GE VERNOVA

THE DATA CENTERS AI & ML TRILEMMA

Ihab Chaaban P.Eng. MBA

Aeroderivatives Global Commercial Development Director
Gas Power – GE Vernova

CONTENTS

INTRODUCTION:

As the world is making efforts to decarbonize at an affordable, reliable, and sustainable aim, the Information technology Industry has been growing at a higher rate than the energy infrastructure. Data Centers, 5G, cloud computing and block chains are just a few examples that are imposing a growing burden on utilities and energy sources. Artificial Intelligence (AI) and Machine Learning (ML) added on the top of the existing challenges more ones that will require a holistic view to prepare for what's coming to fulfill the data centers energy demand while minimizing the environmental impact.

AI and ML have been around since the last century, however, the pace of technological innovations in the last decade substantially accelerated in a way that truly started to affect our daily lives. From reducing the time needed to develop reports, and presentations to creating prototypes and automating other tasks, the evolution of AI and ML proved to be game changers. Nevertheless, upstream, the environmental and energy impacts are crucial to the extent of becoming obvious and challenging for data centers, utilities, and power providers. This paper discusses those challenges, explains the impacts of this evolution, and proposes innovative solutions and concepts to mitigate them.

HOW AI AND ML IMPACT THE ENVIRONMENT AND ENERGY INFRASTRUCTURE

In a previous white paper *(GEA35139A – Greening the Future Data Center Infrastructure via the GE Aeroderivative Technology and Microgrid Controls)*, it was shown how 5G and cloud computing affected the growth of data centers to the extent of the existing hyperscale in the industry. It was also shown how the Aeroderivative gas turbine technology was able to mitigate the footprint, power density, and complexities impact imposed by the old model where the data centers relied on diesel reciprocating gensets for standby power. Additionally, the paper shed some light on microgrids, and thermal hybrid technology as means to reduce the carbon footprint. Now, with AI and ML, a new challenge, other than power density, and carbon footprint has emerged, that is, water footprint. Hence, we're currently facing a Trilemma that we need to solve to continue fulfilling the demand in a reliable and more sustainable way.

The Power Usage Effectiveness (PUE), which is the ratio of the total amount of energy used by a computer data center facility to the energy delivered to computing equipment, is one the most significant Key Performance Indicators (KPI) for any Data Center. The correlation between KW (Kilowatts) and GB (Gigabytes) is always monitored as a design parameter. In an ideal case, the PUE ratio should equal to one, nevertheless, there will always be the inefficiencies that increase that ratio to more than one, especially those related to heat loads and the cooling system. As AI and ML have been making strides rapidly, the conventional CPUs (Central Processing Units) at the heart of the servers were not enough and hence the High-performance Computing (HPC) required adding Graphics Processing Units (GPU's) which have higher thermal power densities when compared to conventional architectures per Vertiv. As the servers' design is stated in the form of "rack density", Data Centers Frontier (DCF) mentioned in one of its articles that 7 KW per rack, is manageable by most data centers with the conventional air-cooling system. Even with the climb towards 8 to 16 KW per rack, we may not need introducing new types of cooling such as liquid cooled servers. DCF also mentioned that some data centers went to mineral oil cooling and other refrigerants, such as CO2, as the rack density kept climbing. According to Vertiv, Air-based cooling systems lose their effectiveness when rack densities exceed 20 KW, at which point liquid cooling becomes the viable approach. That climb in rack densities that could surpass 70 KW in the near future is not a myth anymore.

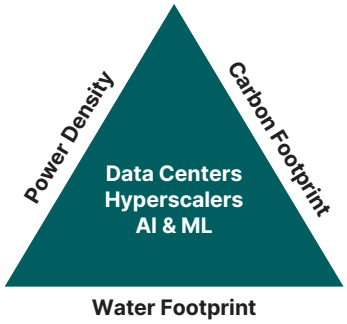


Figure 1: The AI & ML Trilemma

AI utilizes chatbots (such as Microsoft's several versions of ChatGPT and Google's BARD) to train. Per Pengfei Li from UC Riverside and other Authors paper: Making AI Less "Thirsty", training GPT-3 in Microsoft's, state-of-the-art U.S. data centers can directly consume 700,000 liters of clean freshwater which is concerning, as freshwater scarcity has become one of the most pressing challenges for the environment. The Times of India (and as in Pengfei's paper) simplified the analogy in one of its articles to an example of 20-50 question-long conversation with a chatbot like ChatGPT can consume 500 ml of fresh water to cool down the servers. Despite the existence of several technologies for servers cooling per Vertiv, such as servers' immersion cooling and rear door heat exchangers, the water footprint is still an inevitable challenge to overcome, for those data centers required to have liquid cooling, especially in regions with water scarcity.

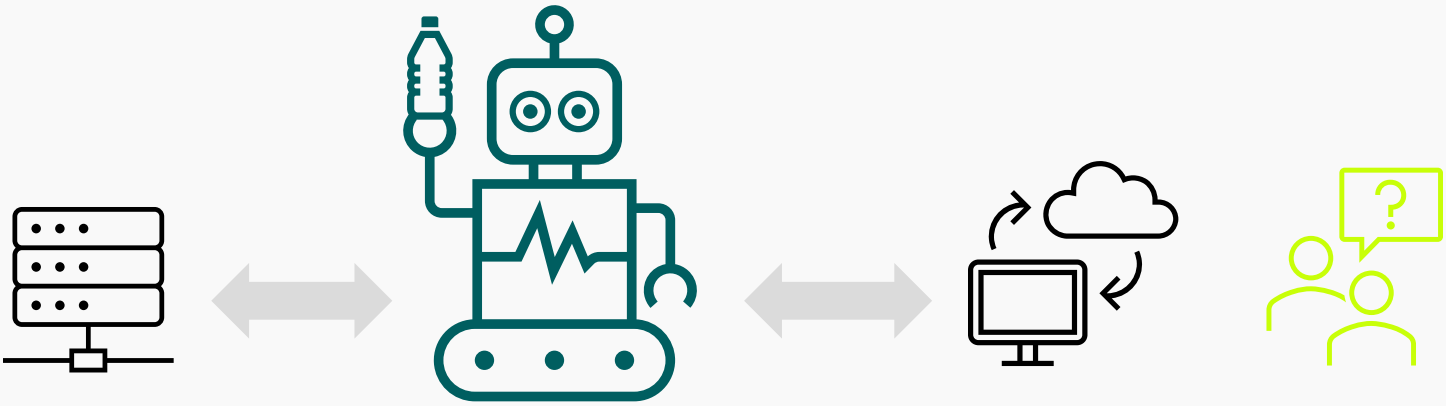


Figure 2: A Thirsty Chatbot

Now, let's not forget the elephant in the room, that is the power (MW) and energy (MWH) to support all this rack density growth. As most of data centers rely on utilities and IPPs (Independent Power Producers) to provide their main supply through redundant substation feeds, the burden on the grids, from Generation to Transmission and Distribution is excruciating. All this is occurring simultaneously with grids supporting electrification and decarbonization. Additionally, let's recall the numerous diesel gensets to back up the utility feed in case of failure to maintain a Tier level that is either mandated by the Uptime Institute or a self-regulated norm within the Hyperscalers.

And last but not least, the Carbon Footprint. According to Stanford Artificial Intelligence Index report 2023, several machine models were selected to scan their CO2 equivalent emissions. As examples of this study, such as Gopher, BLOOM, GPT-3 and OPT. GPT-3 in a 1.1 PUE data center had 429gCO2eq/kwh when consuming 1287 MWH leading to 502 tons of CO2 Equivalent emissions and 552 tons when multiplied by the PUE. Gopher, in a 1.08 PUE data center had 330gCO2eq/kwh when consuming 1066 MWh leading to 352 Tons of CO2 Equivalent Emissions and 380gCO2eq/kwh when multiplied by the PUE. According to MIT research that published in digitally.cognizant.com, it was estimated that ChatGPT-3 consumed 936 MWh, which is enough to power approximately 30,632 US households for one day, or 97,396 European households for the same period. Needless to say, that the numbers are staggering. The fact that most of hyperscale data centers offset their carbon footprint by purchasing RECs (renewable Energy Certificates or Credits), may not be a sustainable solution given the forecasted growth that could lead to unaffordable pricing of those certificates when all ML models are injecting the said levels of CO2 emissions equivalent and beyond, and competing for those RECs to offset them.

ADDRESSING THE TRILEMMA

A. The Power Density and Grid burden

As explained in the preceding paragraphs how the HPC and GPUs in AI and ML models led to the higher rack densities in data centers in general and in hyperscalers in specific, which contributed to larger data centers growth and expansion. The conventional practice of relying on grid power where data centers owners sign long term Power Purchase Agreements (PPA) with IPPs and utilities with renewable assets and import the remainder from other sources while relying on standby power from diesel gensets may not be sustainable at the AI scale that could easily exceed 100 MW. Therefore, the Aeroderivative technology was proposed in a previous paper (GEA35139A) as a contemporary solution when utilized as a standby resource or even as the prime one. A single GE Vernova LM2500XPRESS aeroderivative package with around 35 MW could replace 11-12 diesel gensets approximately in standby applications saving on real estate, switchgear, transformers, and overall footprint.

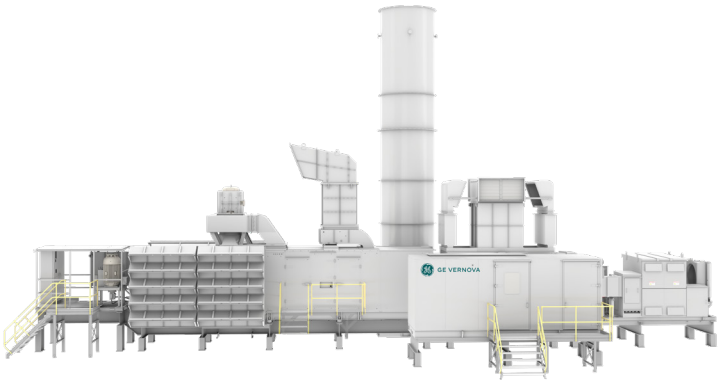


Figure 3:GE Vernova's LM2500XPRESS

As grids try to fulfill the capacity and energy needs of data centers, they are facing challenges to move from conventional resources utilizing fossil fuel, such as coal, that has been supporting the grid for decades, towards Renewables and energy storage technologies, especially Battery Energy Storage Systems (BESS), known for short term storage durations. Those challenges are even magnified when aiming at the Round the Clock (RTC) schemes. The growth of such schemes will lead to an extensive amount of penetration of Inverters Based Resources (IBRs) that behave completely differently from synchronous ones on the grid. The different profile of IBRs in short circuit, loss or increase of loads situations, makes zonal and nodal regions fragile and unstable which decreases the reliability of the grid. Hence maintaining the synchronous inertia in the system reduces substantially the risks of brown outs or black out incidences. Those events could have an economic impact on data centers when looking at the Value of Lost Load (VoLL) and Value of Supply Security (VoSS).

The synchronous resources on the grid, such as gas turbines, have a proven track record to enable a stable and reliable grid. An example of those synchronous resources is aeroderivative gas turbines that have been running on a variety of fuels for decades and where natural gas, that produces less carbon emissions than other fossil fuels, like coal, has been predominantly utilized as a bridge towards decarbonization, balancing Renewables and leading towards a cleaner, compared to other fossil fuels, and more stable grid. Hence, the utilization of the Aeros in data centers, in standby applications, can not only simplify the design in the said scale and reduce the footprint, but it can also support the grid when not serving the data center needs. Synchronous inertia and ancillary services (such as those shown in Fig.4) will grow in value as renewable and IBR assets keep the current growth pattern.

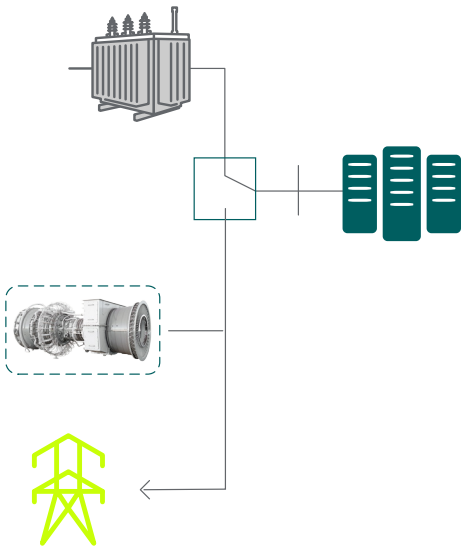


Figure 4: Aeros on Standby and exporting AS to the grid

ANCILLARY SERVICES - AERO

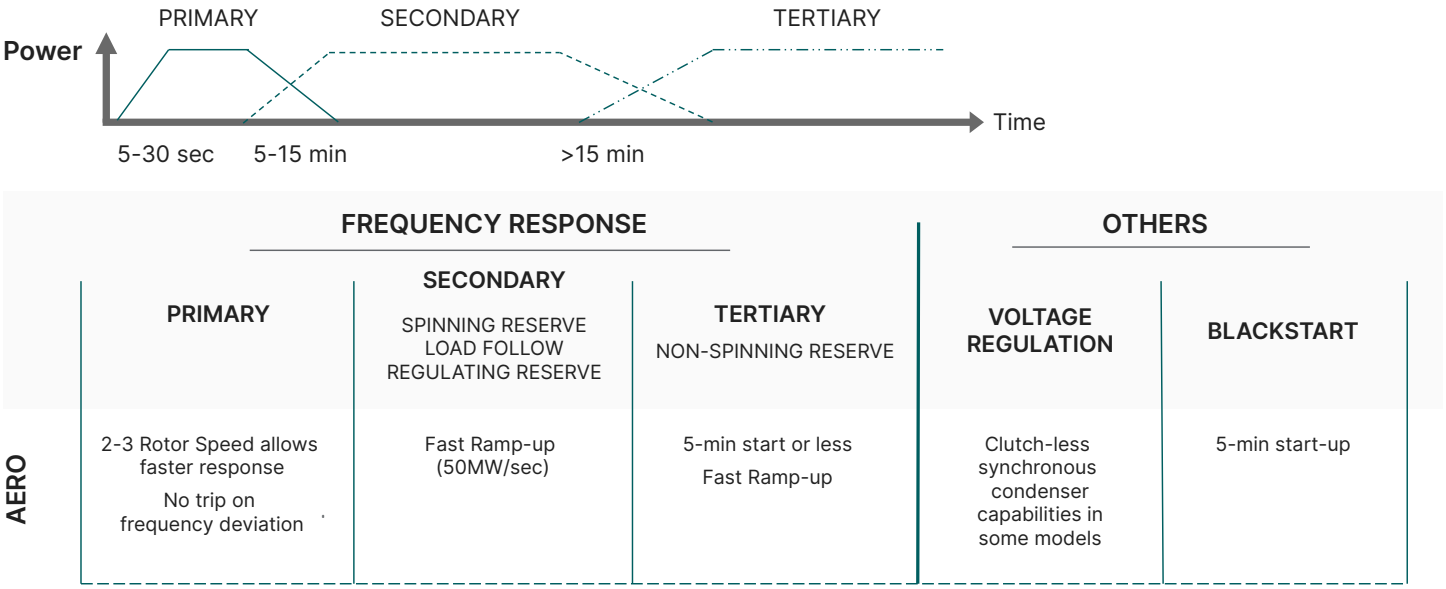


Figure 5: Examples of the types of Ancillary Services that Aeros can support

B. The Carbon Footprint

As explained in the previous sections, the CO2eq emissions were shown for several machine models in grCO2eq/KWH and the corresponding equivalent Tons. While Natural Gas contributes to carbon emissions, it is still considered a cleaner energy source than coal, for example, and capable of supporting the expansion of Renewables long term. The following chart in Fig. 5 shows how the GE Vernova's Aero technology can currently cut the carbon emissions when running on blends of Natural Gas and Hydrogen up to the 100% H2 fuel. The Aero technology, with its fuel diversity attribute, can combust a variety of other biofuels as well.

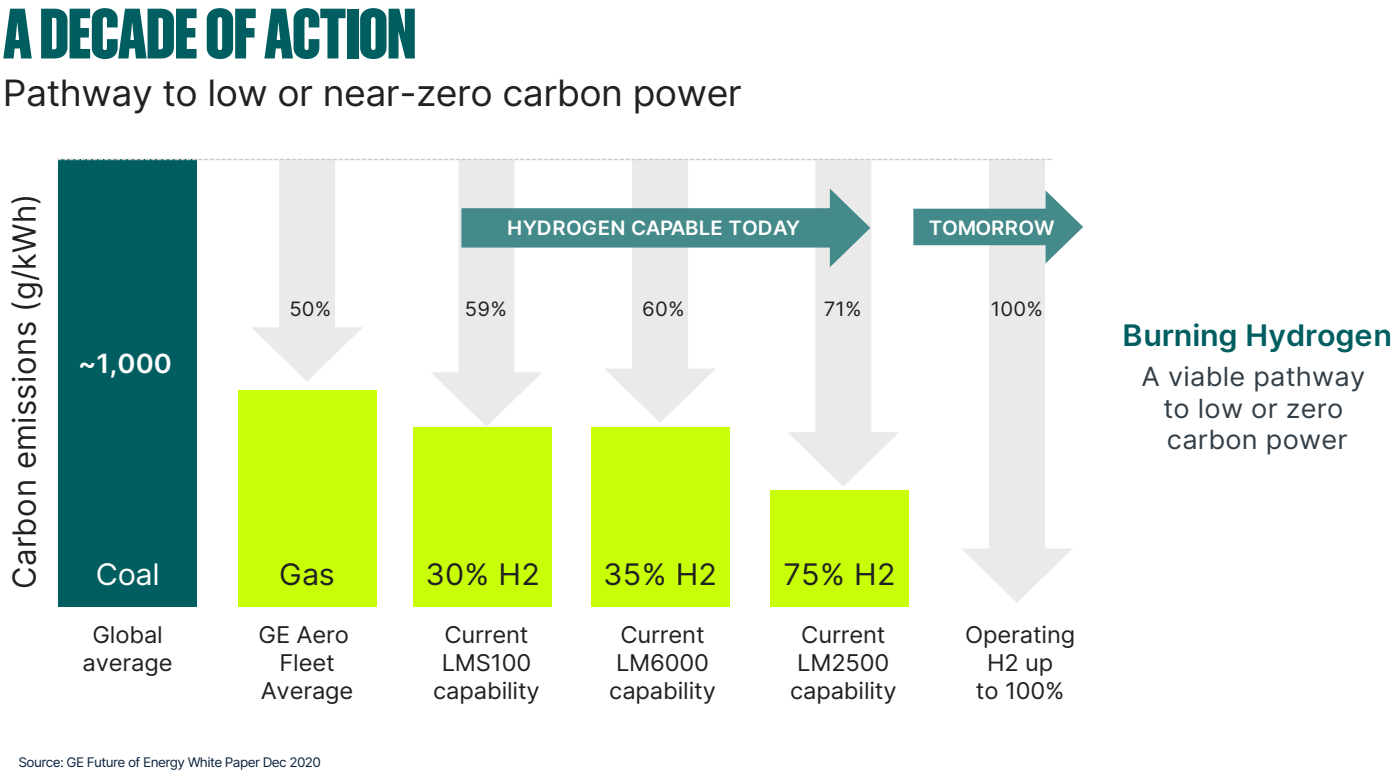


Figure 6: Pathway to low or near-zero carbon power

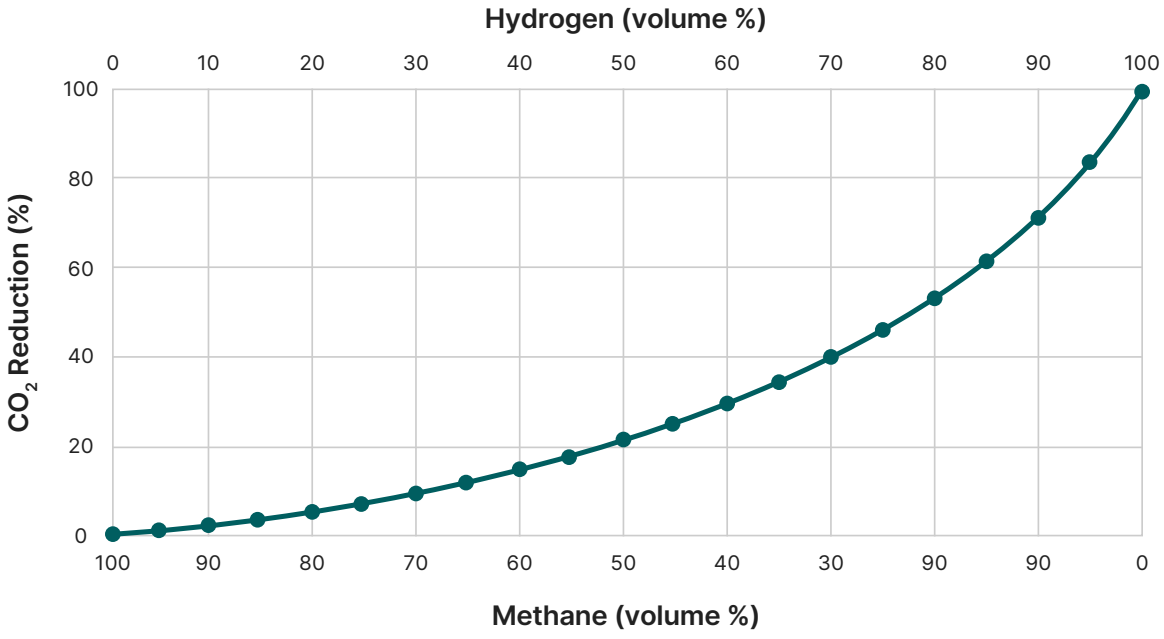


Figure 7: Decarb with H2 Blending

The Green Hydrogen economics remain challenging in terms of the LCOH (levelized cost of hydrogen) that could range between 4 – 7 $/KG subject to the Electrolyzers efficiency, Storage, and compression requirements. Nevertheless, the blends of Natural Gas with Hydrogen in the interim could be a viable option for certain carbon emission reduction targets. In order to achieve a zero-carbon operation, there are many elements needed beside the technology to achieve an affordable business model and look holistically into the hydrogen model from various angles, such as:

A. The Hydrogen economy could be accelerated via policy changes and regulatory subsidies to support an economical LCOH in $/Kg (the IRA in the US with the associated ITC and PTC for example)

B. The above, supported by Industrial off-takers, via use or take agreements, will provide more traction to establish a sustainable and affordable use case. In such structure, the business model could be bankable when supported by many parties other than the OEM / Developer conventional model.

Another alternative for reducing carbon emissions is to consider a thermal hybrid configuration via a Microgrid to model the operation in a way to let the renewable assets (Wind Turbines or Solar PV) to be the main contributor to the point of interconnect (Data Center Transformer) while the Aero technology would be utilized to firm a certain capacity or a capacity factor. This leads to a reduced amount of fuel burnt, lesser emissions in general and CO_2 in specific, as well as a lowered blended LCOE which all improve the bankability of the project. The mentioned RECs that are purchased by Data Centers to offset the carbon emissions of any KWH, from a non-green resource, to fulfill the demand could easily become costly when compared to the cost savings introduced by the said Microgrid scheme of operation. GE can model such operation via its in house developed tool called the Hybrid Architect. The Tool accounts for 52.91 Kg CO_2 /MMBTU (per eia.gov) of Natural gas combusted at the specified hourly conditions and load factor within the designated span of years to provide an estimate for the carbon footprint of the operation. Additionally, the model could combust a blend of Natural Gas and H2 among that hybrid operation, with all controls accounted for in the model for the interfacing between the assets, to reduce further the carbon footprint. This all leads to a substantial cut in the CO_2 emissions, in an economical way, as a bridge towards a zero-carbon operation.

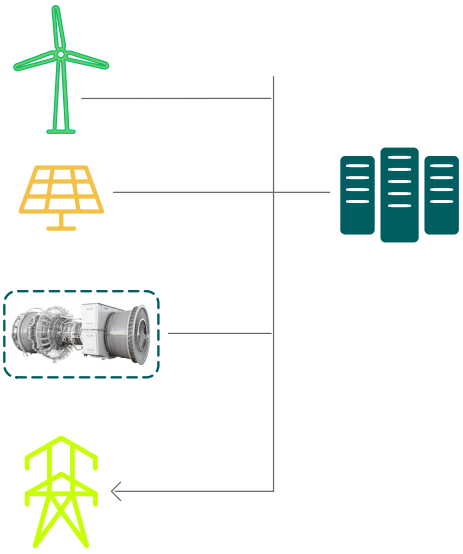


Figure 8: A Microgrid Thermal Hybrid configuration

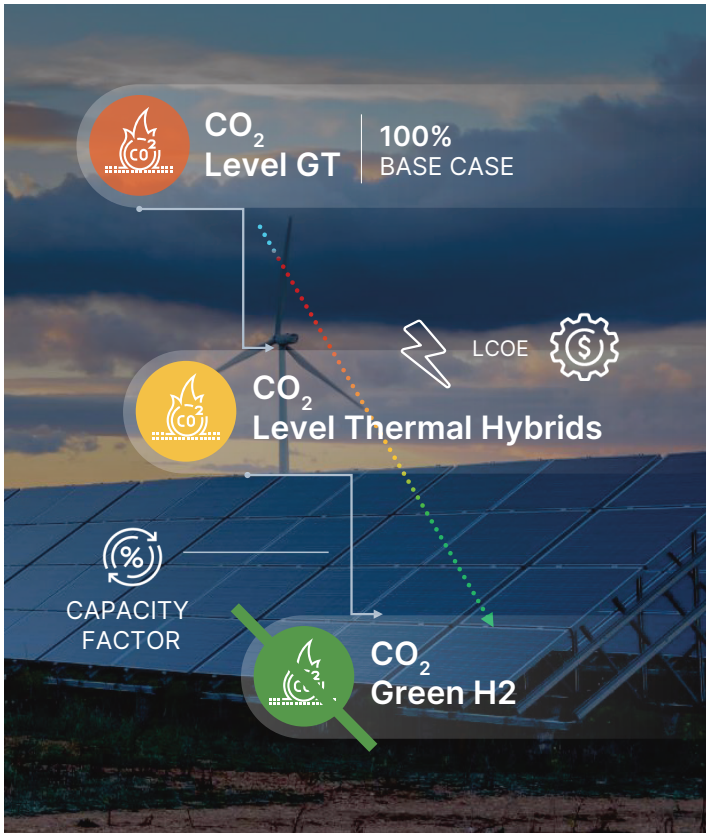


Figure 9: A staged modeling starting from 100% fossil fuel to thermal hybrids in the interim and ending with 100% Green H2 operation for zero carbon emissions

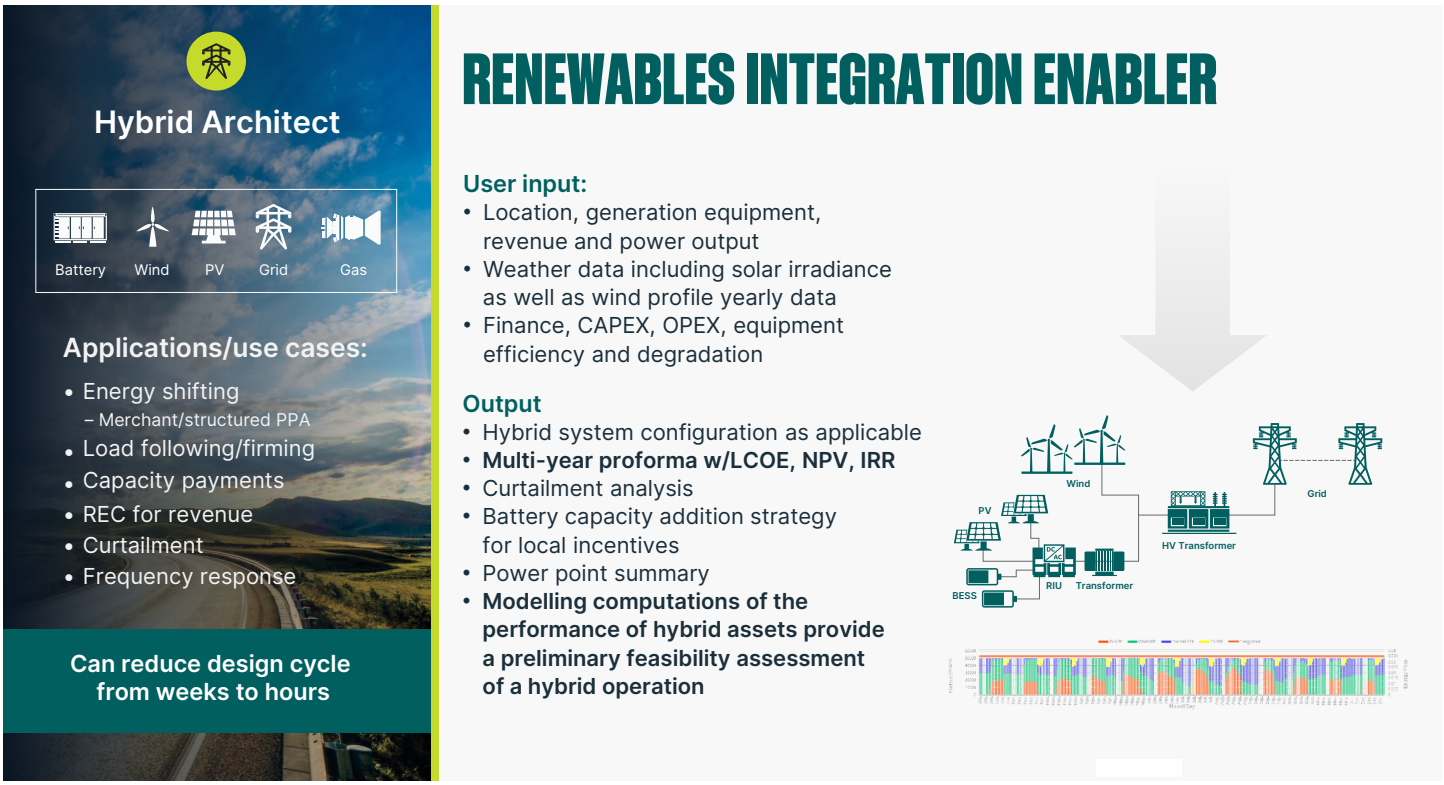


Figure 10: The Hybrid Architect: The renewables integration enabler

C. The Water Footprint

Since water cooling is now a requirement when crossing the 20 KW rack density, per Vertiv, due to HPC and GPU usage, it was a natural consequence to start looking into how that could be accomplished efficiently to reduce the water footprint. Per Pengfei's paper, data centers in general consume a significant amount of freshwater. For example, the paper mentioned an estimate, subject to the climate conditions, ~ 3.8 liters of water are consumed for each kWh of cooling load by an average data center, resulting in a water usage effectiveness (WUE) of 3.8 L/kWh, while some data centers can even use 5.2 L/kWh. When the ML models are training, the mentioned WUE ratios are expected to increase with hefty factors, however, the WUE is a variable figure that changes subject to many parameters such as the ambient conditions and the timing of the day when Renewables can supply power to the grid. So, for example, if the ML machine is training during the day when Solar PV energy is sufficient with a zero carbon (hypothetically) operation, normally at that time of the day, the WUE is at its worst. And, similarly, to conserve on water by training the machines during the night, then the energy from non-fossil sources and carbon footprint would be on the high side. So how do you solve this trilemma?

The previous sections showed how an Aero Gas Turbine could be deployed in a thermal hybrid Microgrid to reduce the carbon footprint and help address reliably the Power needs of a data center for prime power configuration. In addition to this scheme, heat recovery could be added to recover the heat from the Aero exhaust energy and feed a direct heat absorption chiller that creates chilled water that could be stored in an insulated tank to utilize for cooling the data center heat load. In that scheme, the chilled water could feed the servers rear door heat exchangers without the need to immerse the servers in water, which is still not favored by some data centers that prefer a non-direct contact between the cooling medium and the servers. As an example, for a 300 MW Data Center with an assumed 95% efficiency, the estimated heat load could be in the order of 15 MW which is equivalent to approximately 4265 Refrigeration Tons (RT). A single GE LM2500+G5, with a net output of ~30 MW could provide over double this RT when recovering the heat direct from the exhaust via an absorption chiller and reach an overall plant efficiency of up to ~80%. There could be other options, such as combined cycle and steam generation that could be analyzed subject to the use case. GE Vernova has over 800 units installed in cogeneration, with over 26 million

operating hours, which makes the GE Vernova aeroderivative gas turbine one of the most experienced technologies in the cogeneration world. More details on the GE Vernova aeroderivative packageexperience in heat recovery and CHP solutions can be found in the GEA34176 White paper.

The following diagram in Fig.11 provides a full layout of a proposed configuration to help address the trilemma of Power, Carbon, and Water footprints utilizing the Aero technology hybridized with Renewable assets and recovering its exhaust heat to generate chilled water for the Data Center cooling system feed. The configuration shed some light on each system that could be beneficial to data centers actively seeking growth in general and in the AI and ML arena in specific, while taking into consideration the discussed elements.

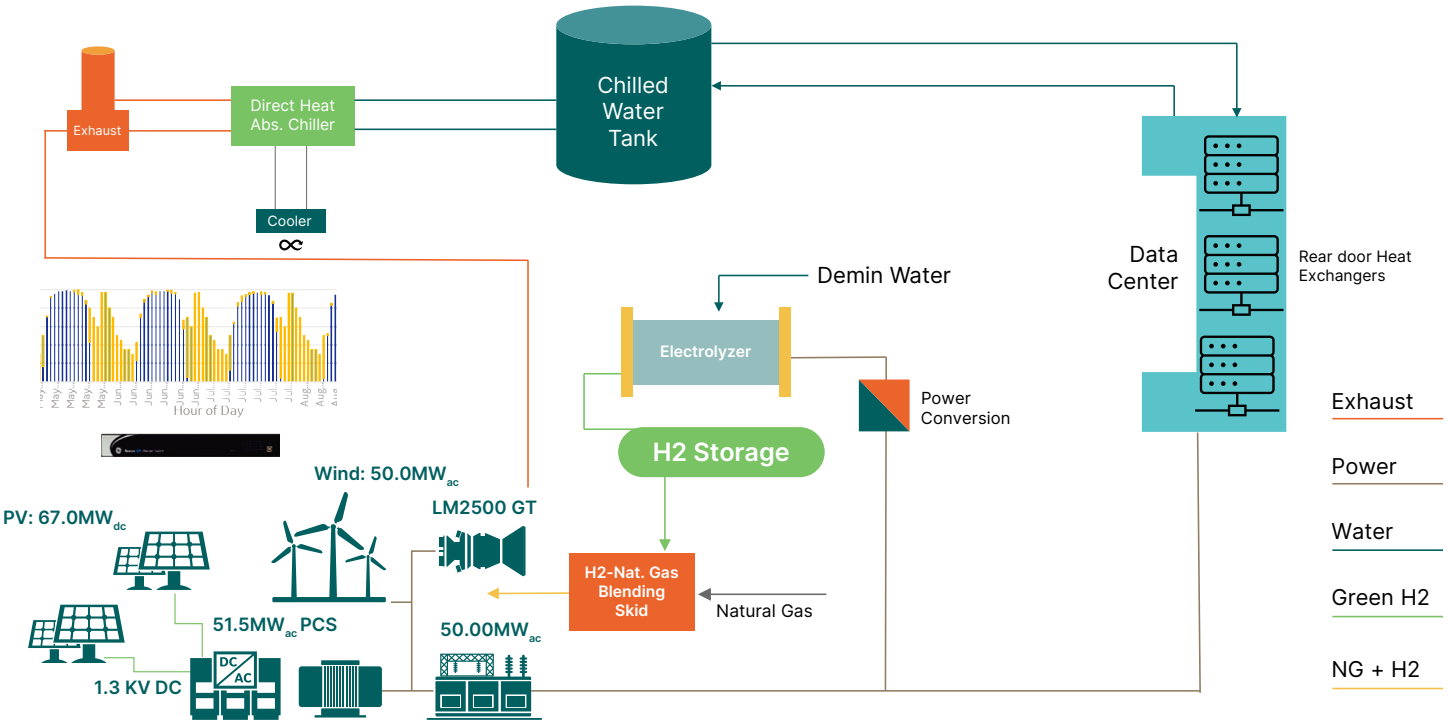


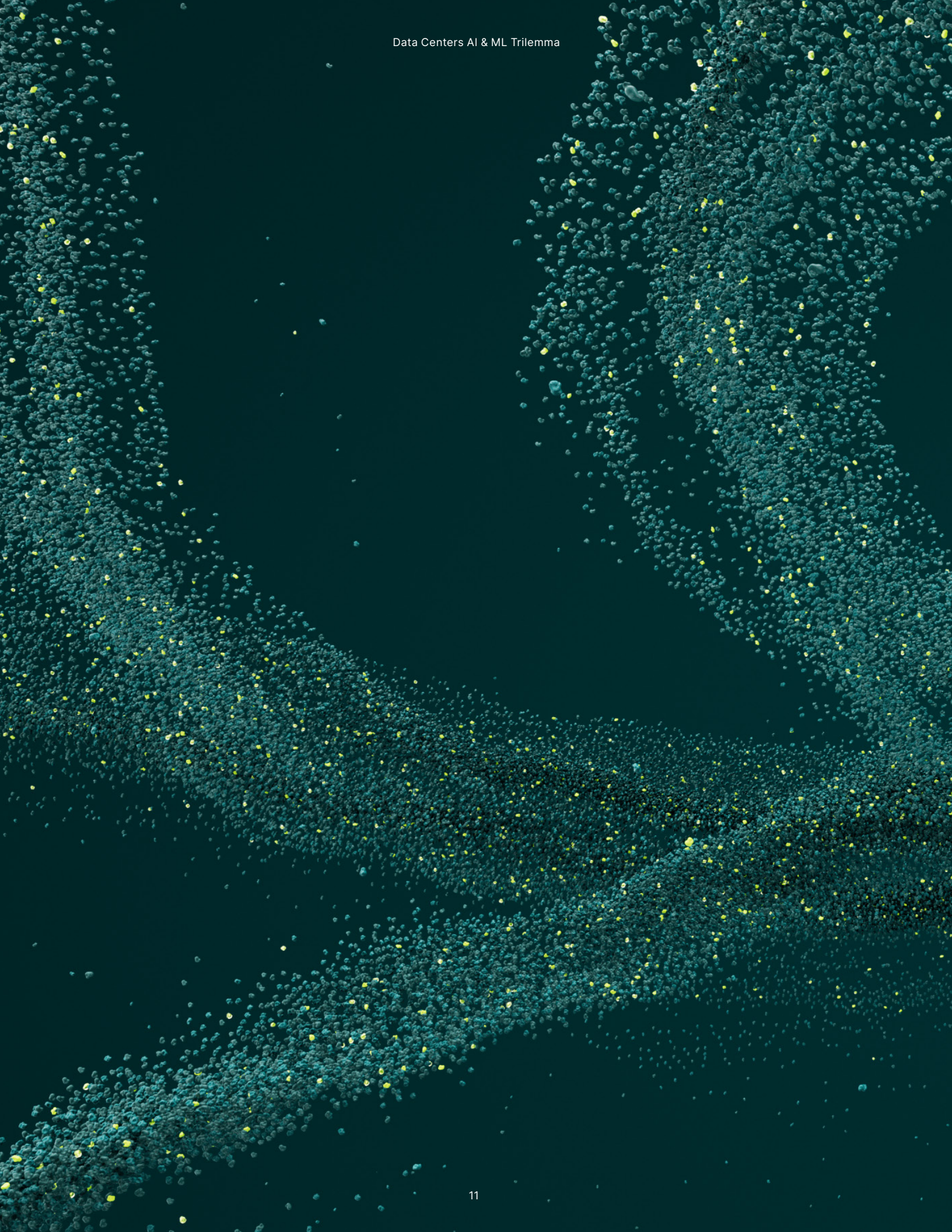
Figure 11: An integrated Thermal Hybrid Microgrid with a Direct Heat Absorption Chiller running on a blend of green hydrogen and natural gas

CONCLUSION

This white paper gave an overview on the impact of AI and ML on the power, carbon, and water footprints for Data Centers. It also provided some proposed schemes to integrate the Aero Technology, with heat recovery option, and Renewable assets in a Microgrid scheme to help address those impacts concurrently. Alternatively, the Aero technology could still be deployed to serve the data center as a standby asset, given its power density and high reliability inherited from its aviation roots while serving the grid in some use cases, when not supplying power to the data center. There are many ways to address the AI and ML challenges, and this paper provided a proposed view on some of those, keeping in mind a reliable, more sustainable, and economical operation. As Albert Einstein's once said, *"You can't solve today's problems with yesterday's solutions"*. Therefore, the data centers industry should consider innovative solutions to address today's challenges and continue the growth pattern efficiently and in a more sustainable manner.

REFERENCES

1. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models
   Pengfei Li UC Riverside – Jiyang Yang UC Riverside – Mohammad A.Islam UT Arlington – Shaolei Ren UC Riverside.
2. GEA35139A – Greening the Future Data Center Infrastructure via the GE Aeroderivative Technology and Microgrid Controls *Ihab Chaaban – Steve Halford.*
3. GEA34176 – Aeroderivative Gas Turbines Driving CHP Applications
   Helmut List – Special Thanks to Helmut List – Product Champion in GE Vernova on his materials on the Heat Recovery Systems.
4. The Hybrid Architect
   Special Thanks to Martin Yan – Value Engineering Leader from GE Vernova on his efforts creating and developing that GE Modeling Platform.
5. Artificial Intelligence Index Report 2023 – Stanford University
6. Understanding Data Center Liquid Cooling Options and Infrastructure Requirements
   Vertiv White Paper
7. Stanford Artificial Intelligence Index Report 2023
   https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf
8. https://www.datacenterfrontier.com/data-center-design/article/33005296/meta-previews-new-data-center-design-for-an-aipowered-future
9. https://techwireasia.com/2023/05/heres-how-generative-ai-is-affecting-data-centers/
10. https://timesofindia.indiatimes.com/india/ai-chatbots-may-be-fun-but-they-have-a-drinking-problem/articleshow/100634953.cms?from=mdr
11. https://www.businesstoday.in/technology/news/story/may-drink-a-500-ml-bottle-of-water-for-20-50-questions-chatgpt-data-centres-consumes-a-lot-of-water-warns-study-377473-2023-04-14
12. https://digitally.cognizant.com/ais-energy-use-isnt-sustainable-enter-tinyml-wf1584550#:~:text=At%20its%20current%20growth%20rate,generate%2C%20according%20to%20MIT%20research
13. https://cybernews.com/editorial/chatgpt-carbon-footprint/#:~:text=To%20illustrate%2C%20a%20single%20data,year%20than%20100%20US%20homes
14. https://www.networkworld.com/article/3454626/8-ways-to-prepare-your-data-center-for-ai-s-power-draw.html
15. https://gizmodo.com/chatgpt-ai-water-185000-gallons-training-nuclear-1850324249
16. https://www.eia.gov/environment/emissions/co2_vol_mass.php



ABOUT THE AUTHOR

Ihab Chaaban, P.Eng., MBA
Aeroderivatives Global Commercial Development Director,
Data Center, Grid Firming, Hydrogen & Hybrids – GE Vernova

Ihab Chaaban is a global technical and commercial development leader with over 30 years of experience in the Power Generation and Systems industry, 15 of which with GE, with deep domain expertise in the Energy, Utility and Electrical Power Businesses. Chaaban started his career as an Electrical Engineer and held several roles in Engineering, Sales and Services in the Caterpillar and Solar Turbines organizations via global assignments in the Middle East, Africa, Brazil, Canada and the USA.

Chaaban started his career with GE as an Application Engineering Lead for the aeroderivative global division followed by other roles in commercial operations leadership and proposal management for aeroderivative, Distributed Power and industrial frames gas turbines. Chaaban assumed the role of GE aeroderivatives global commercial development director to support the grid firming segment including Aero gas turbines in hybrid and green hydrogen operations with renewables assets and battery energy storage systems. Chaaban is a Licensed Professional Engineer in Ontario Canada, holds a B.Sc. in Electrical Engineering from Alexandria University, an MBA from Lansbridge University and speaks five languages (English, French, Arabic, Portuguese proficiently with a good command of Spanish).

gevernova.com

* Trademark of GE Vernova

GE VERNOVA

GEA35355 09/2023