How data center digital twin simulations are set to derisk the transition to AI compute

With AI remodeling today's data centers, IT teams must focus on derisking its impact. A new generation of digital twin simulations will be critical to their success.
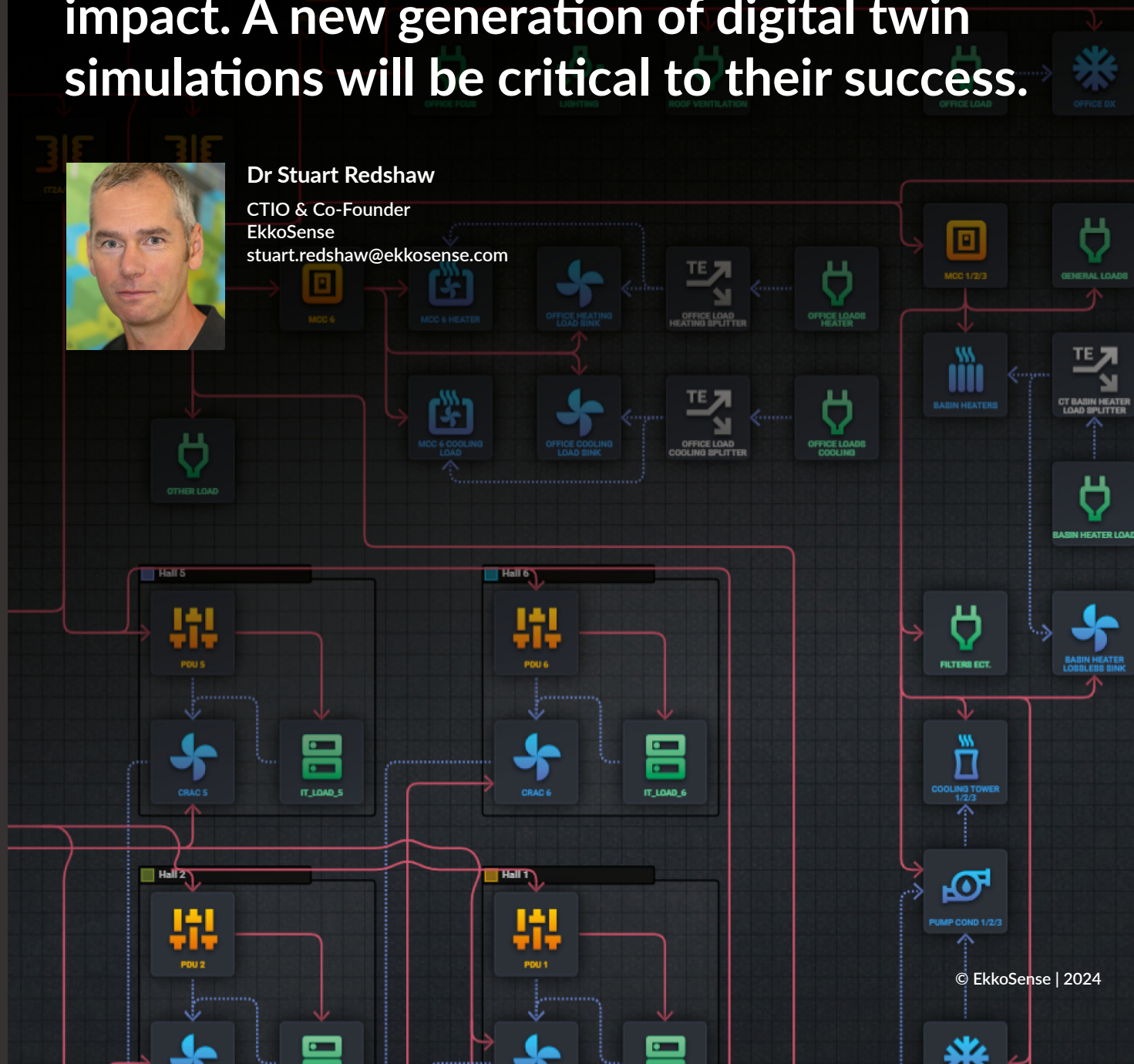
Dr Stuart Redshaw

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

Introduction

The era of gradually evolving data center engineering is at an end as AI compute's sheer power and enormous heatloads start to land. Decisions are needed now. Organizations are moving towards the deployment phase of their AI strategies, and they simply can't afford to get things wrong.

This time round things are very different. Traditionally, adding new data center workloads has largely been straightforward, almost transactional – with additional racks generating a fairly consistent additional heatload. Adapting to this was routine, requiring an evolved engineering approach.

However, with AI compute things are very different as the latest hardware generates heatloads that are orders of magnitude more powerful than current systems. Clearly hybridized liquid cooling will have a vital role to play, but with IT equipment workloads not yet fully understood, CFOs and CIOs set on spending potentially hundreds of millions on new AI compute deployments require much greater levels of re-assurance.

The good news is that an entirely new class of data center digital twin modeling and simulation software – EkkoSIM from EkkoSense - is now available to help organizations address the immediate Big SIM requirements of their AI compute projects.

In this white paper we set out the pressing requirement for this kind of solution from IT leaders and data center management under pressure to support the latest high-density AI loads and evolving cooling approaches. We then detail how EkkoSIM is the first data center digital twin modeling and simulation solution to address this need.
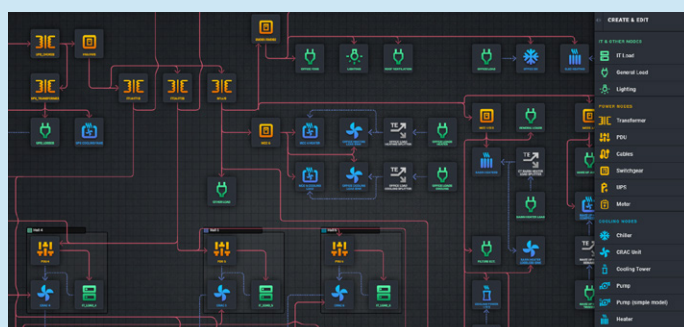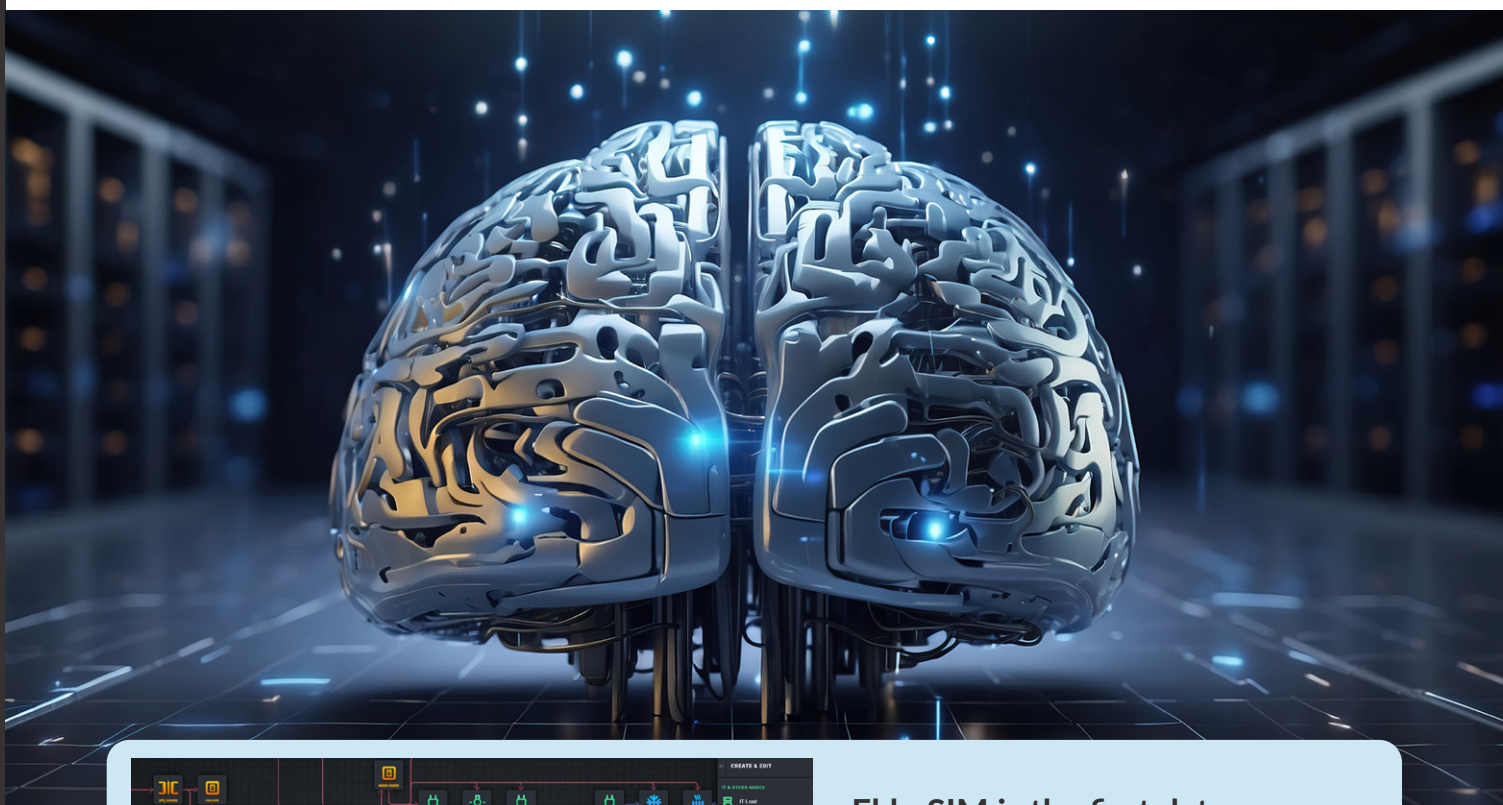
EkkoSIM is the first data center digital twin modeling and simulation solution to help organizations address the immediate Big SIM requirements of their AI compute projects.

Keeping pace with AI's rapid acceleration

It's hard to open any newsfeed today without finding yet more news about why AI is changing everything, and how organizations across the world are positioning themselves to best take advantage.

> **"AI has emerged as the most potent technology since the inception of the Internet era and is capable of propelling worldwide productivity growth."**
>
> Satya Nadella, Microsoft CEO

> **"Generative AI may be the largest technology transformation since the cloud, and perhaps since the internet… the amount of societal and business benefit from the solutions that will be possible will astound us all."**
>
> Andy Jessy, Amazon Chief Executive

> **"the consequences of AI will be extraordinary and possibly as transformational as some of the major technological inventions of the past several hundred years: think the printing press, the steam engine, electricity, computing and the Internet…"**
>
> Jamie Dimon, JP Morgan Chase & Co Chairman & CEO

> **"the table stakes for being competitive in AI are at least several billion dollars per year."**
>
> Elon Musk

Less than a year ago a Goldman Sachs Economic Research report suggested that global AI investments could hit $200 billion by 2025. That's already looking like a low bar projection, with the last month alone seeing Microsoft and OpenAI unveiling a $100 billion plan to build Stargate – an AI supercomputer to run GenAI workloads; Abu Dhabi setting up a new $100 billion AI and semiconductor tech fund; and Saudi Arabia's Public Investment Fund sitting on $40 billion for AI investments.

Then there's all the IT and business leaders competing with each other this year to fuel the AI frenzy. According to Microsoft's CEO Satya Nadella: "AI has emerged as the most potent technology since the inception of the Internet era and is capable of propelling worldwide productivity growth." Amazon's Chief Executive Andy Jassy says: "Generative AI may be the largest technology transformation since the cloud, and perhaps since the internet… the amount of societal and business benefit from the solutions that will be possible will astound us all." While Jamie Dimon, Chairman & CEO at JP Morgan Chase & Co. writes that "the consequences of AI will be extraordinary and possibly as transformational as some of the major technological inventions of the past several hundred years: think the printing press, the steam engine, electricity, computing and the Internet…"

Elon Musk, who is building a $1 billion supercomputer for Tesla, has also pointed out that "the table stakes for being competitive in AI are at least several billion dollars per year". He's recently described the pace of AI compute growth as like Moore's Law on steroids – suggesting that "the artificial intelligence compute now coming online appears to be increasing by a factor of 10 every six months… the chip rush is bigger than any gold rush that's ever existed"

While analyst firms such as McKinsey acknowledge the strong business case for the deployment of AI applications, they also recognize that establishing a clear route towards unlocking GenAI value can be challenging. Google certainly found that when its new Gemini AI chatbot recently returned a series of disturbing results. Markets reacted negatively, with parent company Alphabet's stock price dropping by 5% in a day.

Despite the clear risk involved, major organizations simply can't afford not to get involved. Everyone wants a piece of the AI action, and there's a real worry that if a business isn't developing or launching some form of AI-powered proposition then they could get left behind.

Of course the deployment of a new generation of AI applications represents a significant opportunity for businesses to differentiate and gain market advantage. There's a huge pressure to get things right, so it's critical that your organization's IT infrastructure and data centers are ready to support your AI applications, and that you've made all the right choices in terms of preparation.

How should data centers adapt to accelerating AI adoption?

With the processing of GPU-intensive AI compute workloads clearly set to generate more heat within data centers, operations teams need to look hard at their current infrastructure and how it will need to evolve. Data center operations teams might think – like every time up until now – that they have a massive site that can easily absorb new loads. But this time they can't be so sure.

At EkkoSense we're already seeing AI-driven heat loads showing significant dynamic variability – it's all very different from the reassuring certainty of running traditional enterprise workloads.

Operations teams also need to think hard about anticipated AI workloads, their current data center infrastructure and how it will need to change in terms of cooling. With high density workloads now typically running at 30 kW per rack – and some even hitting 60 kW per rack – it's true that the legacy 3-5 kW per rack data center supported by traditional air cooling is starting to look like it's an infrastructure on borrowed time.

While we've seen air cooling able to support up to around 30 kW per rack, you can sense it's starting to hit the limits of what's achievable. At EkkoSense we've been able to apply our AI-powered 3D visualization and analytics software to remove thermal risk and optimize data center cooling capacity even at these densities, but CIOs and their operations teams know that AI's remodeling of the data center is well under way and shows no signs of slowing down. There's a need to adjust to AI's new engineering realities.

In terms of cooling technologies, air cooling isn't going away any time soon. Data centers are still going to need their existing cooling infrastructure to support their extensive workload commitments. This is proven out even in operational AI data centers with direct-to-chip liquid cooling solutions that are still emitting significant quantities of heat to air.

Data center teams will need to continue to take the time to optimize their existing thermal and cooling performance if they're to unlock capacity for additional IT loads. In many cases the answer to future heat rejection will be to combine both air and liquid cooling in a hybrid approach. Key questions to consider here include the exact blend of air and liquid cooling technologies you'll need, and a clear insight into your plans to accommodate higher density AI racks with their greater power and infrastructure requirements.

Optimizing legacy air cooling performance

Among the many existing data centers that are now being adapted for AI workloads, EkkoSense advises that data center teams first ensure that their legacy air cooling performance is as optimized as it can be to support current loads. Only then should they move quickly to supplement their existing system with a direct-to-chip liquid cooling hybrid as a matter of urgency.

This may take time, but it needs to align with the delivery timeline of the AI compute hardware. Remember, the manufacturers of the AI hardware will be asking for a cooling option before orders can be placed. Once the AI hardware is deployed, you need to be fully prepared so that your AI compute can be operated at 100% from day one. Both the air and liquid cooling solutions need to ramp up and run perfectly. However, even at this pace, there remains room for optimization and a continued focus on operational efficiency - so making sure your digital twin models are fed with live data to verify the design is also vital.

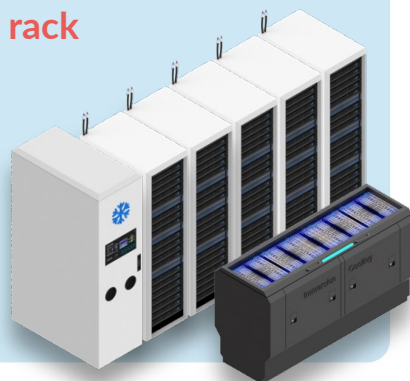## Current maximum enhanced air-cooling capability

30 kW per rack

Enhanced Air Cooling

Current typical high density workloads

60 kW per rack

Liquid Cooling
Direct-to-chip
Immersion
Cooling

Need for absolute real-time white space visibility

At EkkoSense we're already seeing the transition towards AI starting to impact data center infrastructure - both through the sheer volume of compute cycles that developers are now running on AI platforms, as well as infrastructure investment at the many sites that we're busy optimizing around the world.

Investment at this scale can quickly lead to new engineering realities – sometimes the sheer maths involved can be frightening. And the infrastructure decisions that CFOs, CIOs, CEOs, and COOs are taking now have to be right – getting things wrong at this stage has the potential to apply real constraints to an organization's AI compute strategy.

Effective data center optimization an important first stage

So anything that can be done now to reduce noise and stress levels for data center teams is clearly important. Top of the list is knowing exactly what's likely to happen from an infrastructure and engineering perspective when organizations launch their AI services.

This requires absolute real-time infrastructure visibility. The ability to look across your whole data center estate and be able to check that the assumptions that you're making on cooling, power and capacity are actually standing up. And, if they're not, getting insights as early as possible so they can sort things in readiness for scaling up AI compute and the volume of GenAI applications being delivered.

Applying best practice AI-powered optimization is an important first step in getting your existing infrastructure operating as efficiently as it can. At EkkoSense we believe that machine learning and AI are changing the game for data center optimization, yet – despite the current global acceleration towards an entirely new generation of AI compute capabilities – very few software vendors currently apply machine learning and AI to support their critical data center optimization activities.

Having access to this kind of insight is increasingly important, particularly the ability to understand exactly where you stand in terms of your current cooling, power and capacity needs. However, given the acceleration towards AI computing, there are now some more profound engineering questions that CEOs, CFOs and CIOs need to be answered, including:

> **How can we predict and maintain optimal performance across our full infrastructure?**

> **Where are the infrastructure capacity bottlenecks for increasing IT loads?**

> **What spaces can we use for high-density loads such as AI?**

> **What equipment do we need to replace – and when?**

> **What would be our best cooling strategy?**

Decisions are needed now, and organizations simply can't afford to get things wrong. These are questions that need very precise answers – and quickly.
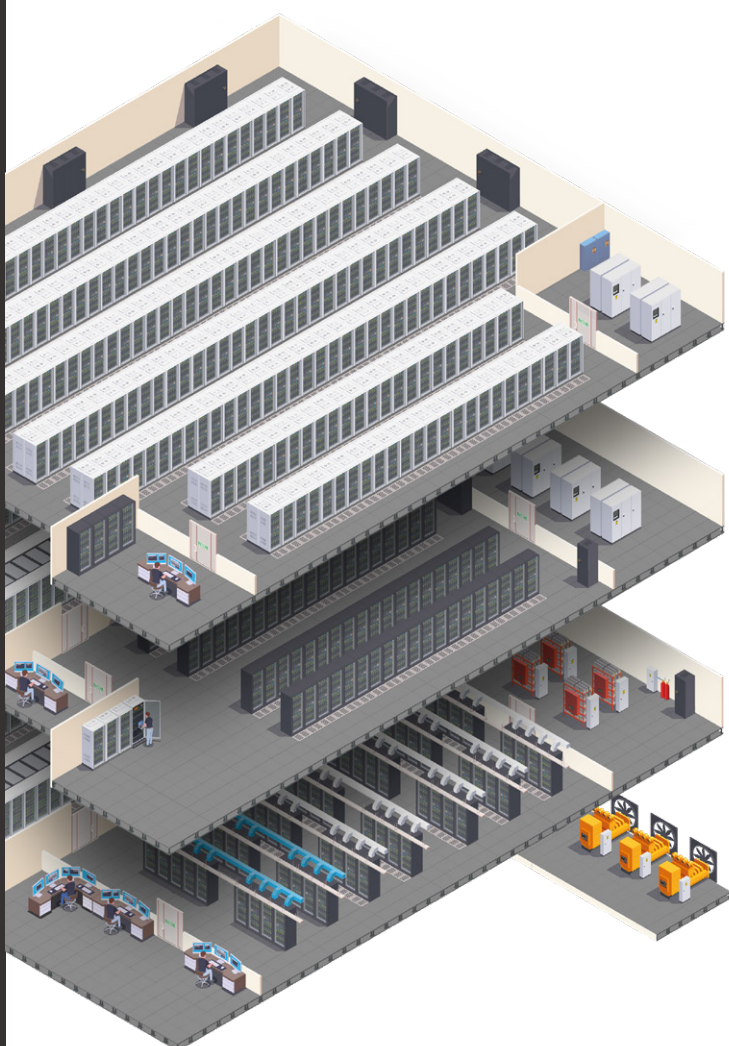
Unlocking the innovation answers you need with data center digital twins for simulation and modeling

The deployment of a new generation of AI applications represents a significant opportunity for businesses to differentiate and gain market advantage. Consequently, there's huge pressure to get things right! It is critical that your organization's IT infrastructure and data centers are ready to support GenAI deployments, and that you've made all the right choices in terms of preparation.

So, for data center teams tasked with responding to the rapid pace of change triggered by continued AI innovation, what's really needed is straight answers to some basic questions. That's particularly the case when it comes to planning for high-density AI workloads. These can already exceed 30 kW per rack – and will keep on rising. Everyone knows data center infrastructures will inevitably be stretched, but very few operations team know by how much exactly.

Given the scale and cost of this challenge it's simply not practical for IT teams to build their own full-scale test facilities, particularly when there are so many questions and variables involved that need addressing fast. So, whether you're a CFO validating potential AI investments, a CIO trying to identify capacity bottlenecks for increasing IT loads, or a data center manager looking to specify the most appropriate cooling strategy, some way of simulating potential AI data center scenarios before committing to major infrastructure investments is essential.

> **At EkkoSense we're closely aligned with the Industry 4.0 approach where the physical data center and grey space world is fully digitized with infrastructure connected as intelligent assets within clearly defined cooling, power and capacity process scenarios.**

Enabling modeling and simulation across end-to-end data center infrastructures

The whole point of an innovation tool for data centers is to allow businesses to simulate change quickly, note the broad effects and results, and draw whatever conclusions they need to.

At EkkoSense we're closely aligned with the Industry 4.0 approach where the physical data center and grey space world is fully digitized with infrastructure connected as intelligent assets within clearly defined cooling, power and capacity process scenarios. EkkoSense's distinctive combination of Internet of Things, machine learning, AI analytics, 3D visualization and digital twin representations combines to enable the delivery of an entirely new class of **data center digital twin simulation and modeling** capabilities.

A key element of this approach has been to provide organizations with comprehensive modeling and simulation across their full, end-to-end data center and grey space infrastructure. A digital twin data center that extends beyond the white space, enables CIOs and their IT Teams to respond to their AI infrastructure concerns by carrying out 'what-if' analysis queries, tracking the performance of their current capital investments, simulating extensions to existing capacity, or even planning entire new builds.

Drawing on first principles maths and physics models

The target has been to rapidly simulate data center change in response to AI deployment plans, drawing on first principles maths and physics models so that businesses can create precise cooling, power and capacity simulations with accuracy levels running at 99% plus. However, this kind of AI sandbox simulation can't currently deliver if too much AI compute processing, cooling, power and capacity data is being dialled in. While it would be possible to create data center digital twins that simulated every single switch position, setpoint and power circuit, the models would have to be unbelievably sophisticated to get every detail exactly right.

That's why in building EkkoSIM the focus has been on creating a Big SIM innovation tool for data centers – one that uniquely enables people to ask questions about rapidly evolving issues such as ultra-high density AI compute deployments, but that doesn't overlook all the critical associated heat transfer, thermodynamics, energy and cost components that need to underpin every decision.

EkkoSIM – the first digital twin solution to provide comprehensive modeling and simulation across full end-to-end data center infrastructures

EkkoSIM helps data center teams to zero in on the right decisions by creating, testing and analyzing the results of multiple data center 'what if?' scenarios.

Thanks to the accessibility of EkkoSIM's data center digital twin software, IT and operations teams can accurately analyze and predict the most suitable designs, map metered data against predictive models, and triage those data center assets that would benefit most from configuration assessments.

Critically, this new approach also equips customers with a rapid simulation capability that supports changes to their IT load so they can understand exactly how increased GenAI workloads might impact their cooling, power and capacity infrastructure.

Visibility at one of the most pivotal points in data center history

The value of EkkoSense's digital twin approach for data center modeling and simulation is that it provides CEOs, CFOs, CIOs and senior data center management teams with the visibility they need at one of the most pivotal points in data center history. Of course, it's still underpinned by really complex statistical mathematics and the power the EkkoSense data lake – but that's all hidden within the engineering.

What CIOs and CFOs and data center management get to see though, is whether it's possible to take their infrastructure forward so that they can bring in the capacity needed to deliver AI compute at scale within their business timeframe. So, when they're asked: 'are we ready for that?', they will – for the first time – be able to answer with confidence. EkkoSIM unlocks the data center and grey space answers they need.

EkkoSim 1.0
MODEL | PREDICT | DEPLOY

> **When CIOs, CFOs and data center managers are asked: 'are we ready for that?', they will – for the first time – be able to answer with confidence. EkkoSIM unlocks the data center and grey space answers they need.**

# Zeroing in on the right data center infrastructure decisions with EkkoSIM

EkkoSense is focused on helping organizations to navigate the inevitable pressures associated with delivering support for the latest ultra-high-density AI compute workloads.

Leveraging its powerful Digital Twin technology, the company's new EkkoSIM solution provides organizations with comprehensive modeling and simulation across full, end-to-end data center and grey space infrastructures.

EkkoSIM supports unlimited simulation scenarios, with examples including the addition of new plant; increased IT loads; failover scenarios; reconfiguration of cooling systems, and viability of new power distribution systems. Other simulation options include quick simulations, graphical analysis, configuration of electrical and mechanical systems, alternative cooling configurations for chilled water, split DX units, air-cooled chillers, liquid cooling IT racks and air economizers, as well as N+R and 2N+R resilience testing.

The EkkoSIM interface clearly displays the capacity for each unique asset class based on the provisioned (maximum) and utilized (existing) site loads. EkkoSIM also integrates with the award-winning EkkoSoft Critical 3D visualization and analytics platform, drawing on EkkoSense's 50 billion point data lake to provide live Actual vs. Expected performance data in order to review operation performance against OEM specifications.

Key EkkoSIM benefits:
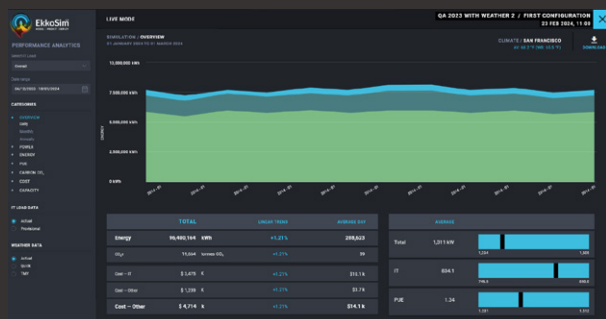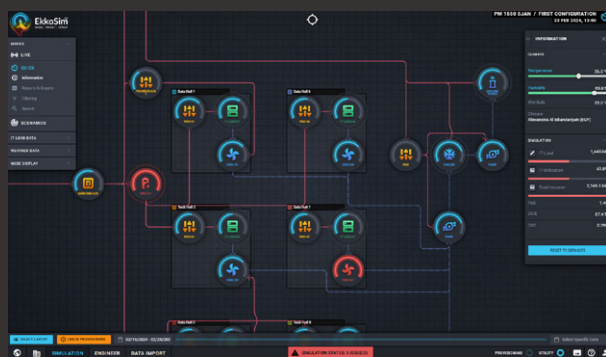
Reduce capital investments

By accurately analyzing and predicting the most suitable designs

Decrease operating expenses

By continually analyzing metered data against predictive models

Analyze the performance of critical capital purchases

**By triaging assets that would most benefit from configuration assessments**

Track Total Cost of Ownership (TCO)

Calculate data center equipment and power TCO

EkkoSIM provides organizations with comprehensive modeling and simulation across full, end-to-end data center and grey space infrastructures.
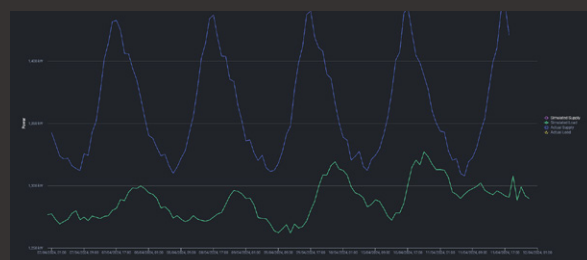
Quick and Advanced scenario generation

## EkkoSIM offers a broad range of additional simulation options, including:
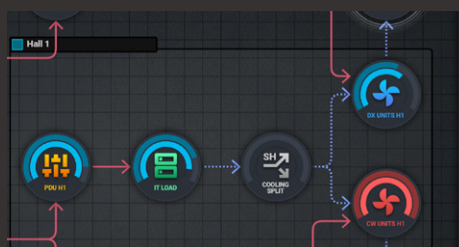
Quick Simulations

Using Quick SIM to make changes to the IT load, adjust the number of assets or test site performance under different conditions
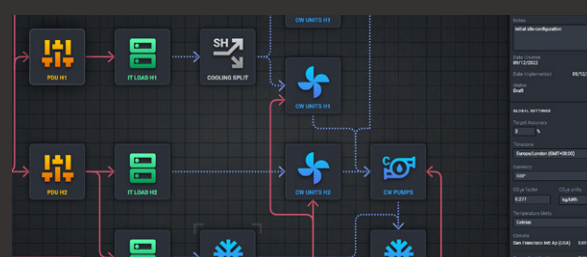

Performance Analysis

EkkoSIM can run these for periods up to five years, with KPI analysis including energy usage, seasonal PUE, CO2 emissions, costs and capacity
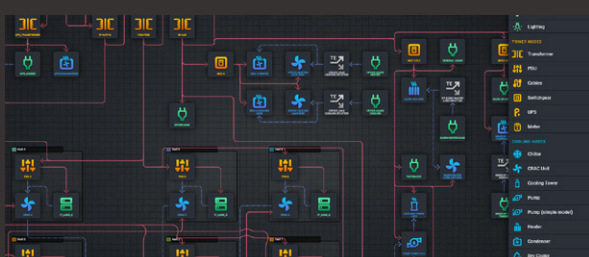

Configuration

Created for electrical and mechanical systems based on performance design data from manufacturers

Alternative Cooling Configurations

Simulating cooling configurations such as air and water-cooled chillers, liquid-cooled IT racks, chilled water, split DX units, air-cooled chillers, liquid cooling IT racks, immersion cooling and air economizers
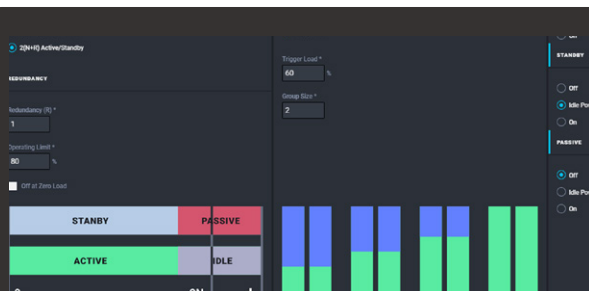
Resilience Testing

Supporting N+R and 2N+R resilience testing for both shared or staged loading, with power and cooling distribution circuits able to split and combine to simulate performance under failure scenarios

EkkoSIM also integrates with the award-winning EkkoSoft Critical 3D visualization and analytics platform, drawing on EkkoSense's 50 billion point data lake to provide live Actual vs. Expected performance data in order to review operation performance against OEM specifications.

Contact EkkoSense to continue the conversation and discuss your AI business challenges.
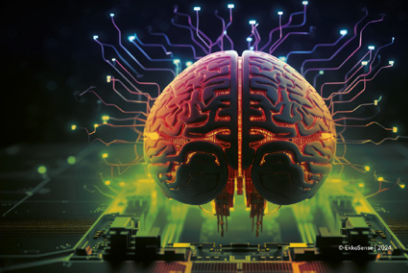

Click or scan

Have you read...

Click or scan

Click or scan