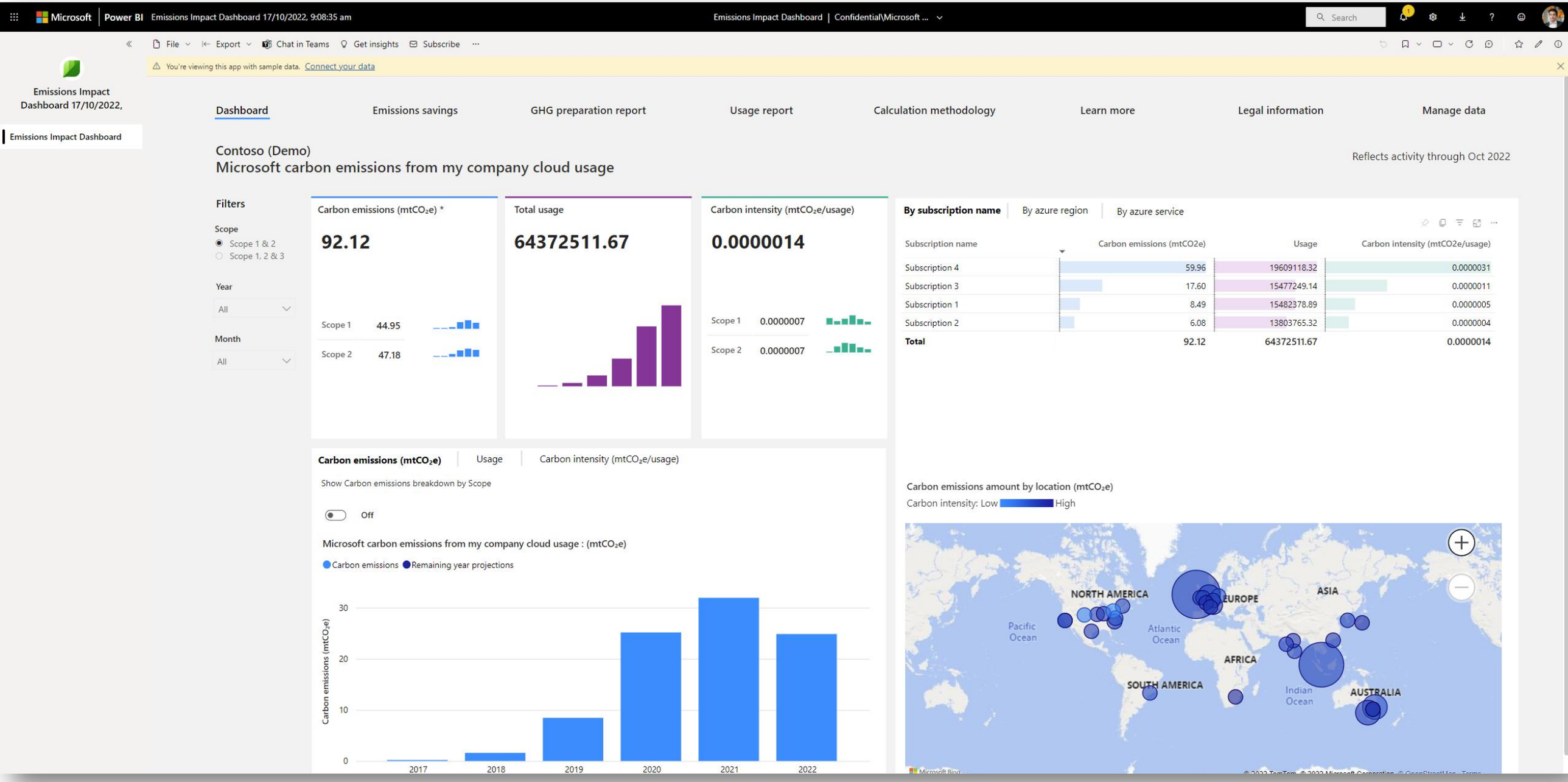
Calculate your carbon emissions from cloud-based computing

The Emissions Impact Dashboard provides our cloud customers with transparency into their greenhouse gas emissions associated with cloud usage. Our Generally Available version and is available for anyone to download.

Requirements for getting started

- The Emissions Impact Dashboard runs on Power BI Pro. Get the Power BI Pro free trial.
- The Emissions Impact Dashboard can support EA, Microsoft Customer Agreement, and CSP for Azure customers. Enrollment ID or Billing Account ID, each with Admin access, is required to connect your company data to the calculator. For more info on how to connect the calculator, visit this site.


Microsoft
Power BI
Install this Power BI app?
Emissions Impact Dashboard
by Microsoft Corporation
View in AppSource
Apps may contain security or privacy risks. Only install apps from trusted authors and sources.
Learn more
By installing this app I agree to the publisher's privacy policy and terms of service
Install
Cancel


Microsoft | Power BI Emissions Impact Dashboard 17/10/2022, 9:08:35 am
Emissions Impact Dashboard | Confidential\Microsoft ...
Search
File Export Chat in Teams Get insights Subscribe
You're viewing this app with sample data. Connect your data
Emissions Impact Dashboard 17/10/2022,
Emissions Impact Dashboard
Dashboard
Emissions savings
GHG preparation report
Usage report
Calculation methodology
Learn more
Legal information
Manage data
Contoso (Demo)
Microsoft carbon emissions from my company cloud usage
Reflects activity through Oct 2022
Filters
Scope
Scope 1 & 2
Scope 1, 2 & 3
Year
All
Month
All
Carbon emissions (mtCO2e) *
92.12
Scope 1 44.95
Scope 2 47.18
Total usage
64372511.67
Carbon intensity (mtCO2e/usage)
0.0000014
Scope 1 0.0000007
Scope 2 0.0000007
By subscription name
By azure region
By azure service
Subscription name
Carbon emissions (mtCO2e)
Usage
Carbon intensity (mtCO2e/usage)
Subscription 4 59.96 19609118.32 0.0000031
Subscription 3 17.60 15477249.14 0.0000011
Subscription 1 8.49 15482378.89 0.0000005
Subscription 2 6.08 13803765.32 0.0000004
Total 92.12 64372511.67 0.0000014
Carbon emissions (mtCO2e)
Usage
Carbon intensity (mtCO2e/usage)
Show Carbon emissions breakdown by Scope
Off
Microsoft carbon emissions from my company cloud usage : (mtCO2e)
Carbon emissions
Remaining year projections
Carbon emissions (mtCO2e)
30
20
10
0
2017
2018
2019
2020
2021
2022
Carbon emissions amount by location (mtCO2e)
Carbon intensity: Low High
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA

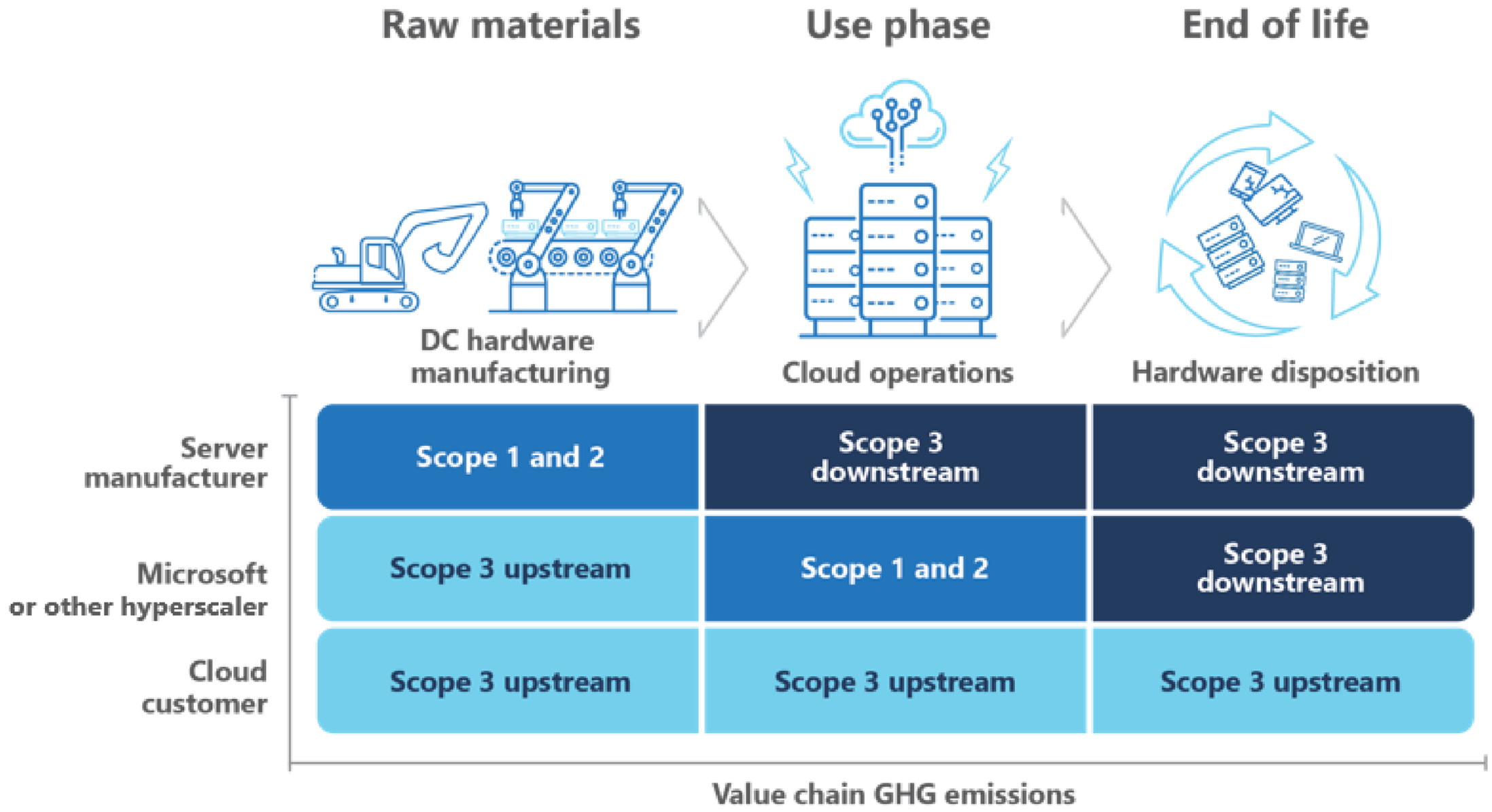


Figure 3: Scope 1, 2, and 3 sources of emissions for a cloud ecosystem.

Source : https://www.opencompute.org/documents/lca-sop-in-ocp-document-submission-template-docx-pdf

Microsoft Azure