Simplifying liquid cooling deployments with EkkoSense

AI compute growth is leading to some very real data center engineering challenges

Dr Stuart Redshaw
CTIO & Co-Founder
EkkoSense
stuart.redshaw@ekkosense.com

We're quickly moving towards a world where almost all organizations will have their own AI applications – some deployed tactically, while others will be used more strategically to remodel core processes and unlock new levels of efficiency.

McKinsey suggests that generative AI could change the whole 'anatomy of work' - unlocking productivity benefits of up to $4.4 trillion annually, while Elon Musk has suggested that the current pace of AI compute growth is like "Moore's Law on steroids".

Given the sheer pace of change, it's hardly surprising that increasing demands from the business to process GPU-intensive generative AI, training and inference workloads are now imposing huge challenges for todays' data center infrastructure and operations teams.

This is leading to some very real data center engineering challenges – particularly as analyst growth projections are only looking one way. Goldman Sachs, for example, estimates that data center power demand will grow by 160% by 2030, while New Street Research projects that AI-specific data center spend will grow 11x during the 2023-27 period.

Goldman Sachs estimates that data center power demand will grow by 160% by 2030.

How ready are data center operations teams to accommodate these solutions?

There's no doubt that growing demands to process GPU-intensive AI workloads are placing huge pressure on current data center infrastructure and operations.

However, the laws of physics don't change. Broadly, for every watt of electrical energy we put into a computer, we create one watt of heat through energy transfer.

For many years, as an industry, we've seen standard 1U/2U pizza box X86 server deployments operating in the 4-5kW per rack range – and the accepted cooling design approach has been around air-based cooling. This has primarily involved perimeter cooling units blowing air into a raised floor plenum to create a positive static pressure under the floor. Cold air is forced out of a floor grille in front of the rack which is then consumed by the fans in the IT equipment, with heat being removed by the device through conduction and convection. Hot air is exhausted at the back of the rack, with the heat rising into a ceiling plenum before finding its way back to the CRAC or CRAH.

Data center teams have worked to make this process a bit more efficient by adding hot and cold air separation techniques such as blanking panels and cold aisle containment. But, ultimately, there's a limit to the amount of heat that can be removed because of limits on the volumetric amount of air that can be moved through this type of architecture through the raised floor and floor grilles. That's why for higher density loads such as blade servers and HPC deployments that can require up to 10kW per rack or more, there's been a requirement to deploy enhanced air cooling strategies such as inrow or rear door cooling products. But, with AI loads forecast to grow to 60kW per rack and above, everything changes, air-cooling strategies alone won't be enough to remove the considerable increases in heat load.

Data center cooling has to evolve rapidly to keep pace.

With the processing of GPU-intensive workloads such as AI clearly set to generate more heat within data centers, operations teams need to think hard about their current cooling infrastructure and how it will need to evolve. While air-based cooling enhancements have helped, it's unlikely that this will be enough to support likely AI compute needs.

As AI compute loads grow dramatically with wider deployment, we'll also see ultra-high-density AI racks that can potentially require up to 100 kW of power for equipment and cooling that could be worth up to $10 million plus per rack. Not surprisingly, this growth in demand for more powerful server hardware is also seeing a parallel boom in the sales of liquid cooling systems. Analyst firm Omdia, for example, is now projecting $2bn liquid global cooling revenues by the end of 2024, with this rising to some $5 billion by 2028. With the wider deployment of ultra high-density AI racks, air cooling alone is clearly no longer enough to meet requirements the intense cooling requirements of this new generation of AI computing.

With AI loads forecast to grow to 60kW per rack, air-cooling strategies alone won't be enough to remove the considerable increases in heat load.

Understanding the different cooling options – and how they might work together

First, it's worthwhile noting that, despite all the excitement around liquid cooling, the technology itself is nothing new.

Liquid cooling has been around since Cray launched its X-MP supercomputers in the early 1980s – hence its 'bubbles' nickname, while a second wave of liquid cooling followed to support the introduction of blade servers by vendors such as HP some 15 years ago.

So what are the options facing data center management now? Let's consider the likely technical scenarios that today's data center operations teams typically face when considering evolving cooling requirements.

Standard data center rack deployments

3-5 kW per rack

Most of these IT systems have been running at between 3-5 kW per rack, supported by traditional air cooling, and already dealing with the challenge of increasing workloads at around 20% per year.

Initial AI workload deployments

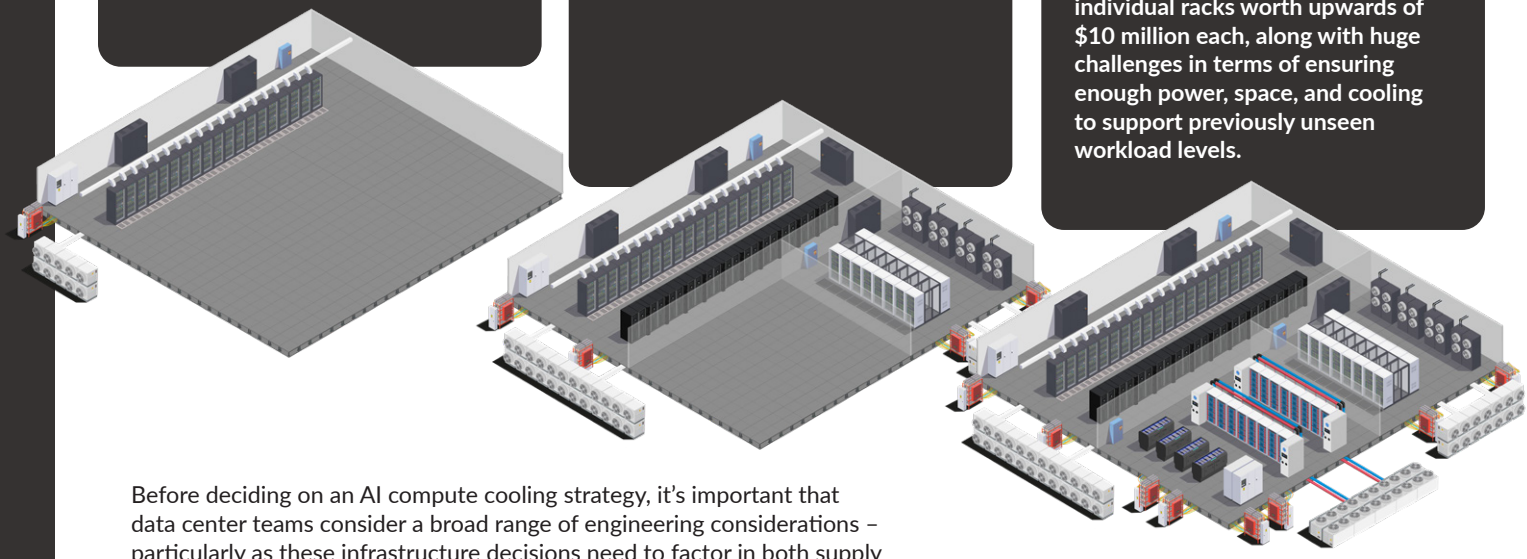
10 kW per rack

With high density AI workloads starting to hit 10 kW per rack, existing infrastructure is starting to be stretched and there's a requirement for increased cooling. Initially this will involve enhanced air-cooling approaches such as in-row or rear-door cooling.

Where are we heading?

125-150 kW per rack

As AI compute loads extend dramatically with wider and deeper deployments, we're starting to see ultra-high-density AI racks in place. These increasingly require 125-150 kW per rack; however, we're also starting to see high-performance AI factory vendors talking about a new generation of 250-500 kW racks. This combination of high-density compute and liquid cooling will see individual racks worth upwards of $10 million each, along with huge challenges in terms of ensuring enough power, space, and cooling to support previously unseen workload levels.



Before deciding on an AI compute cooling strategy, it's important that data center teams consider a broad range of engineering considerations – particularly as these infrastructure decisions need to factor in both supply chain realities as well as ongoing corporate ESG and sustainability goals.

Other factors that data center teams need to consider include:

Establishing the exact hybrid blend of air and liquid cooling technologies that are needed

Current and future plans to accommodate higher density AI racks within their increasing power and infrastructure requirements

Precise timings for AI compute workloads going live, and consideration of potential supply chain limitations around procurement

Potential liquid cooling limitations still need to be addressed

Ambitious annual growth statistics might suggest that liquid cooling is simply going to replace the hundreds of thousands of air cooling systems already deployed around the world.

However, there's a number of very practical reasons why this isn't likely to happen, including the fact that only the high power loads justify the cost and complexity of liquid-cooled solutions. There are still many lower-powered 'conventional' servers that are being deployed.

Of course immersion cooling can deliver great performance, but there are still potential concerns around oil spills, the lack of ability to make fibre connections, as well as issues with liquid interfering with the light interface. Some components and PCBs can degrade in the liquid cooling medium, and there are practical concerns around maintenance difficulties around the need for oil replacement and changing out fans, heat sinks and the thermal paste on chips.

Direct-to-chip (DTC) liquid cooling can also introduce issues such as the risk from fluid leaks, and the likelihood of a limited thermal buffer should liquid circulation fail – with potential associated resilience concerns, With direct-to-chip there is still the requirement for significant heat rejection to air from server components – such as power supplies - that are not connected to the DTC circuit.

Installing liquid cooling will also require the use of external heat rejection plant to get excess heat out of the building. Two considerations here include will existing air cooling systems accept the additional heat load, and whether further changes will be needed to air cooling systems to optimize heat transfer from liquid systems. This is often overlooked when it comes to immersion cooling in particular, and needs to be planned and costed into any direct liquid cooling upgrade projects. Data center teams also need to consider whether there are any opportunities for efficiency gains with higher facility water temperatures or by using alternative heat rejection techniques.

Some components and PCBs can degrade in the liquid cooling medium, and there are practical concerns around maintenance difficulties.

Air cooling and liquid cooling both have an important part to play

Given that it's not possible to run completely liquid-cooled data centers, the reality for most data center operators is that liquid cooling and air cooling will both have an important role to play in the cooling mix – most likely as part of an evolving hybrid cooling approach.

Air cooling certainly isn't going away, as data centers will largely still have to rely on their current air cooling infrastructure to support their existing IT workloads. Current thermal and cooling performance will also require continued optimization if they are to successfully unlock further capacity for additional IT loads. Once liquid cooling is deployed you will need to ramp it up and run at an optimum temperature – and backfill where needed with air cooling to create the most efficient hybrid model.

The transition to hybrid cooling will require careful management, with the procurement of high-density AI compute hardware and liquid cooling infrastructure only proving the starting point. Introducing liquid-cooled solutions adds a new dimension to capacity planning, with the placing of new loads now requiring consideration of space, electrical power, air cooling capacity, and liquid cooling capacity.

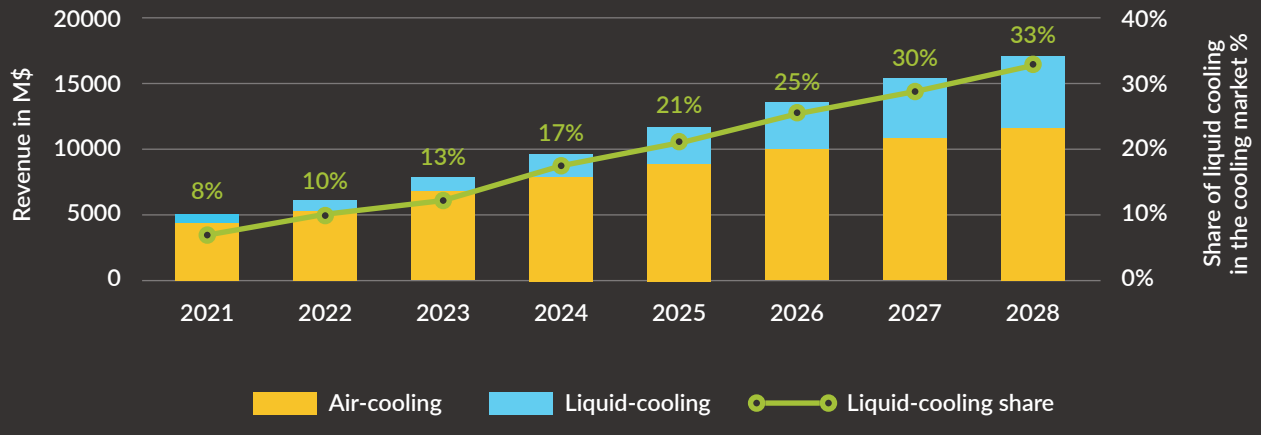
Any hybrid deployment will need to remain fully optimized, particularly as workloads continue to scale – making the requirement for absolute real-time white space visibility more important than ever.

Managing the transition to this kind of hybrid cooling approach will be challenging for data center teams looking to provide an effective cooling solution to support their potentially hundreds of millions of dollars investment in AI compute hardware and cooling infrastructure.

Before simply deploying this equipment, key questions need answering, including establishing the exact blend of air and liquid cooling technologies you'll need, and also recognising the complexity of managing the operation of a hybrid air cooling and liquid cooling approach within the same room. Investment at this scale can quickly lead to new engineering realities – increasing the need for absolute real-time white space visibility. Anything that can be done to dial down the stress levels for data center teams proves really important at this stage, particularly when it comes to understanding and managing hybrid cooling complexities.

Key questions need answering, including establishing the exact blend of air and liquid cooling technologies, and recognising the complexity of managing the operation of a hybrid air cooling and liquid cooling approach within the same room.

Liquid-cooling versus air-cooling forecast 2021-2028



Source: Omdia

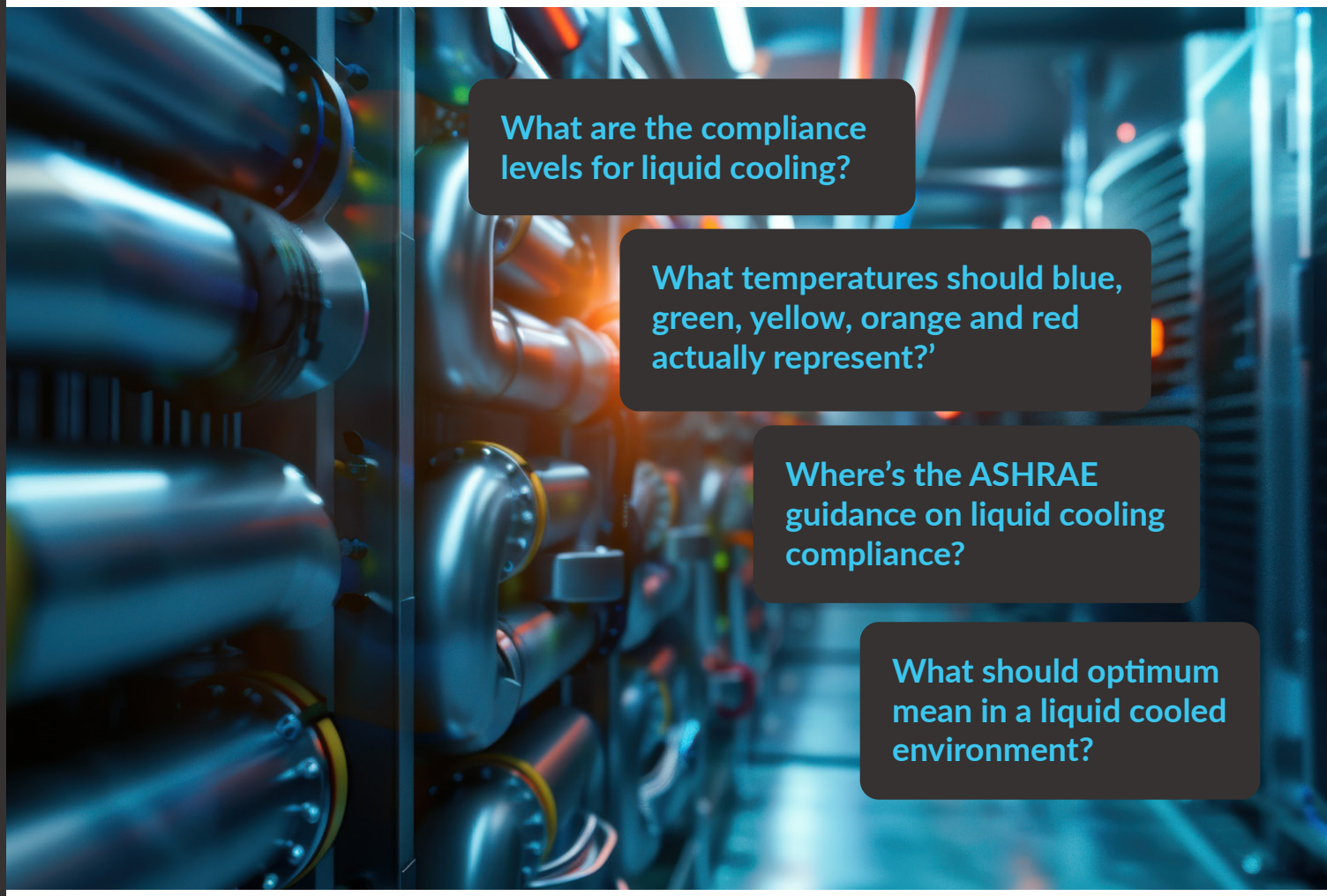
Understanding liquid cooling compliance metrics is key for optimization

Given the unprecedented value of organizational AI compute and hybrid cooling investments, IT and data center management simply can't afford to let it fail.

Business pressures also mean that generative AI, inference and learning applications will be running at 100% right from initial deployment, so it's essential that operations teams first understand that their AI compute will run safely, and also that it will continue to perform optimally with maximum workloads.

This presents a challenge. As the global leader in AI-powered data center optimization software, EkkoSense has worked hard to establish and track what optimal means for today's data center operators. When we monitor a data center's power, capacity and thermal performance we provide a clear 3D visualization with simple colors that illustrate data center performance in real-time. Essentially this takes a very complex infrastructure management challenge and simplifies it so that operations team members can focus on other added value activities such as compliance, ESG reporting, and unlocking further capacity.

So, whether it's air cooling, liquid cooling or a hybrid environment, the requirement for absolute real-time white space visibility only gets more important. That's been fine for traditional air-cooled data centers, but until now it's been hard to apply the same level of optimization simplicity to liquid cooled facilities. When data center management asks itself critical questions such as 'what are the compliance levels for liquid cooling?', 'what temperatures should blue, green, yellow, orange and red actually represent?', 'where's the ASHRAE guidance on liquid cooling compliance?', and 'what should optimum mean in a liquid cooled environment?', they will perhaps be shocked that there currently aren't any straightforward answers to these questions. That should worry them, particularly if they've just signed off on a number of $10m AI compute racks to support their CEO-driven AI initiative.



That's where EkkoSense can help

At EkkoSense we know that this is currently a key concern for data centre teams focused on AI compute deployments. We've spent the last ten years developing and refining an innovative AI-powered light touch solution for data center optimization that's used by many of the world's leading brands. However we also recognize that operations teams need the same level of absolute real-time white space visibility for their liquid side monitoring as they do for their existing air-side cooling.

That's why EkkoSense offers liquid cooling as part of its distinctive AI-powered EkkoSoft Critical optimization approach. Whether it's air-side or liquid-side cooling, EkkoSense measures the specific cooling duties of air cooling and liquid cooling equipment for real-time reporting within our 3D visualization and analytics software. AI compute operators will now have a clear indication of how their business-critical AI engines are performing thermally within their data centers – providing immediate re-assurance instead of uncertainty.

While EkkoSense is working to define clear thresholds for liquid cooling compliance, the industry also has work to do. Once we know what optimal actually means in a liquid cooling environment we can begin to establish headroom and what's safe in terms of optimization. However, with liquid-side monitoring established, the work on keeping AI compute deployments fully optimized in terms of power, capacity and thermal performance can really take off.

Hybrid cooling optimization with EkkoSense

Founded in 2014, EkkoSense is a rapidly-growing global SaaS company that's increasingly recognized as the smart choice for data center teams looking to improve their operational performance while removing power, capacity and thermal risks.

EkkoSense's combination of powerful 'what-if?' scenario simulations, low-cost air-side and liquid-side monitoring sensors, web-based 3D visualizations with analytics, AI-powered advisory and detection tools, and simple installation makes for light-touch deployments, minimal overheads and results within weeks. Typical ROI for an EkkoSense project is under a year, with project costs typically financed by cooling energy savings. Continuous data center optimization also helps to unlock further value as IT loads change.

Because the latest AI compute hardware and hybrid cooling infrastructure introduces new levels of complexity, EkkoSense deploys the power of AI to capture, visualize and analyse data center performance. EkkoSense's disruptive AI-powered optimization software draws on the power of over 50 billion machine learning datapoints to simplify and provide real-time insight into how this latest generation of data center equipment is performing.

Key components in the EkkoSense data center optimization portfolio include:

Monitoring

Ultra-low cost wireless sensors for liquid and air cooled infrastructure complemented by flexible integration offerings

EkkoAir

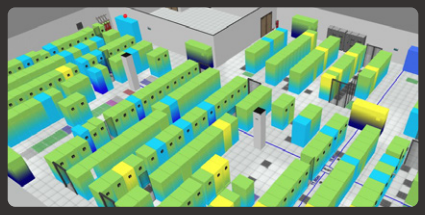
Provides real-time wireless tracking of air-side cooling duty loads for CRAC/AHU units

EkkoFlow

For Liquid Cooling – cost-effective measurement of liquid cooling duty loads with measurement of chilled water liquid cooling flows and temperatures to begin optimising performance in dense AI compute environments

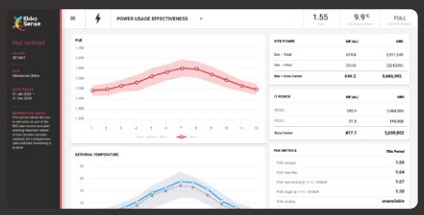
EkkoSim Simulation

End-to-end data center infrastructure modeling and simulation tool for 'what if?' scenario planning and cooling configuration modeling. Draws together capacities, operational parameters, resilience, energy flows, heat transfer pathways and capacity changes over time to simplify hybrid cooling planning



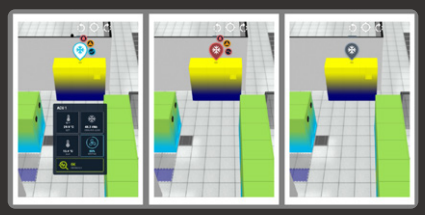
EkkoSoft Critical

3D visualization and AI analytics software to track performance of air, liquid and hybrid data center cooling architectures with 3D objects for digital twin



Compliance Automation

Automated capacity and ESG reporting to free up valuable data center operations resource to focus on added value activities. ESG Reporting, for example, is an embedded solution to collect, trend, analyse and report on ISO 30134 ESG Reporting requirements such as PUE, CER, CUE and WUE



AI-powered Cooling Advisory and Anomaly Detection tools

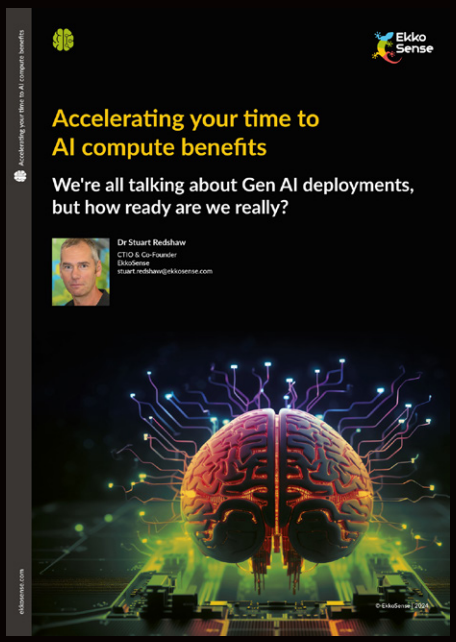
To help operations teams unlock cooling energy savings from their systems, and identify M&E cooling equipment performance anomalies ahead of a potential equipment failure

Simplifying liquid cooling deployments with EkkoSense

Contact EkkoSense to continue the conversation and discuss your liquid cooling challenges.

Click or scan

Have you read...



Click or scan

Ekko Sense

UK Headquarters: +44 (0) 115 678 1234
North America: 1-833-921-3335
Germany: +49 89262025276
Australia: +61 2 8358 0031
info@ekkosense.com www.ekkosense.com

DCS AWARDS 2024 WINNER
BusinessGreen UK Green Business Awards 2024 WINNER
DATACLOUD GLOBAL AWARDS 2024 WINNER
The CLOUD Awards WINNER 2023-24

ekkosense.com