DCS

DATACENTRE SOLUTIONS

DEVELOPING DIGITAL INFRASTRUCTURE IN A HYBRID WORLD

ISSUE II 2024

HTTPS://DATACENTRE.SOLUTIONS

FIVE DATA CENTRE PREDICTIONS FOR 2024

Cloud | Design + Optimisation | Energy Management | Infrastructure + Asset Management
Hosting + Colocation | Infrastructure + Hardware | Power & Cooling

AN ANGEL BUSINESS COMMUNICATIONS PUBLICATION

centiel

continuous power availability

StratusPower™ Helps data centres achieve net-zero targets

Green UPS | DARA TECH | SWISS MADE

- Powered by Centiel's Distributed Active-Redundant Architecture (DARA) for minimizing downtime
- 9-nines availability (99.9999999%)
- 30 years design life
- Built on proven semiconductor technology for increased reliability

StratusPower UPS goes beyond power with a commitment to peace of mind and operational excellence.

Experience the future of data centres today!

www.centiel.com

VIEWPOINT

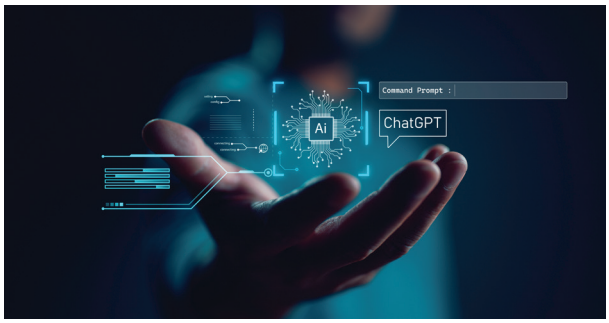
By Phil Alsop, Editor

All those in favour say AI

AI'S IMPACT on the data centre is twofold. The demand for AI across all industry sectors means new levels of compute are required, and this means power-hungry GPUs, more dense racks, more cooling required, faster networks and storage... all of which require a new type of data centre. The hyperscalers and colos are busy building out these new facilities – many of which do not need to be in the main FLAPD markets, but can be built near to where there is available power.

Legacy facilities are unlikely to be capable of housing AI-intensive workloads, but that isn't stopping many organisations looking at carrying out the necessary upgrades. Or, just maybe they are accepting that older data centres can best cope with the 'traditional' workloads and leave the AI ones to the new builds. So, the IT required to run AI applications is transforming the way in which data centres are designed, built and operated.

Secondly, AI is being used inside the data centre to help improve and optimise data centre performance – both the day to day operations and longer term maintenance and planning. The potential is enormous – indeed, it is already being realised in many data centres as smart new software solutions leverage AI and Machine Learning to deliver truly efficient performance. The 'holy grail' of the lights out data centre is closer than ever before.



It's with great pleasure that I recommend to you our major AI focus in this issue of DCS. Hopefully, the combination of articles will provide plenty of information as to the ways in which AI will impact the data centre. Yes, getting to grips with the technology is a significant challenge, but, primarily, it's a great double win-win opportunity for data centre owners and operators.

Providing the right environment for AI workloads is going to be big business – although a note of caution as per the Uptime Intelligence article, which suggests that there's something of an AI bottleneck for now at least. And using AI to continue to improve data centre performance and efficiency is a win for any organisation wanting to eliminate unnecessary spending and also to meet its sustainability goals.

All those in favour of AI, say aye!

CONTENTS

ISSUE II 2024

DCS DATACENTRE SOLUTIONS

The data centre trade association

DCA News

NEWS

DCS DATACENTRE SOLUTIONS

Editor
Philip Alsop
+44 (0)7786 084559
philip.alsop@angelbc.com

Senior B2B Event & Media Executive
Mark Hinds
+44 (0)2476 7189701
mark.hinds@angelbc.com

Sales Manager
Peter Davies
+44 (0)2476 718970
peter.davies@angelbc.com

Director of Logistics
Sharon Cowley
+44 (0)1923 690200
sharon.cowley@angelbc.com

Design & Production Manager
Mitch Gaynor
+44 (0)1923 690214
mitch.gaynor@angelbc.com

Marketing & Logistics Executive
Eve O'Sullivan
+44 (0)2476 823 123
eve.osullivan@angelbc.com

Publisher
Jackie Cannon
+44 (0)1923 690215
jackie.cannon@angelbc.com

Circulation & Subscriptions
+44 (0)2476 718970 circ@angelbc.com

Directors
Sukhi Bhadal: Chief Executive Officer
Scott Adams: Chief Technical Officer
E: info@angelbc.com

Published by: Angel Business Communications Ltd, 6 Bow Court, Burnsall Road, Coventry CV5 6SP. UK
T: +44 (0)2476 718970 E: info@angelbc.com

Angel BUSINESS COMMUNICATIONS

Datacentre Solutions is published 6 times a year on a controlled circulation basis in Europe, Middle East and Africa only. Subscription rates on request. All information herein is believed to be correct at time of going to press. The publisher does not accept responsibility for any errors and omissions. The views expressed in this publication are not necessarily those of the publisher. Every effort has been made to obtain copyright permission for the material contained in this publication. Angel Business Communications Ltd will be happy to acknowledge any copyright oversights in a subsequent issue of the publication.

The paper used within this magazine is produced by chain of custody certified manufacturers, guaranteeing sustainable sourcing. ISSN 2756-1143 (Online)

Growth of AI creates 'unprecedented' demand for global data centres

Increased need for data storage prompts real estate investors and developers to uncover additional, sustainable sources of energy.

AS THE global economy continues to rapidly adopt Artificial Intelligence (AI), infrastructure to support these systems must keep pace. Consumers and businesses are expected to generate twice as much data in the next five years as all the data created over the past 10 years. This growth presents both an opportunity and a challenge for real estate investors, developers and operators. JLL's Data Centers 2024 Global Outlook explores how data centers need to be designed, operated and sourced to meet the evolving needs of the global economy.

With the growing demands of AI, data center storage capacity is expected to grow from 10.1 zettabytes (ZB) in 2023 to 21.0 ZB in 2027, for a five-year compound annual growth rate of 18.5%[1]. Not only will this increased storage generate a need for more data centers, but generative AI's greater energy requirements – ranging from 300 to 500+ megawatts – will also require more energy efficient designs and locations. The need for more power will require data center operators to increase efficiency and work with local governments to find sustainable energy sources to support data center needs.

"As the data centre industry grapples with power challenges and the urgent need for sustainable energy, strategic site selection becomes paramount in ensuring operational scalability and meeting environmental goals," said Jonathan Kinsey, EMEA Lead and Global Chair, Data Centre Solutions, JLL. "In many cases, existing grid infrastructure will struggle to support the global shift to electrification and the expansion of critical digital infrastructure, making it increasingly important for real estate professionals and developers to work hand in hand with partners to secure adequate future power."

Sustainable Data Center Design and Operations Solutions

AI-specialized data centers look different than conventional facilities and may require operators to plan, design and allocate power resources based on the type of data processed or stage of generative AI development. As the amount of computing equipment installed and operated is expected to continue increasing with AI demand, heat generation will surpass current standards.

Since cooling typically accounts for roughly 40% of an average data center's electricity use, operators are shifting from traditional air-based cooling methods to liquid cooling. Providers have shown that liquid cooling boasts significant power reductions – as high as 90% – while improving capability and space requirements.

"In addition to location and design considerations, data center operators are starting to explore alternative power sourcing strategies for onsite power generation including small modular reactors (SMRs), hydrogen fuel cells and natural gas." said Andy Cvengros, Managing Director, U.S. Data Center Markets, JLL. "With power grids becoming effectively tapped out and transformers having more than three-year lead times, operators will need to innovate."

Global Investment in Data Centers and Power

To support these requirements, critical changes need to be made across the globe to increase power usage.

- In Europe, one-third of the grid infrastructure is over 40 years old, requiring an estimated €584 billion of investment by 2030 to meet the European Union's green goals.
- In the United States, meeting energy transition goals to upgrade the grid and feed more renewable energy into the power supply will require an estimated $2 trillion.
- Data centers' rapid growth is also putting pressure on limited energy resources in many countries. In Singapore, for example, the government enacted a moratorium to temporarily halt construction in certain regions to carefully review new data center proposals and ensure alignment with the country's sustainability goals.

The global energy conundrum presents both opportunities and challenges to commercial real estate leaders with a stake in the data center sector. Generative AI will continue to fuel demand for specialized and redesigned data centers, and developers and operators who can provide sustainable computing power will reap the rewards of the data-intense digital economy.

Eco-digital economy expected to double in the next five years to almost $33 trillion

Nearly eight in ten organizations say they are experiencing a dual transition towards a more digital and sustainable world, yet they have only harnessed around 25% of the overarching potential of mainstream tech.

THE UNTAPPED potential of digital technologies is vast, and the eco-digital economy, driven by digital and sustainability, is expected to double by 2028. That's according to the Capgemini Research Institute's latest report, 'The Eco-Digital Era™: The dual transition to a sustainable and digital economy' developed in collaboration with the Digital Value Lab at the Digital Data and Design Institute at Harvard.

Implementing digital technologies has enabled organizations to reduce their energy consumption by almost a quarter and delivered a 21% reduction in greenhouse gas (GHG) emissions in the past five years, cites the report. In this new era of a dual transition to an eco-digital economy that delivers not only economic value, but also environmental and social value, the scaling up of digital adoption will propel economic growth with sustainability at its core.

More collaborative and platform-driven than ever before, this eco-digital era™ is giving rise to new business models and revenue streams, as well as enhanced cost efficiencies, all driven by data utilization, cloud technology, collaborative ecosystems, and connected products and services. According to the report, seven in 10 organizations agree that digitally driven business models will become a key contributor of revenue growth in the next three to five years. Furthermore, 60% expect digitally driven business models to generate more revenue than their traditional business models.
"In the eco-digital era, there is greater exploration of digital technologies' value to business – for instance by the scaling of data and cloud, and by having digital technologies play a crucial role in achieving sustainability goals. There is also a fast evolution of emerging tech such as generative AI and synthetic biology, and greater collaboration giving rise to digital ecosystems," comments Dr. Suraj Srinivasan, Philip J. Stomberg, Professor of Business Administration at Harvard Business School and Head of the Digital Value Lab at the Digital Data and Design Institute at Harvard. "This shift is truly fundamental, cross-sectoral and global in nature. One of the biggest questions that organizations have to address and manage, as they scale, is knowing what to centralize and what to decentralize in terms of platform architecture, and most importantly, data governance."

Mainstream technologies at scale set to deliver most value

Investment in digital transformation – from scaling-up mainstream technologies and implementing cybersecurity measures, to reskilling the workforce and automating business processes - is expected to result in the most significant returns over the next five years, from 4% at present to 14% in 2028.

According to the report, around half of organizations (48%) are either at the planning stage or actively developing strategies to harness the potential of emerging technologies such as edge computing and the much-hyped generative AI. However, it is the mainstream technologies such as data and analytics and cloud at scale that organizations believe will deliver the most powerful business benefits over the next five years.

"The eco-digital economy is unlike anything that has come before it, and society has harnessed only a fraction of the overarching potential that mainstream technologies such as cloud, AI, and automation hold," said Fernando Alvarez, Chief Strategy and

Development Officer at Capgemini and Group Executive Board member. "Organizations will need to leverage focused efficiencies in their core business, enabled by digital, in order to free up investment to support their dual transition. We are at the dawn of a new transformative era and we have only scratched the surface of how digital technologies can help expedite the delivery of substantial economic, environmental, and societal benefits."

In the past five years alone, implementing digital technologies has enabled organizations to reduce their energy consumption by almost a quarter (24%) and delivered a 21% reduction in GHG emissions. The report estimates that the reduction of global GHG emissions through the use of digital technologies by 2028 will outweigh the expected increase of emissions attributed to digital.

Almost 40% of the total workforce due to be dedicated to digital initiatives in the next 3-5 years

The global workforce will require significant transformation to keep pace with technological advancements at scale across industries. With 64% of organizations investing in reskilling their existing workforces, there is a need for flexible frameworks that allow for rapid evolution.

Telcos can save millions with advanced edge sustainability solutions

New report from STL Partners projects potential annual energy savings of up to 5% equating to ~$40 million in energy costs and over one million tons of carbon dioxide equivalent of emissions for a large group operator.

TO HELP TELCO providers navigate sustainability at the edge, STL Partners has released a new report, "Sustainability Insights: Navigating Sustainable Edge Strategies". The report finds that advanced edge sustainability strategies can reduce telco providers energy usage by 3-5% and one million tons of carbon dioxide equivalent (tCO2eq) of emissions for a typical large group operator. These strategies, including Precision Liquid Cooling, could save a typical large group telco operator up to $40 million a year in energy costs.

While traditional data center sustainability has made significant strides, edge sustainability remains an overlooked concern amid anticipated growth in edge computing. More so than in centralised cloud computing, sustainability at the edge poses its own unique challenges due to varied environments. Tailored, situation-specific strategies become essential for sustainable edge operations, considering each operator's unique infrastructure, environment, and resources.

In the report sponsored by Iceotope, STL Partners looks at five subcategories of specialised cooling technologies – Precision Liquid Cooling - Liquid to Air, Precision Liquid Cooling - Liquid to Liquid, Tank Immersion (Single and Two Phase), Direct to Chip Water Cooling, and Specialised Hardware - Ruggedised Servers. Precision Liquid Cooling - Liquid to Air technology - adopted by Iceotope for its KUL RAN solution – ranked highest amongst the cooling technologies, having been found to score well across six of the seven measurement criteria - legacy site reuse, energy use, water consumption, longevity of equipment, limited maintenance and COTS hardware adaptation.

"With edge computing projected to grow to over $450 billion in 2030, it is critical this growth is developed in as sustainable manner as possible – arguably more so than traditional centralised cloud," said Philip Laidler, director of consulting for STL Partners. "This report examines the multitude of factors that need be considered when defining your edge sustainability strategy. There is no one-size-fits-all answer and edge operators must select the optimal advanced sustainability strategy based on their unique context."

While traditional data center sustainability has made significant strides, edge sustainability remains an overlooked concern amid anticipated growth in edge computing

"As more data moves to the edge, new challenges for telco providers emerge. Greater sustainability, energy efficiency and serviceability across distributed workloads is needed more than ever before," said Nathan Blom, Chief Commercial Officer for Iceotope. "We designed KUL RAN specifically to reduce power consumption and maintenance costs for telco providers as they bring data-center class compute to the extreme Far Edge. This report from STL Partners validates that advanced sustainability strategies can be a game changer for the Telco Edge."

Precision Liquid Cooling efficiently captures nearly 100% of the heat generated by servers, cutting energy use by up to 40% and any water consumption by 96% across the entire telco data center estate from the cloud to the edge.

This innovative technology allows for greater flexibility of IT solutions by eliminating hotspots, minimizing wasted space, and reducing water consumption. Compatible with the same rack-based architecture as air cooled systems, it simplifies serviceability in both data centers and edge locations.

Legacy data centres - opportunity or liability?

Latest BCS Report lands with unique insight from over 3,000 industry players.

IN THE rapidly evolving landscape of information technology, legacy data centres are increasingly facing a plethora of challenges that hinder their ability to meet modern demands. This is according to the latest independent bi-annual industry survey by BCS, a leading provider of integrated IT asset consultancy solutions.

This latest report, the 27th edition, highlights the multifaceted issues confronting these traditional data centres, illuminating the complexities that organisations must navigate to remain competitive and efficient in today's digital era. According to the IDC, the average data centre is 9 years old, while Gartner states that any site more than 7 years old is obsolete. One-third of respondents have at least some proportion of their facilities which are between six and ten years old and around 17% operate stock which is ten years old or more.

Most respondents cited multiple challenges affecting them, with 56% of participants stating that the operational costs per sqm was too high to be competitive and would be problematic in the future. The lack of sustainable and renewable power closely followed in second place, illustrating the difficulty in meeting CSR and ESG targets when there is a lack of available renewable power to meet modern IT environment demands.

Additional challenges include: addressing the issue of embedded carbon in legacy data centres which is critical in the context of global efforts to reduce greenhouse gas emissions and combat climate change; inadequate disaster recovery and data backup systems; energy efficiency; maintenance costs and a shortage of the specialist skills needed to support the facility.

Jim Hart, CEO at BCS explains: "Legacy data centres were designed during a period when current technological advancements were not anticipated. These facilities now struggle to cope with the escalating requirements of modern computing, such as higher data volumes, faster processing speeds, and the need for robust cybersecurity measures. However, in some cases the perception of the issues faced by these facilities is more negative than the reality. For example, options such as retrofitting key M&E areas to address ESG concerns should be considered and certainly at BCS we have seen missed opportunities to address issues around scalability.

"That said there is no doubt that legacy data centres are at a critical juncture where they must overcome a myriad of challenges to stay relevant in the digital age. At BCS we have helped many clients navigate potential pathways for transformation and innovation to deliver the best possible outcomes to modernise their digital built assets. This includes helping organisations to figure out whether a given site warrants a closer look for investment by undertaking high level modelling, saving time and money in studies that may prove a site unsuitable for expansion; de-carbonisation; continued maintenance and operation," concludes Jim.

The BCS survey captures the views of over 3000 senior datacentre professionals across Europe, including owners, operators, developers, consultants and end users. It aims to gauge both the current and ongoing prospects for the data centre industry across Europe and covers insights into: supply and demand; expansion plans; drivers of change; power concerns and the ongoing skills shortage.

"Legacy data centres were designed during a period when current technological advancements were not anticipated. These facilities now struggle to cope with the escalating requirements of modern computing, such as higher data volumes, faster processing speeds, and the need for robust cybersecurity measures"

Hyperscalers and cloud data centres tackle Scope 3 emissions

As hyperscalers and cloud data centers grapple with the dual challenges of reducing Scope 3 emissions and managing escalating power consumption, the technology sector finds itself at a critical juncture.

WITH THE TECH SECTOR accounting for 2-4% of global GHG emissions and 7% of worldwide power consumption, hyperscalers and managed cloud service providers are responsible for a considerable proportion of the total. Rising energy costs and demand have become an increasingly important strategic and political issue.

Robert Pritchard, Principal Analyst, Enterprise Technology and Services at GlobalData, comments: "Cloud services play a role in everyone's life in the 21st century – and especially amongst businesses and government bodies. A very significant part of providing these services is energy – and the acceleration of adoption of artificial intelligence is only going to drive further rapid growth in demand."

GlobalData analysis reveals that Scope 3 emissions, stemming from indirect sources, constitute a substantial 70-90% of greenhouse gas emissions. In response, hyperscalers and managed cloud providers are endeavoring to mitigate their consumption. Notably, they have achieved considerable success in managing Scopes 1 and 2, where they exert more direct control.

Pritchard adds: "Most large tech companies are aiming to be carbon net zero by 2030 or 2040, and progress has been made by using more efficient equipment and by moving to renewable energy sources.

Achieving Scope 3 will see tech companies demand similar efforts from their supply chain to meet their targets,

which is a very tall order."
Pritchard continues: "Many tech companies are looking to differentiate on their 'green' credentials and sustainability is a strategic issue. Validation is key, as are targets. A growing number of customers, partners, and governments require specific achievements as stakeholder entry stakes."

Enterprises to access digital infrastructure using subscription-based services

By 2026, 80% of new enterprise digital infrastructure investment is forecast to be operated through a subscription-based model, according to the Global Interconnection Index (GXI) 2024 released recently.

GXI 2024 predicts that in order to meet the ever-growing demands of data-dense technologies such as AI, 5G and edge computing, IT decision-makers are increasingly shifting away from long-term purchases of physical equipment, such as servers, routers and storage arrays, in favour of flexible subscription-based models.

This shift from CAPEX to OPEX started with multicloud adoption but is now becoming the norm across all infrastructure out to the edge, providing enterprises with greater agility in architecting their infrastructure everywhere while ensuring they have access to the most efficient technologies.

"Industry patterns have shown that the traditional procurement process of buying your own IT hardware, if that is not your business, is becoming a competitive disadvantage," said Steve Madden, Vice President of Digital Transformation & Segmentation, Equinix. "The pace of hardware innovation is increasing (especially with GPU technologies), putting pressure on price-performance ratio and infrastructure efficiency. Globally, digital transformation requires businesses to become more agile while adapting to dynamic changes. Subscription models can offer continuous improvement and easier adoption of new technologies already in place."

Additional Forecast and Trends from GXI 2024

- **Digital economy continues to expand:** Global interconnection bandwidth is forecast to grow at a 34% five-year compound annual growth rate (CAGR), reaching 33,578 terabits per second (Tbps) by 2026.
- **Accelerating growth of ecosystems:** Organizations are connecting with 30% more business partners in twice as many locations.
- **Digital proximity drives business at the edge:** Edge infrastructure has shown the highest growth rate, and is expected to expand at over two times the rate of core through 2026.

Legacy infrastructure creating sustainability nightmare for 63% of IT leaders

Despite sustainability and energy efficiency ranking as important, Daisy study finds many IT leaders lack real confidence in meeting their green targets.

ALMOST two-thirds (63%) of IT leaders say their organisation's legacy infrastructure is causing them a huge sustainability nightmare according to new research from Daisy Corporate Services (Daisy). At a time when sustainability is under the microscope, the survey of 250 senior IT decision makers reveals that legacy hardware currently contributes to more than a third (37%) of organisations' overall power consumption.

Almost nine in ten IT leaders (86%) say that sustainability and energy efficiency is important to their operations, with 84% stating that their organisation has IT efficiency targets in place. However, only half (51%) of those surveyed are "very confident" they can meet these targets.

"Sustainability is a vital component of any modern business, and IT departments have a growing role in helping the wider organisation achieve green targets. But legacy technology is a cause for concern amongst IT teams, with ageing equipment still contributing to a significantly to power consumption," comments Andy Bevan, Head of Propositions and Strategy Consulting at Daisy. "Organisations can benefit from the sustainability features of their cloud providers but are being held back by the challenges of migrating their legacy hardware. Here is where modern hybrid cloud platforms can help bridge the gap between on-site infrastructure and cloud to deliver performance and sustainability benefits."

IT budgets remain under scrutiny

In addition to creating sustainability challenges, legacy technology also remains a significant cost centre. IT leaders admit that almost a third (29%) of their budgets are still being used to support, maintain, and manage inefficient legacy hardware. At the same time, many IT leaders are being asked to reevaluate their IT spending. More than two-thirds (69%) of survey respondents describe the pressure to reduce IT capital expenditure as "significant."

As a result, IT leaders are weighing up approaches optimise costs. Of those surveyed, 86% believe moving to a consumption-based IT infrastructure model will benefit their organisations, with increased flexibility driving lower costs. In addition, 82% of IT decision makers think the use of AIOps (Artificial Intelligence for Operations) will enhance their operations in the future.

"Over the last few years, it has been a tough operating environment for many organisations. Driving efficiencies is a big part of businesses' survival strategies today, which is having a huge impact on IT teams," adds Bevan. "At a time when IT leaders are under pressure to reduce capital expenditure many organisations are still incurring significant maintenance and support costs on their legacy hardware. By moving to the cloud and a consumption-based pricing model, organisations can reduce ongoing costs and increase flexibility by paying for what they use. For cost-constrained IT departments this should be their nirvana."

New roadmap to position Europe as the 'Quantum Valley' of the world

Quantum Flagship unveils new roadmap with calls to end reliance on outside nations for developing essential components and hardware and position Europe as the world's first 'Quantum Valley'.

QUANTUM experts met policymakers and representatives at the European Commission in Brussels recently to present the new Strategic Research and Industry Agenda SRIA 2030: Roadmap and Quantum Ambitions over this Decade.

The new strategy published by the Quantum Flagship shows Europe is building an autonomous ecosystem for scientific and industrial applications in the quantum space, with thousands of researchers, a thriving workforce and the highest level of public funding for quantum technologies in the world. Strong themes in the roadmap include economic and technological sovereignty, with calls to strengthen the EU's role as a global player in this transformative field.

The roadmap seeks to position the European Union with its innovation hubs, research centres, startups, ready workforce, and multi-billion Euro investments as the quantum equivalent to Silicon Valley – or the world's first 'quantum valley'. This call follows the recent European Declaration on Quantum Technologies in December 2023, with the ultimate aim of making Europe the leading global region for quantum excellence and innovation.

Salvatore Cinà, coordinator of the SRIA, said, "This new strategic research agenda recognises quantum technologies as an emerging global strategic domain with immense potential for tackling societal challenges. By consolidating our research efforts and forging strong links between academia and industry, we are poised to lead in quantum technologies globally. With contributions from more than 300 experts, the SRIA defines the strategy and the roadmap for Europe in Quantum Technologies."

Coordinated by the French Atomic Energy and Alternative Energies Commission (CEA) within the European Coordination and Support Action 'QUCATS', the new roadmap, supported by specific recommendations for future programs, seeks to merge numerous quantum agendas from research, industrialisation, computing, and communication, which co-exist today into a single coherent strategy. This streamlined roadmap will allow the European Commission to optimise all quantum investments in the future. Quantum Computers, Simulators, Communications and the Quantum Internet

If the recommendations in the roadmap are met, Europe could be at the forefront of this competitive global race to improve the performance and maturity of quantum technologies by 2030.

The SRIA provides a clear strategy and recommendations to develop quantum computing and simulator devices that outperform or accelerate existing classical computers to solve specific problems relevant to industry, science, and technologies.

It also discusses the strategy to establish quantum communication. The agenda provides short-term plans and recommendations to enhance the network with advanced quantum cryptography protocols, with security based on the laws of quantum physics. Such a network would also share and distribute quantum resources like superposition and entanglement.

The long-term goal is to realise the quantum internet – a network connecting different quantum devices, from computers and simulators to sensors – at a European scale.

Expanding the Quantum Community
The Strategic Agenda seeks to bolster Europe's quantum workforce by broadening and strengthening the quantum community. The roadmap recommends actively championing equality, diversity, and inclusion to ensure a wide range of skilled professionals can enter Europe's quantum workforce.

Beyond the quantum community, the SRIA suggests embracing other industrial and academic sectors looking to integrate quantum technologies into their products and services.

In particular, the high-performance computing (HPC) industry is poised to harness quantum hardware as a powerful accelerator within its infrastructures. This collaboration will improve computational power and capabilities to address fundamental problems for society, like drug discovery, new materials for energy generation and storage, and optimise logistics and transport. The communications industry is also gearing up to enhance the security and capabilities of future networks by leveraging quantum technologies. Meanwhile, the semiconductor industry is integrating quantum technologies into its roadmap to address the needs of pilot and production lines dedicated to quantum chips, as well as the enabling technologies, such as classical chips for quantum.

DCS AWARDS 2024

CELEBRATING 14 YEARS OF SUCCESS

The DCS Awards: 36 Categories across 5 Themes

23 MAY
Leonardo Royal Hotel
London St Pauls

KEY DATES:

03 MAY: VOTING CLOSES

23 MAY: AWARDS CEREMONY

SPONSORSHIP PACKAGES

The DCS Awards offer extensive branding and sponsorship opportunities through online advertising in our Datacentre Solutions & Digitalisation World publications, and of course at the awards ceremony itself.

HEADLINE SPONSOR

HUAWEI

CATEGORY SPONSORS

H2H COMMUNICATIONS

oryx align

PowerControl A brand of legrand

Schneider Electric

For sponsorship opportunities and/or to book your awards table please contact: awards@dcsawards.com or call +44 (0)2476 718970

VOTE HERE: https://dcsawards.com/vote

Supported by

dca The data centre trade association

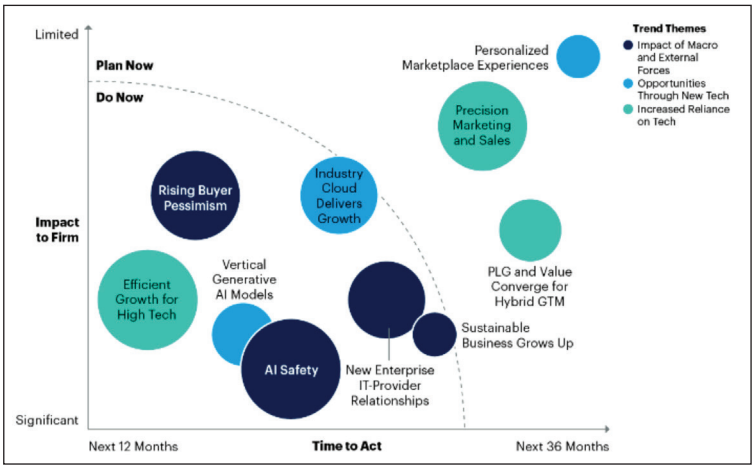
Gartner identifies top trends impacting technology providers in 2024

Gartner, Inc. has highlighted the top trends that will impact technology providers in 2024. Worldwide IT spending and GenAI and cloud are also covered.

"GENERATIVE AI (GenAI) is dominating the technical and product agenda of nearly every tech provider," said Eric Hunter, Managing Vice President at Gartner. "The technology reshapes a tech provider from its growth and product strategy down to the everyday tools used by its associates. Despite the potential for GenAI to reshape providers, it is not the only influence facing technology leaders. There are new points of friction in growth plans, new points of fusion in marketing and sales, and new relationships opening up to technology and service providers (TSPs)."

The immediate and long-term implications of these issues require product leaders to balance between short-term opportunity and long-term advantage and strategies based on economic recovery or recession. Gartner's top trends for 2024 reflect these dualities (see Figure 1).

Efficient Growth for High Tech
Significant growth in IT spending over the last decade benefited high-tech companies. Capturing that growth led high-tech firms to pursue growth without a full measure of the costs. This is a "growth at all costs" strategy. High-tech firms anchored their product, organization and employment plans on a hypothesis of continued strong growth.



➤ Figure 1: 2024 Gartner Tech Provider Top Trends.
Note: The bubble size for each trend conveys the relative magnitude of disruption for a given trend, relative to the other top trends. Source: Gartner (February 2024)

As macroeconomic conditions create uncertainty among buyers and increasing costs of capital shift investor focus to margin growth, Gartner analysts see a trend toward tech providers focusing on efficient growth. Efficient growth strategies recognize the value in growing in ways that strengthen current margins and future revenue opportunities.

New enterprise it-provider relationships

Increased business and technical demands require enterprise IT to cover more ground at a deeper level and a faster pace, eroding enterprise IT's capacity and capabilities. This creates a trend for product leaders at tech providers to create new relationships and revenue opportunities across the enterprise, including expanded provider roles within enterprise IT and the business, outcome-centric provider-enterprise relationships and enterprise-wide tier-1 relationships.

Sustainable business grows up

Sustainability efforts and managing the ESG impact have been unilaterally focused on mitigating internal risk and ensuring compliance. Product leaders must evolve by embracing double materiality and holistic leverage of emerging technologies to meet sustainability objectives.

AI safety

Responsible AI and AI safety are not new concepts, but the unprecedented rapid development of GenAI technologies has fueled the discussion around risk management and how to address growing issues such as content provenance and hallucination. Product leaders must build solutions that incorporate safety principles with a focus on model transparency, traceability, interpretability and explainability aspects. Preempting regulatory and compliance issues will be

critical to staying competitive in this vibrant GenAI market by creating trust.

Rising buyer pessimism

Over the past three years, tech providers have increasingly observed negative sales pipeline effects due to new buyer behaviors that are colliding with outdated go-to-market (GTM) models. Without adapting sales and marketing approaches to detect and respond to buyer pessimism, technology providers will see their own GTM operations decline in both internal and external perspectives.

Vertical generative AI models

While general-purpose models perform well across a broad set of GenAI applications, they can be impractical for many enterprise use cases that require domain-specific data. Tech providers must explore industry-focused models that can be adapted to specific user requirements using available resources more efficiently. Those failing to do so will face increased costs and complexity in the creation and leverage of models.

Personalized marketplace experiences

Specialized, niche, digital marketplaces are emerging to help buyers navigate the complexity of procuring, implementing and integrating solutions. Product leaders who do not offer their services through personalized digital marketplaces limit their findability for their target customers. Gartner predicts that 80% of sales interactions between suppliers and buyers will occur in digital channels by 2025.

Industry cloud delivers growth

Service providers, hyperscalers, ISVs and SaaS providers are turning to vertical solutions to deliver the customer outcomes that will drive provider growth. By 2027, Gartner predicts that more than 50% of tech providers will use industry cloud platforms to deliver business outcomes, up from less than 5% in 2023.

PLG and value converge for hybrid GTM

Product-led-growth (PLG) focuses on showing value to product users, creating intent signals that go-to-market (GTM) teams can use with prospective buyers. But most companies using a PLG GTM have begun to realize that, in most cases, a 100% self-serve GTM motion isn't tenable. At some point, sellers must be involved to convert deals.

Buyer needs for business value and outcome justification — for new or expansion business — will meld PLG tactics with value management and realization initiatives in hybrid GTM strategies.

Precision marketing and sales

Rapidly evolving technology advances, such as GenAI, digital buying and the metaverse, are changing how tech providers market and sell technology. Tech providers failing to adopt new approaches will see the erosion of overall deal quality combined with the loss of relevance and limited growth within established accounts.

Worldwide IT spending to grow 6.8%

WORLDWIDE IT spending is expected to total $5 trillion in 2024, an increase of 6.8% from 2023, according to the latest forecast by Gartner, Inc. This is down from the previous quarter's forecast of 8% growth.

While generative AI (GenAI) had significant hype in 2023, it will not significantly change the growth of IT spending in the near-term. "While GenAI will change everything, it won't impact IT spending significantly, similar to IoT, blockchain and other big trends we have experienced," said John-David Lovelock, Distinguished VP Analyst at Gartner. "2024 will be the year when organizations actually invest in planning for how to use GenAI, however IT spending will be driven by more traditional forces, such as

Table 1. Worldwide IT Spending Forecast (Millions of U.S. Dollars)

	2023 Spending	2023 Growth (%)	2024 Spending	2024 Growth (%)
Data Center Systems	243,063	7.1	261,332	7.5
Devices	699,791	-8.7	732,287	4.6
Software	913,334	12.4	1,029,421	12.7
IT Services	1,381,832	5.8	1,501,365	8.7
Communications Services	1,440,827	1.5	1,473,314	2.3
Overall IT	4,678,847	3.3	4,997,718	6.8

Source: Gartner (January 2024)

profitability, labor, and dragged down by a continued wave of change fatigue."

IT Services Becomes largest segment of IT spending in 2024

IT services will continue to see an increase in growth in 2024, becoming the largest segment of IT spending for the first time. Spending on IT services is expected to grow 8.7% in 2024, reaching $1.5 trillion (see Table 1). This is largely due to enterprises investing in organizational efficiency and optimization projects. These investments will be crucial during this period of economic uncertainty. "Adoption rates among consumers for devices and communications services plateaued over a decade ago.

Consumer spending levels are primarily driven by price changes and replacement cycles, leaving room for only incremental growths, so being surpassed by software and services was inevitable," said Lovelock. "Enterprises continue to find more uses for technology – IT has moved out of the back office, through the front office and is now revenue producing, until there is a plateau for how and where technology can be used in an enterprise, there cannot be a plateau in enterprise IT spending."

CIOs' change fatigue continues to impact IT spending

The overall IT spending growth rate for 2023 was 3.3%, only a 0.3% increase from 2022. This was largely due to change fatigue among CIOs. Momentum will regain in 2024, with overall IT spending increasing 6.8%.

Even with the expected regained momentum in 2024, the broader IT spending environment remains slightly constrained by change fatigue. Change fatigue could manifest as change resistance — with CIOs hesitating to sign new contracts, commit to long-term initiatives or take on new technology partners. For the new initiatives that do get launched, CIOs require higher levels of risk mitigation and greater certainty of outcomes.

Sustainability and sovereignty top GenAI cloud criteria

BY 2027, 70% of enterprises adopting generative AI (GenAI) will cite sustainability and digital sovereignty as top criteria for selecting between different public cloud GenAI services, according to Gartner, Inc.

"Because of its scale and shared services model, cloud technology is best-suited for the delivery of GenAI-enabled applications at scale and the development of general-purpose foundation models," said Sid Nag, Vice President Analyst at Gartner. "However, certain aspects must be addressed, including digital sovereignty, or the ability to control where data is stored and where operations are executed, and sustainability issues so that organizations can operationalize GenAI."

Digital sovereignty's role in public cloud decisions

The advancement of foundation models (FMs) and large language models (LLMs), which are the core of GenAI capabilities, are driving rapid and continual evolution of GenAI capabilities and use cases. Implementation of GenAI in the enterprise poses significant regulatory challenges, including regulations on the data contained within LLMs, as well applications that leverage these FMs and LLMs.

"Specialty cloud providers will become an important consideration for many enterprise cloud architectures as organizations extend their cloud operations to cover diverse locations and use cases," said Nag. "Digital sovereignty will drive the need to include cloud providers that can meet the evolving and unique requirements of sovereign operations no matter the region they operate in."

Sustainability's role in public cloud decisions

Organizations deploying GenAI services will look to the public cloud, given the scale of the required infrastructure, but they will also require cloud providers to address nontechnical issues related to sustainability.

The pressure from investors, customers, regulators and governments about sustainability is forcing organizations to manage and optimize their IT carbon emissions to achieve their environmental sustainability goals. New processes, capabilities and tools will be introduced, oriented to the monitoring and management of energy consumption and carbon emissions for GenAI workloads deployed on cloud.

"Cloud computing plays a pivotal role today in supporting sustainability and GenAI business applications by providing scalable infrastructure, enabling eco-friendly practices and allowing cost-effective resource management," said Nag. "Therefore, cloud is the platform that most IT leaders rely upon to support their sustainability journey when it comes to overall GenAI implementation."

DCS ROUNDTABLE

Developing Digital Infrastructure In A Hybrid World

Not every discussion is a **heated debate**

- Based around a hot topic for your company, this 60-minute recorded, moderated ZOOM roundtable would be a platform for debate and discussion
- Moderated by Editor, Phil Alsop, this can include 3 speakers
- Questions prepared and shared in advance
- There would be an opportunity to view and edit in advance

This event would be publicised for 4 weeks through all our mediums including:

- A banner on the Digitalisation World Magazine homepage for 8 weeks
- 4x weekly dedicated HTMLs
- 4x news pieces which would also appear on the weekly e-newsletters
- Promoted through our social media platforms for 8 weeks (pre and post event)
- Available as an on-demand asset through all mediums
- All registered attendees' details would be made available to you

Cost: £5995

Contact: Jackie Cannon at jackie.cannon@angelbc.com

ANGEL EVENTS

Five data centre predictions for 2024

The critical digital infrastructure sector continues to enjoy robust growth. Rapidly evolving technologies will further drive and sustain this trend in 2024 and beyond — but will also create new challenges for operators.

BY JACQUELINE DAVIS, RESEARCH ANALYST; JOHN O'BRIEN, SENIOR RESEARCH ANALYST; DOUGLAS DONNELLAN, RESEARCH ANALYST; ANDY LAWRENCE, EXECUTIVE DIRECTOR OF RESEARCH; DANIEL BIZO, RESEARCH DIRECTOR; MAX SMOLAKS, RESEARCH ANALYST, UPTIME INTELLIGENCE

A RECENT Uptime Intelligence report highlights some of these challenges and their implications. These include greater scrutiny over sustainability commitments; the adoption of power-hungry AI; the need for (and limitations of) direct liquid cooling; the evolution of data center management software; and the emergence of data center campuses that redefine the meaning of hyperscale. Readers are encouraged to read the full report, but for the purposes of this article, we are focusing on just one of the five predictions - Demand for AI will have a limited impact on most operators.

At the beginning of each calendar year, Uptime Intelligence compiles a short list of trends or predictions that will be relevant to the digital infrastructure sector for the year (and years) ahead. These lists aim to highlight crucial, yet often overlooked industry topics, which encourage a closer examination.

The predictions for 2024 highlight the opportunities arising from the sustained surge in demand for IT and the progress in new IT and facility technologies. However, we also address the accompanying challenges: how to cool high-density racks at scale, how to meet escalating IT demand, and the difficulties around expanding capacity while also meeting ever-stricter sustainability commitments.

Despite the notable growth of the digital infrastructure sector over the past five years, external global events (including the COVID-19 pandemic, Russia's invasion of Ukraine and extreme weather events) have disrupted supply chains and energy prices, and raised the cost of capital projects. Many of these challenges, however, result from the ongoing success of the IT sector itself, with the development of new software (including artificial intelligence, AI) and processors, and the economical delivery of IT services.

Investment by owners and operators of data centers, including cloud and hyperscale operators, is set to increase. Uptime Institute survey data reveals that more than half (61%) of enterprise operators and almost three-quarters of colocation operators (71%) expect their data center spending budgets to increase in 2024, and this is primarily driven by the need for capacity growth.

Laws passed in 2023 — and others slated for the years ahead — are creating additional hurdles for operators. These regulations primarily focus on reporting climate risks, improving energy performance and lowering carbon emissions. While not all organizations will be affected by these regulation, data center operating costs are still likely to increase.

In addition to the increasing legislation around climate change, 2023 also exposed the industry's lack of preparation in responding to extreme weather events as heat waves overwhelmed data center cooling systems in some regions. Recent record-breaking temperatures are expected to climb even higher in 2024 and will force many operators to reassess their resiliency strategies. Many are already increasing investment in this area.
To help make operational decisions relating to resiliency and energy efficiency, many organizations may plan to leverage AI technologies in 2024. However, according to Uptime survey data, trust in AI to make operational decisions has decreased over the past year. This is likely due to some of the unpredictable and inaccurate results of large language models. Even so, innovation in other forms of AI and machine learning is beginning to make an impact in the data center sector.

Securing and accommodating the necessary infrastructure for AI training models will be expensive and will require power-hungry IT and facility equipment. Given the costs and supply chain constraints, these deployments may be limited to only a few large-scale operators in the near term. The year ahead will task operators with balancing new technology integration against costly infrastructure updates and sustainability pressures — while also managing greater complexity and minimizing operational risk.

Demand for AI will have a limited impact on most operators

Key trends

- Advances in generative AI models have created runaway expectations of demand for data center capacity and high-density racks.
- A limited supply of chips for training large models will cap the significant pace of capacity growth for AI.
- AI's impact will be widely felt throughout the industry, but indirectly through straining equipment chains, pushing server chip power levels and making operators rethink their facility resiliency posture.

The recent intrigue and fascination with artificial intelligence (AI) marks a new high in the evolution of the technology. Now the data center industry is bracing itself for a significant increase in demand for net new capacity and the technically challenging requirements of providing sufficient power and cooling.

The intense activity in developing AI-powered services results largely from two developments: one in neural network architecture, called the "transformer"; the other in compute hardware, spearheaded by Nvidia's continued development of AI acceleration in silicon.

The launch of OpenAI's ChatGPT chatbot in November 2022, demonstrating previously unseen machine-learning sophistication, triggered an "arms race" as organizations competed to develop sophisticated AI-powered applications.

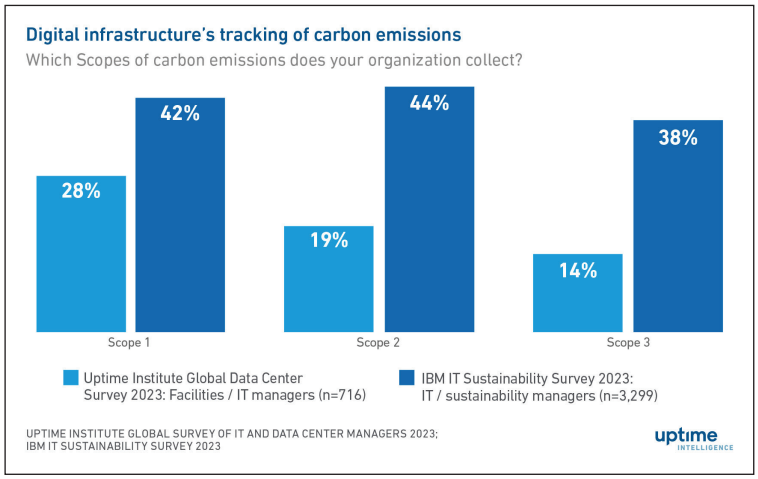
Demand for AI-accelerators is outstripping supply, and industry stakeholders are predicting a significant increase in new data center capacity to accommodate the higher rack densities required to drive generative AI workloads. This could also place strenuous demands on power and cooling infrastructures, as well as drive up the size and weight of extreme density cabinets.

While there is little doubt that demand for compute power and density will increase, Uptime Intelligence considers some of these expectations to be overblown. For many operators, the impact on data centers will be more indirect through AI's effect on supply chains, chip designs and attitudes towards facility resiliency.

New capacity demand and densification will be muted

There are two key reasons for new capacity demand and muted densification. First, the speed and scale at which AI training infrastructure grows is capped by chipmaking capacity. Even though Nvidia has been able to scale up production of its chips at its manufacturing partner TSMC, it is managing output carefully to avoid a potential supply glut following pent-up demand. Currently, Nvidia is forecast to ship up to 600,000 of the H100 / H200 chips by the end of 2023, and another 1.5 million to 2 million in 2024. These chips will form the basis of the large majority of AI training infrastructure for large generative models.

Assuming that most of these chips will be running AI workloads, Uptime Intelligence considers 2,000 megawatts (MW) to 2,500 MW of additional IT load capacity (uplift compared with baseline data center demand) to be a plausible scenario between the start of 2023 and the first quarter of 2025. Even if the whole sector adds hundreds of megawatts of data center capacity in the next 18 months (both net new and through planned expansions)



to accommodate for a surge in AI infrastructure demand, this would be a significant — but not dramatic — undertaking. Moreover, relatively few operators and sites will see the bulk of this uptake. Currently, only a handful of customers account for the majority of AI chip sales, with major cloud infrastructure and web services companies (such as Microsoft, Google and Amazon Web Services, as well as their counterparts in China) topping the rankings.

Second, while AI-optimized hardware can be much denser than typical IT equipment, it does not need to take extreme form. Nvidia's reference designs call for up to 50 kilowatts (kW) of power in its densest implementation, but hardware can also be spread out to meet power delivery or airflow limitations. Although much denser systems exist (some well above 100 kW per cabinet), they tend to be offered to a few large, multi-megawatt supercomputer installations to limit the footprint and to mitigate the cable run length limitations of high-speed interconnects.

In reality, most recently built facilities with modern power distribution (3-phase, higher voltages) will be able to handle all but the largest AI training clusters. This will occur through a combination of spreading out systems, upgrading breakers where necessary and adding more power circuits to reach the desired level of capacity per rack (e.g., 20 kW or 40 kW).

Generative AI's broader impact: slow but lasting
Besides adding demand for capacity and the adoption of high-density IT systems, Uptime Intelligence expects generative AI (and other forms of AI) to affect data center operators indirectly. In chronological order of the expected time of impact:

- Prolonged demand-supply imbalance. Lead times of data center equipment remain long. This is particularly evident when it comes to large systems, such as engine generators (which often have two-year waiting lists), switchgears, transformers, uninterruptible power supplies — but also some mechanical equipment and smaller components. The additional capacity needs of AI training clusters will contribute to a wider demand-supply imbalance by tying up even more of the equipment supply in the hands of a relatively few large and hyperscale data center operators.
- AI arms race will push chip power envelopes further. Server silicon power ratings have escalated markedly since around 2017. The scale of integration has outpaced transistor energy gains and many larger IT customers prefer more performant systems, even at the cost of higher power consumption. There is a renewed bout of fierce competition between chipmakers for performance supremacy in AI and other technical computing tasks. Cloud infrastructure will also inevitably push up silicon design power further and faster. In a few years, mainstream servers with up to 1 kW of realized power use will be common.

This trend affects the design of all current and future server processors and accelerators, and shapes model line-up and pricing decisions. New generations of servers may only bring tangible benefits in application performance or economics for those with specific and suitable IT workloads.

Equally, keeping up with best-in-class server chips is costly: the same class of products (relative in the porfolio) can cost 50% to 100% more today than they did five or six years ago, despite growing competition. Chipmakers will offer only the largest buyers sufficient discount to offset this silicon inflation.

- Promoting mixed-tier facilities. Most data centers are designed and built to a single level of resiliency, typically aiming for very high levels of service availability despite its high cost (low resiliency facilities tend to be dedicated to supercomputing). The real possibility of AI-heavy services becoming a meaningful application category in mission-critical enterprise and colocation facilities has led to a reconsideration of mixed-tier facilities.

Arguably, the criticality of IT racks running AI workloads (training in particular) does not warrant the costly overheads from delivering conditioned power backed up by batteries and engine generators, let alone all the component redundancy required to achieve concurrent maintainability.

The same applies to many other applications serving low criticality functions, yet receiving gold-plated facility services. A wider spread of enterprise AI training systems might change the resiliency posture of some data center operators from uniform resiliency standards to a multi-tier service approach.

To download the executive summary of this report, and access related resources please visit: https://uptimeinstitute.com/resources/research-and-reports/five-data-center-predictions-for-2024

DCS AWARDS 2024

CELEBRATING 14 YEARS OF SUCCESS

The DCS Awards: 36 Categories across 5 Themes

23 MAY
Leonardo Royal Hotel
London St Pauls

KEY DATES:

03 MAY: VOTING CLOSES

23 MAY: AWARDS CEREMONY

SPONSORSHIP PACKAGES

The DCS Awards offer extensive branding and sponsorship opportunities through online advertising in our Datacentre Solutions & Digitalisation World publications, and of course at the awards ceremony itself.

HEADLINE SPONSOR

HUAWEI

CATEGORY SPONSORS

H2H COMMUNICATIONS

oryx align

PowerControl A brand of legrand

Schneider Electric

For sponsorship opportunities and/or to book your awards table please contact: awards@dcsawards.com or call +44 (0)2476 718970

VOTE HERE: https://dcsawards.com/vote

Supported by

dca The data centre trade association

Navigating power demand in the age of AI

As the world increasingly relies on digital services and AI, data centres face a growing problem in terms of power demand. ABB explores what data centre professionals can do to meet surging demand, while staying competitive and meeting their sustainability goals.

DANEL TURK, DATA CENTRES PORTFOLIO MANAGER AT ABB

THE CONTINUOUS GROWTH of AI services since the launch of ChatGPT in 2022 means that data centre power demands will continue to rise. In fact, the International Energy Agency's (IEA) Electricity 2024 report predicted that global data centre electricity use could double by 2026, reaching 1000 terawatt-hours (TWh) – that's equivalent to the annual electricity consumption of Japan.

So, data centre professionals face several challenges when looking to meet that demand in a cost-effective, sustainable way. Here are a few considerations.

Think modular

If you're building or enlarging a data centre, take advantage of a modular approach. Modularity offers a way of scaling sustainably in a way that meets the demand for power and availability while simplifying the specification and installation process.

Rather than constructing or expanding a data centre all at once, you can build it up in blocks. As an example, you can build a projected 200-megawatt (MW) data centre in 20 MW sections – onboarding customers as demand increases. That makes it more cost-effective because it avoids wasted capacity along the way.

It's also an approach that minimises downtime and risk, especially with prefabricated solutions – like skids and eHouses – that are ready for quick and easy installation as they are constructed off-site and then factory-tested before delivery. Besides, buying one prefabricated product instead of a range of individual components that require assembly saves money and time. This approach is favoured by local utilities and governments too, with the smaller incremental growth of a data centre more likely to be approved than the full-scale build of a new one.

Consider shifting to medium voltage

Medium-voltage (MV) equipment has become increasingly viable and cost-effective especially in the case of rising power demand.

MV UPSs, for example, can be installed modularly. As an example, ten 2.5 MW UPS blocks can be wired in parallel to create a 25 MW system. This enables faster deployment and increases overall system capacity without any additional complexity. It also avoids the challenge of having multiple LV UPSs that need to be regularly maintained and serviced. In addition, the lower currents at MV level mean cables can have a smaller cross-section, leading to additional savings.

Furthermore, MV UPSs are often more energy efficient than their low-voltage (LV) counterparts, and can provide power protection for the entire data centre, not just the server racks. So, switching to MV can offer long-term cost savings.

Be more sustainable with BESS

Meeting sustainability goals is a challenge for data centre operators when it comes to the increase in power demand.
One solution is using battery energy storage systems (BESS) to reduce reliance on diesel gensets as they can integrate renewables, like wind and solar energy, into the local energy mix. By using a BESS, excess energy from solar panels on the roof can be stored and used at another time, optimising renewables usage.

Using BESS can also provide load shifting and frequency response services to the grid, further aiding negotiations with the local utility, and creating potential new revenue streams via give-back schemes during peak demand.

First, though, it's important to plan carefully and consider what your energy demands are to avoid over-specifying – after all, specifying for five hours of uptime would require a lot of batteries.

Make use of AI
The onward development of AI might be driving the increase in power demand, but it can also help to make data centres more reliable and efficient. This could happen through optimised cooling, predictive condition-based maintenance, data access and transfer, and demand balancing.

The key to this is connectivity. An automation system running an AI suite can keep track of the hundreds of thousands of monitoring points deployed in a typical mid-to-large data centre. This provides operators a 360 overview of the data centre's performance, energy use and asset health at any given point.

This data can then be used to make efficiency improvements. Take the example of a cooling system. While the upstream chiller and the

distribution system are often viewed separately – leading to operational inefficiencies when attempts are made to make it more efficient – an automation system lets operators see holistically how one part affects the other. As a result, operators can make more informed decisions to improve overall efficiency.

Invest in SF6-free equipment

Moving to SF6-free equipment now will ensure that you're prepared for incoming regulations on the use of electrical equipment which contains the insulating gas. This will make your sustainability reporting easier. Regulations have been proposed because SF6 can leak to the atmosphere, where it has a global warming potential around 25,000 times greater than CO2.

Adopt a TCO mindset

When considering how to grow efficiently and sustainably, data centre managers need to adopt a total cost of ownership (TCO) mindset.
This involves calculating and assessing all the direct and indirect costs of an asset over its entire lifecycle to determine its cost, as opposed to a more traditional view of separating capital and operational expenditure. By viewing expenditure with a TCO mindset, future operating cost savings are seen as net present value.

A simple way of looking at this is where you have a machine that will be running continuously for around 20 years; by investing in a more energy efficient model now, you will see significant cost and emission savings over the long term.

Power demands might be growing as AI and digital services expand exponentially, but data centre managers and operators can meet these challenges with the right strategies in place. While these challenges will become more urgent as demand increases, acting now, with a TCO mindset, will help to ensure data centres continue to make a positive impact.

AI and data centres -

Why AI is so resource hungry

As we wait for the numbers to emerge for past and existing power use for ML and AI what is clear is that it is once models get into production and use, we will be in the exabyte and exaflop scale of computation. For data centre power and cooling, it is then that things become really interesting and more challenging.

BY ED ANSETT, FOUNDER AND CHAIRMAN, I3 SOLUTIONS GROUP

AT THE END OF 2023 any forecast of how much energy will be required by generative AI is inexact.

Headlines tend towards guesstimates of "5x, 10x, 30x power needed for AI" and "Enough power to run 100,000s of homes" etc. Meanwhile reports in specialist publications such as the data centre press talk of power densities rising to 50kW or 100kW per rack.

Why is generative AI so resource hungry? What moves are being made to calculate its potential energy cost and carbon footprint? Or as one research paper puts it, what is the "huge computational cost of training these behemoths." Today, much of this information is not readily available.

Analysts have forecast their own estimates for specific workload scenarios (see below), but with few disclosed numbers from the cloud hyperscalers at the forefront of model building, there is very little hard data to go on at this time.

Where analysis has been conducted, the carbon cost of AI model building from training to inference has produced some sobering figures. According to a report in the Harvard Business Review " researchers have argued that training a 'single large language deep learning model' such as OpenAI's GPT-4 or Google's PaLM is estimated to use around 300 tons of CO_2. Other researchers calculated that training a medium-sized generative AI model using a technique called "neural architecture search" used electricity and energy consumption equivalent to 626,000 tons of CO_2 emissions."

So, what's going on to make AI so power hungry?

Is it the data set, i.e. volume of data? The number of parameters used? The transformer model? The

encoding, decoding and fine tuning? The processing time? The answer is of course a combination of all of the above.

Data

It is often said that GenAI Large Language Models (LLMs) and Natural Language Processing (NLP) require large amounts of training data. However, measured in terms of traditional data storage, this is not actually the case.

For example, ChatGPT used www.commoncrawl.com data. Commoncrawl says of itself that it is the Primary training corpus in every LLM and that it supplied 82% of raw tokens used to train GPT-3: “We make wholesale extraction, transformation and analysis of open web data accessible to researchers. Over 250 billion pages spanning 16 years. 3 - 5 billion new pages added each month.”

It is thought that ChatGPT-3 was trained on 45Terabytes of Commoncrawl plaintext, filtered down to 570GB of text data1. It is hosted on AWS for free as its contribution to Open Source AI data.

But storage volumes, the billions of web pages or data tokens that are scraped from the Web, Wikipedia and elsewhere then encoded, decoded and fine-tuned to train ChatGPT and other models, should have no major impact on a data centre. Similarly, the terabytes or petabytes of data needed to train a text to speech, text to image or text to video model should put no extraordinary strain on the power and cooling systems in a data centre built for hosting IT equipment storing and processing hundreds or thousands of petabytes of data.

An example of a text to image model is LAION (Large Scale AI Open Network) - a German AI model with billions of images. One of its models, known as LAION 400m, is a 10TB web data set. Another, LAION5B has 5.85 billion clip filtered text image pairs.

One reason that training data volumes remain a manageable size is that it’s been the fashion amongst the majority of AI model builders to use Pre-Training Models (PTMs), instead of search models trained from scratch. Two examples of PTMs that are becoming familiar are Bidirectional Encoder Representations from Transformers (BERT) and the Generative Pre-trained Transformer (GPT) series - as in ChatGPT.

Parameters

Another measurement of AI training that are of interest to data centre operators are parameters.

AI parameters are used by generative AI models during training. The greater the number of parameters, the greater the accuracy of the prediction of the desired outcome. ChatGPT-3 was built on 175bn parameters. But for AI, the number of parameters is already rising rapidly. WU Dao, a Chinese LLM first version used 1.75 trillion parameters. WU Dao, as well as being an LLM is also providing text to image and text to video. Expect the numbers to continue to grow.

> “In terms of macro numbers, by 2030 AI could account for 3% to 4% of global power demand. Google said right now AI is representing 10% to 15% of their power use or 2.3 TWh annually”

With no hard data available it is reasonable to surmise that the computational power required to run a model with 1.7 trillion parameters is going to be significant. As we move into more AI video generation, the data volumes and number of parameters used in models will surge.

Transformers

Transformers are a type of neural network architecture developed to solve the problem of sequence transduction, or neural machine translation2. That means any task that transforms an input sequence to an output sequence. Transformer layers rely on loops so that where the input data moves into one transformer layer the data is looped back to its previous layer and out to the next layer. Such layers improve the predictive output of what comes next. It helps improve speech recognition, text-to-speech transformation, etc.

How much is enough power? What researchers, analysts and the press ARE saying

A report by S&P Global titled POWER OF AI: Wild predictions of power demand from AI put industry on edge quotes several sources: “Regarding US power demand, it’s really hard to quantify how much demand is needed for things like ChatGPT,” David Groarke, managing director at consultant Indigo Advisory Group, said in a recent phone interview. “In terms of macro numbers, by 2030 AI could account

for 3% to 4% of global power demand. Google said right now AI is representing 10% to 15% of their power use or 2.3 TWh annually."

S&P Global continue: "Academic research conducted by Alex de Vries, a PhD candidate at the VU Amsterdam School of Business and Economics [cites] research by semiconductor analysis firm SemiAnalysis. In a commentary published Oct. 10 in the journal Joule, [cited by de Vries] it is estimated that using generative AI such as ChatGPT in each Google search would require more than 500,000 of Nvidia's A100 HGX servers, totaling 4.1 million graphics processing units, or GPUs. At a power demand of 6.5 kW per server, that would result in daily electricity consumption of 80 GWh and annual consumption of 29.2 TWh."

A calculation of the actual power used to train AI models was offered by RI.SE – the Research Institute of Sweden. It says: "Training a super-large language model like GPT-4, with 1.7 trillion parameters and using 13 trillion tokens (word snippets), is a substantial undertaking. OpenAI has revealed that it cost them $100 million and took 100 days, utilizing 25,000 NVIDIA A100 GPUs. Servers with these GPUs use about 6.5 kW each, resulting in an estimated 50 GWh of energy usage during training."

This is important because the energy used by AI is rapidly becoming a topic of public discussion.

Data centres are already on the map and ecologically focused organizations are taking note: According to the site 8billiontrees "There are no published estimates as of yet for the AI industry's total footprint, and the field of AI is exploding so rapidly that an accurate number would be nearly impossible to obtain. Looking at the carbon emissions from individual AI models is the gold standard at this time... The majority of the energy is dedicated to powering and cooling the hyperscale data centers, where all the computation occurs."

Conclusion

As we wait for the numbers to emerge for past and existing power use for ML and AI what is clear is that it is once models get into production and use, we will be in the exabyte and exaflop scale of computation. For data centre power and cooling, it is then that things become really interesting and more challenging.

FURTHER READING / REFERENCE

- 1 https://www.sciencedirect.com/science/article/pii/S2666651021000231
- 2 https://towardsdatascience.com/transformers-141e32e69591#:~:text=Transformers%20were%20developed%20to%20solve,%2Dspeech%20transformation%2C%20etc

INTELLIGENT COOLING FOR CRITICAL ENVIRONMENTS

AIRSYS

WWW.AIR-SYS.UK
ENQUIRIES@AIR-SYS.UK

How AI & cloud can help data centres become more sustainable

What is the environmental impact of data centres, how can they become more sustainable, and what is the role of AI?

BY JO DEBECKER, MANAGING PARTNER AND GLOBAL HEAD OF WIPRO FULLSTRIDE CLOUD

A WIDE RANGE of industries are growing at a rapid pace – however, as they grow, so does their data centre energy consumption. Because climate change is such a pressing global issue, it's imperative that businesses have a plan in place to reduce their environmental impact. Data centres account for nearly 1% of energy-related greenhouse gas emissions, and in some cases, cooling systems use up to 40% of the total energy a data centre needs. But the amount of energy and water required to run data centres is no longer the only obstacle they need to overcome: with the rapid adoption of AI and GenAI, the power required for computing resources (and associated power for data centres, either owned or cloud based) is only going to increase, and will have a direct impact on the environment.

However, despite the strain AI currently places on data centre capacity, the future is bright. Used correctly, it has the potential to become a source of savings by optimising and bringing efficiency to energy consumption. As an example, some of the tech industry's biggest players are already developing AI-powered solutions to reduce the volume of energy dedicated to cooling servers, with Google DeepMind developing a way to reduce the volume of energy dedicated to cooling by 40%, and the company building a $1 billion data centre in the UK last month to boost the growth of AI.

What are the benefits of cloud migration?

To keep up with accelerated innovation and near-constant industry disruptions, companies must adapt. Exiting their data centres (which may have been built decades ago using legacy technologies and are thus not eco-efficient in the way they operate) for modern technologies and solutions is one way forward. While successful implementation of AI is the (near) future, adopting the cloud is an important first step, and with the right guidance it can be a successful, seamless transition.

Modern-day cloud infrastructure (both from hyperscalers and top green data centre providers) handles energy requirements and carbon emissions in a highly sophisticated and efficient manner. In most cases, these providers operate almost fully on renewable energy, and sometimes, the heat generated by the IT systems is ploughed back to power cities, thereby creating 100% zero-carbon emission centres.

How can business leaders overcome sustainability challenges and work to reduce their carbon footprint?

While there is a general appreciation and consensus across IT and business leaders towards reducing CO_2 emissions, the challenge has always been around measurements and the ability to clearly quantify the baselines as well as improvements required.

Over the years, there has been a growing understanding around what constitutes the baseline for measuring IT sustainability. It is a comprehensive approach of Scope 1, 2 and 3, with Scope 1 being direct emissions that are owned or controlled by the company, and Scope 2 and 3 being indirect emissions that are a consequence of the activities of the company, but which occur from sources not owned or controlled by it. This makes it slightly complex for organisations driving sustainability to not only control what they can but also to have a deep comprehension of what their suppliers and supply chain are contributing to in terms of eco-friendliness.

In the cloud era, a business's cloud supplier is wholly responsible for how power is regulated or managed, so working with a trusted, vetted organisation – which promotes the use of renewable energy and supports your low/zero waste strategy – is the only way to guarantee that your business's energy utilisation is efficient.

In other parts of IT, businesses need to work with partners who are committing to sustainable manufacturing practices as well as to a 'circular economy'.

It is clear from the above two scenarios that Scope 2 and Scope 3 will be the biggest contributors to overall emissions, and thus it is important for companies to carefully assess partners and providers who bring comprehensive methodologies, measurable practices and new-age solutions to help them achieve their sustainability goals.

What's next for data centres in the near future?

With AI and GenAI, use cases will become more prevalent across businesses over the next year. There will be a huge surge in demand for energy by cloud data centres as well as edge data centres, and providers who can demonstrate a clean and green way of delivering that energy will emerge as the winners. Further, the Climate Neutral Data Centre Pact will dictate that all these providers ensure their data centres (powering the cloud or otherwise) are compliant with the optimal power usage effectiveness (PUE) standards, thus assuring a sustainable future.

AI – An industry snapshot

Last month we held the first BCS Breakfast Club at Moorgate in London which was attended by a select group of investors, developers, architects, consultants, engineers and end users from the data centre sector. We chose the topic of AI and asked participants to share their current usage, challenges, plans and predictions for the future. The discussion gave us a snapshot of views and insights from across the industry.

BY JAMES CARMILLET, DIRECTOR OF COST MANAGEMENT AT BCS

ALMOST ALL OF THE GROUP were in agreement that AI, in some form, can be used to take on time consuming administrative tasks. The hope was that this would free up time for senior executives to train the next generation of data centre professionals to support clients and help address the skills shortage. For many of the organisations around the table, this was the aim for the short to medium term.

ChatGPT is being used regularly

Interestingly, ChatGPT is being used by most of the group members, mainly for reports and research. One of the attendees from a leading investment house said that they are regularly using it for detailed investor reports and memos. They encourage their teams to use it by asking a question, embracing the response but then critically reviewing it.

Another explained that their organisation is currently using ChatGPT to automate some admin, for example, scope to service and any repeatable documents. These are then reviewed by senior staff as they have found that it's easier to do this than

draft from scratch. This review has now been built into their process.

Others in the group don't see ChatGPT as AI but more of an enhancer and they use it to produce reports in just one week that would normally take six months. One organisation is currently mapping their whole business process to then evaluate where AI tools can help. This is being done by a separate team of specialists – not engineers who have different skill sets. But in the future will they be the same people?

There are concerns around its usage

Some members of the group expressed some concerns around security especially around client data and stressed that clear policies on its use are a must. Those that didn't have those in place were in the process of developing them and discussion was given to the need to keep reviewing these as AI continues to evolve.

The importance of thoroughly reviewing any information was underpinned by a recent example in the US where two lawyers and a law firm were fined $5,000 (£3,935) after fake citations generated by ChatGPT were submitted in a court filing. The judge P Kevin Castel said in a written opinion there was nothing "inherently improper" about using artificial intelligence for assisting in legal work, but lawyers had to ensure their filings were accurate.

Some organisations are building their own

Some of our participants were working for organisations that are developing their own AI products, investing heavily in data science and resources. This underlines the strong belief that it will create valuable efficiencies and by building their own they are using data they know and trust that can be enhanced by AI. It was noted that no-one was at the stage where they were using it to do the thinking for you, as yet.

It will change the skill sets needed

Some interesting points were raised about whether this approach to AI will change the skill sets needed in the sector and even prevent junior staff from learning. For example, when you are starting out you might have to file the scope to service documents and have to read every one - that is how you learned the basics. 'If AI gets rid of these 'menial tasks' – how will the new generation learn? Will they just learn to manage the AI systems? Do we risk getting rid of the building blocks of our industry? Is this a good thing?'

Whilst acknowledging these concerns the Group felt that as technology moves forward we have to go with it or get left behind – that maybe we are just changing the building blocks and this is an opportunity to become more intentional in what we train.

The general view was that if we can feed in the correct data we should be able to use AI to speed up cost models and also use AI to stress test schedules

It's not necessarily the silver bullet

The general view was that if we can feed in the correct data we should be able to use AI to speed up cost models and also use AI to stress test schedules. However, is it feasible to input all the knowledge and nuances developed over a 20 plus year career - probably not! Some of the architect attendees felt that design was also an area that might not be suited to AI as you should never put forward a design without understanding why every element is as it is, and AI won't necessarily explain the why.

One participant had asked ChatGPT to design a datacentre to see what the outcome was and it put forward 'an OK process.' However, there were concerns that you can't trust the data – it's free and public. For that reason, some of our participants see AI as a big risk and their current approach is cautious as although they don't want to hold people back and appreciate that AI can spark ideas - we are operating in mission critical scenarios. We could lose control.

Overall the consensus was that as AI is only in its infancy we have to check and review it but as it evolves, the level to which we check it will become less. We will start to trust based on experience and it is then that our jobs will be different. Automation in manufacturing was flagged as an example where we trust machines to churn out complex parts and dimension check them – only manually checking about 1 in 100.

Looking forward

Whilst it is too early to predict the growth and impact of AI, it was generally agreed that early adopters will be the winners and success will be driven by those that invest. In the near future consultants and suppliers will be expected to use and embrace it as they do other tools and it will become the norm. In an ideal world, there should be standard industry AI developed, and to be useful, it needs buy in from all parties or we risk a VHS/Betamax debacle.

Finally, everyone agreed that the end goal is to use AI and reap the benefits because staff shortages are still a real issue. This is a compelling need to be open to it and see if it can help alleviate the workload for existing staff to enable them to refocus their efforts on training and upskilling others.

The AI burden will be met by datacentres. **Can they cope?**

For all the attention being given to AI making things easier and automating tasks, there's a great deal of work that goes into creating AI tools. It's estimated that 80 percent of that work relates to collecting and preparing data. According to IBM, some businesses have embarked on AI projects only to give in after a year spent gathering and cleaning data and having nothing to show for it.

BY NICHOLAS COLE, DATA CENTRE SOLUTION MANAGER, EXFO

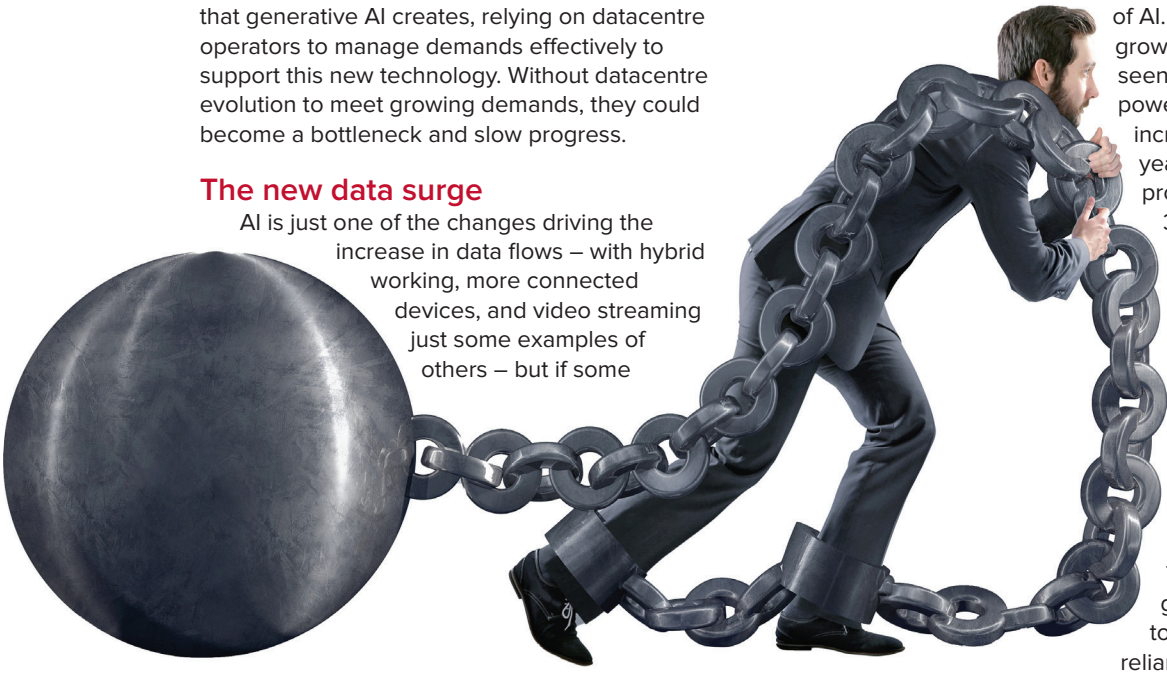
AI IS NOT ONLY data-hungry, it's also energy-hungry and component-hungry. If the current rate of growth continues, the AI industry will use as much energy as the whole of the Netherlands, or around half a per cent of our total global electricity consumption – according to a study from the VU Amsterdam School of Business and Economics. The burden of maintaining the infrastructure to support the AI revolution will fall to datacentres, an integral part of the data economy in terms of storing, managing and processing data.

Datacentres will be the unsung heroes of AI if they can handle the increasing levels of data that generative AI creates, relying on datacentre operators to manage demands effectively to support this new technology. Without datacentre evolution to meet growing demands, they could become a bottleneck and slow progress.

The new data surge

AI is just one of the changes driving the increase in data flows – with hybrid working, more connected devices, and video streaming just some examples of others – but if some predictions hold true, AI will be the change with the biggest effect. We've already seen a big increment in the amount of data used every day. According to McKinsey, 2025 will see the world produce around 463 exabytes of data every day, or 463,000,000 terabytes, which constitutes over 150 times as much as we collectively produced a decade ago.

When people think about the technology required to support increasing amounts of data, it's usually in terms of capacity of transport networks, or wireless technologies like 5G. Often datacentres are forgotten, despite the key role they will play in the ongoing evolution of AI. Yet datacentre growth can be seen through their power consumption, increasing each year by 10%, and projected to reach 35GW by 2030, up from 17GW in 2022 – a figure that does not account for the power efficiencies achieved by the IT industry in recent years.



The accelerated growth of AI tools, and their reliance on huge

data sets, will put a great deal of pressure on datacentres to rethink datacentre infrastructure and the services customers will expect in the future.

Building, upgrading and testing network infrastructure With major cloud and internet service firms leading the way Synergy research forecast hyperscale datacentre capacity to almost triple in next six years. This will be realized through adding to existing sites and building new ones, with Synergy citing 427 projects in the pipe.

These new datacentres will set the blueprint for the AI era as they become supercharged computing platforms for the next generation of app designers. New servers, switches, racks, and cabling will be needed at huge scale to create ultra-high speed, low latency networking fabrics capable of handling massive data sets.

These new servers optimized for AI will house multiple graphical processing units (GPUs) each with high-speed optical transceivers. We could see data rates of up to 1.6TB which is a massive increase over the 25/150/100G connections we see in datacentres optimized for cloud services. This need for speed is due to the collaborative nature of GPUs working on complex AI workloads such as deep learning and neural networks.

GPUs will be connected via transceivers to containerized lead and spine fabrics within pods each transmitting 4 or 8 optical signals packaged in new Quad Small Form Factor Pluggable Double Density (QSFP) and Octal Small Form Factor Pluggables (OSFP). These transceivers may operate with VCSEL-based lasers giving rise to multimode fibre over short distances up to 100m.

Short range multimode fibre at speeds above 100G opens up the need for parallel optical fibre cabling which is referred to as base-8. This uses higher fibre count cables and multi-fibre connectivity often reserved for connections at higher switching tiers. With massive volumes of networking hardware present within AI pods, operators will be faced with weeding out poor performing components such as faulty transceivers and connectors.

New switching capacity will also be necessary to interconnect GPUs in pods and pods to clusters containing 10,000s of GPUs. This is being met by the industry at a rapid rate as vendors have released switches with chipsets capable of moving 51.2 Tb/s of data across 512 lanes of traffic. This is a key enabler for AI systems as operators will be able to build with different configurations such as 32x1.6Tb, 64x 800Gb, 128x 400G or 256x 200Gb through physical ports or virtual breakouts.

Rack design is also getting a makeover, and a pretty major one at that! AI needs power, lots of power and power creates heat, lots of heat; therefore we will see liquid flooding into data halls. Don't worry though because this is not a catastrophic accident – it's by design. Air-cooled systems may not be ideal in all situations so liquid fed through specialist pipes will help fight the battle against heat within racks directly into hardware.

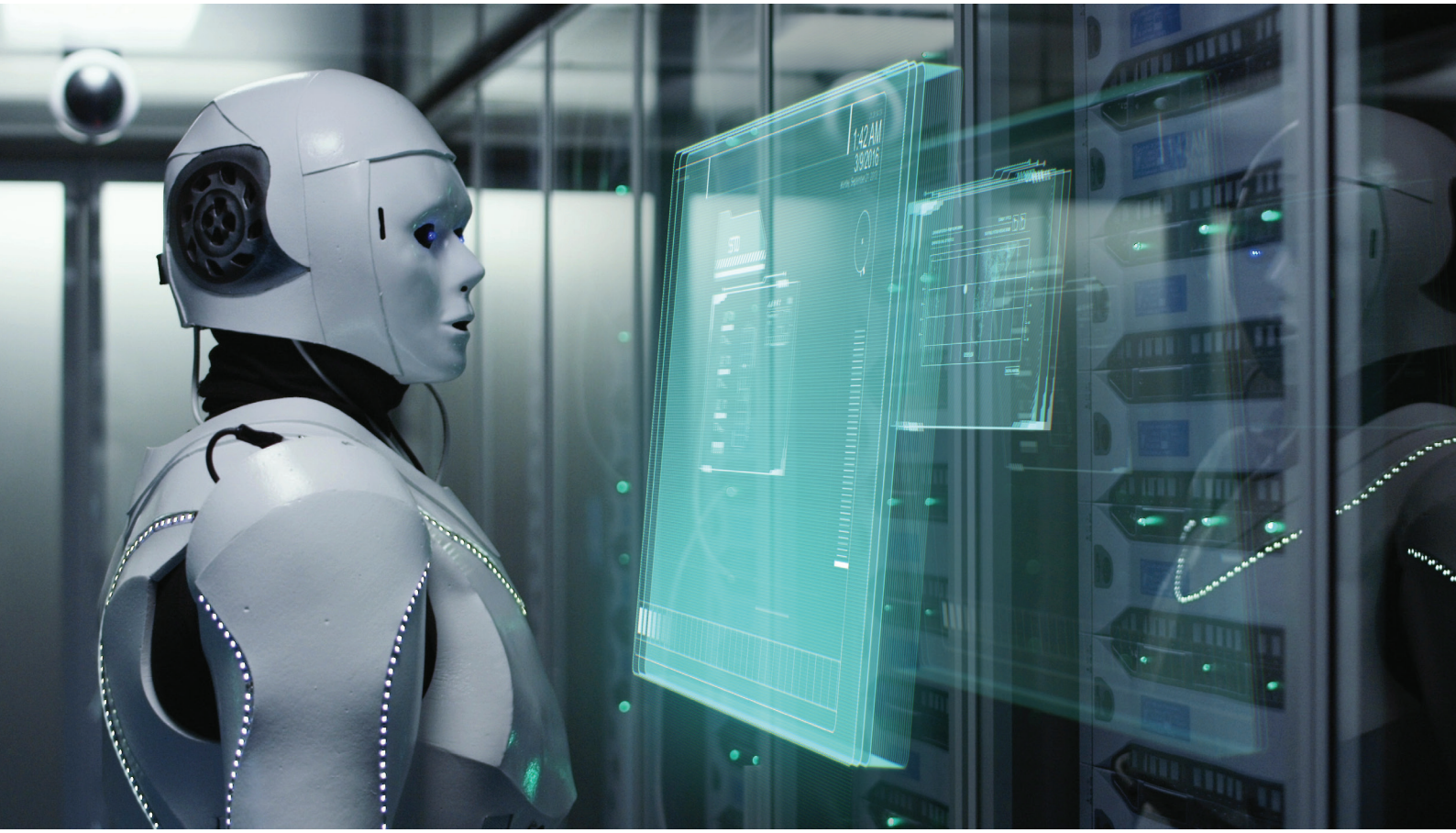
Fibre connectivity and cabling will play a critical part in connecting pods to clusters and clusters to the wider datacentre fabric. Singlemode fibre is expected to continue as the medium of choice as it can be used over longer distances typically up to 500m. Due to challenges with space we may see the emergence of new fibre very small form factor connectors (VSFF) and cassette/panel design. These connectors may terminate anywhere between 2 and 24 fibres depending on optical transceiver selection and system architecture.

With the growing importance of "networking" within datacentres the pressure on component and system performance will intensify. Faulty optical transceivers, fibre reconnectors, and cabling will need to be identified during installation or located quickly as downtime is public enemy number one in datacentres! Multi-fibre certification will ensure loss limits are respected and TX/RX issues are eliminated. Fibre connector inspection should fight contamination problems and transceiver/link testing will identify errors in bit streams and issues related to our new foe: latency.

These new datacentres will set the blueprint for the AI era as they become supercharged computing platforms for the next generation of app designers. New servers, switches, racks, and cabling will be needed at huge scale to create ultra-high speed, low latency networking fabrics capable of handling massive data sets

Bring on 2024 and AI at scale!

Today datacentres, which are already a vital part of the communications ecosystem, have augmented importance given the advent of AI. AI can only thrive with massive sets of data to leverage, and datacentres will ultimately determine their success. As companies jockey for position, adopting solid test practices will make the difference between the average and the elite. So, let's buckle up for the excitement to come as AI makes its way to centre stage in 2024.



Generative AI and the data centre industry

The explosion of generative AI and the associated increase in demand for computing power to implement it is already having an impact on the data centre industry. We are just at the start of the wide-spread adoption of generative AI so we can expect the impact to gather pace over the coming months and years.

BY STUART FARMER, SALES DIRECTOR, MERCURY POWER

AS WITH ANY big technological advance, the rise of generative AI brings significant opportunities, but also risks. The biggest risks in the data centre industry are co-location facility providers becoming overloaded and challenges securing the space and power required for new developments. There is an urgent need for increased investment to mitigate this.

Demand outstripping supply

The headline from JLL's recently published North America Data Centre Report is that record demand is meeting limited supply and this is largely driven by generative AI and the continued adoption of cloud services. We are seeing a comparable situation emerging in the UK and this looks set to continue as businesses across all sectors embrace generative AI.

The JLL report highlights a number of key issues.

Supply chain slowing down

The JLL report points to the supply chain slowing as an exacerbating factor in the imbalance of supply and demand, stating: Supply chain issues and the lack of available land with onsite and scalable power necessary to satisfy future requirements have lengthened data centre development timelines to three to five years, or more.

AI use is increasing

AI use is on the rise. Market research firm Gartner reports that synthetic data will completely overtake real data in AI models by 2030. The Business and Insights Conditions Survey (BICS) from the Office for National Statistics, which was published in June 2023, showed that approximately one in six UK

businesses are already implementing at least one AI application. That figure is sure to rise as more organisations trial generative AI and weave it into different areas of their operations.

Demand for higher density and larger data centre requirements

The JLL report notes that the increasing adoption of AI has caused demand for higher density and larger data centre requirements. Just one example highlighted in the report is that Investment Bank TD Cowen recently reported that 2.1GW of US data centre leases were signed in the second quarter of 2023.

Impact on data centre infrastructure

Outlining the impact on data centre infrastructure, the JLL report notes that AI adoption will need high power density server clusters, with densities of 50-100kW per rack. The report also flagged the fact that many colocation providers have reduced the upfront cost to deliver power to these high-density clusters by increasing the voltage delivered to the floor to 415 volts (in the UK we do this anyway). The report concludes that to meet sustainability goals, innovations will be needed to improve cooling and energy efficiency. We are looking at projects where they are talking about racks requiring 200kW per rack.

Impact on co-location facility providers

The growth of generative AI results in a risk that co-location facility providers will become overloaded.

When I look at the UK specifically, there are many fallow halls that clients have moved out of. In response, Mercury Power is currently designing platforms in these data centres to enable AI and Super Compute to be deployed.

However, there will become a time where power may restrict these deployments as Direct Liquid Cooling reduces PUE and increases the compute power. The physical space required is a fraction of what a traditional deployment has been. Where hundreds of racks were used now this is in 30 or 40 racks. Multiple cooling systems will need to be installed to make the gains in efficiency.

Impact on customers with basic data centre requirements

Not all customers have fully embraced generative AI, and many still have more basic requirements.

The JLL report highlights that major cloud service providers are growing at “breakneck speed” to support workloads. The report goes on to emphasise that these large requirements are causing challenges for users with smaller requirements, due to the difficulty finding space and power. I agree with the report’s findings that planning ahead is crucial for users with smaller requirements, so that they can find space before they need it.

The race is on to push the boundaries on the infrastructure to maximise the opportunities that exist. At Mercury Power we are seeing investment from the global players all the way through to the UK wide Data Centres and regional players, which is encouraging

With the shortages highlighted in the JLL report, which we are also starting to see in the UK along with more deployments being planned and commissioned, there is an opportunity for smaller regional players to offer services to meet the needs of customers looking for smaller or more basic data centre requirements.

Investment requirements

It is likely that we will see increased investment from co-location providers to accommodate for the increased demand.

Inevitably everyone would like to ride the crest of this wave and be the most powerful, fastest to deploy and of course offer the most power dense facilities.

The race is on to push the boundaries on the infrastructure to maximise the opportunities that exist. At Mercury Power we are seeing investment from the global players all the way through to the UK wide Data Centres and regional players, which is encouraging. We are seeing large investment in the smaller data centres, and they are assessing how they can scale to meet the new demands. Looking at flexible M&E designs that can adapt to the ever-changing needs of the clients. This is through driving efficiencies as well as infrastructure investment.

Investment and product development is required from the Cooling OEM’s. They need to develop products that can enable our designs to be more efficient and offer more capacity. This needs collaboration between the AI OEM’s, Data Centre Operators, Data Centre Designers, M&E OEM’s and of course the clients using the data centres.

FURTHER READING / REFERENCE

- Business and Insights Conditions Survey (BICS) from the Office for National Statistics

How eCommerce and fintech firms are leveraging the transformative power of AI-enabled data centres

Andy Connor, EMEA Channel Director at Subzero Engineering, examines the impact AI, blockchain and process automation is having on the eCommerce and fintech sectors.

THE FINTECH and eCommerce sectors are booming. By 2026, the global eCommerce market is expected to be worth more than $8.1 trillion1 annually. Meanwhile a recent study by McKinsey revealed there are now more than 272 fintech unicorns, with a combined market cap of $936 billion – a sevenfold increase in just five years.

Technologies such as artificial intelligence (AI), blockchain and process automation remain critical to sustaining this growth, with companies such as NVIDIA turning to GPU-powered servers to overcome the limitations of conventional data centre infrastructure.

Designed originally to accelerate computer graphics and image processing, GPUs perform complex three-dimensional vector calculations, enabling them to perform multiple operations simultaneously.

Making sense of vast amounts of data

This ability to multitask on a huge scale makes GPUs an ideal tool for managing and analysing vast volumes of data. For example, GPUs, can process neural network training data up to 250 times faster than conventional CPUs. They can also do the job more accurately, making them particularly well suited to high-quality data-driven decision making. For example, credit specialist Capital One uses a suite of GPU-optimised data libraries to accelerate its data science and analytics pipelines. The firm has not only achieved a 100-fold increase in data model training times, it has also reduced its costs by nearly 98%2.

This uptick in processing power means forward-thinking firms now have an opportunity to explore data models faster and with greater confidence. They can also do so in a more cost-effective and energy-efficient manner, and with a faster time to ROI.

Automating and improving processes in financial services

Financial services firms are among those with the most to gain from the huge potential of GPU-powered AI, and many companies are already leveraging this technology to automate and improve mission-critical processes. These use cases include.

- **Algorithmic trading:** analysing historical market and stock data to generate investment strategies, build portfolios and automatically buy and sell investments. Established banks which have developed algorithmic trading strategies include BMP Paribas, Deutsche Bank and Credit Suisse.
- **Detecting fraud:** combatting the most sophisticated types of transaction and identity fraud, increasing fraud detection accuracy, and boosting anti-money laundering and know-your-customer regulation. American Express, BNY Mellon and PayPal are already using a form of AI called natural language processing (NLP) to detect and prevent financial fraud.
- **Accelerating payments:** Fintech payment firms such as PayPal are using machine learning to improve payment authorization rates on their platforms. One way it does this is by predicting

and efficiently managing instances where a bank could decline a payment.

Achieving competitive advantage in eCommerce
eCommerce companies are also waking up to the power of AI, using the technology in innovative ways to achieve competitive advantage. Among some of the innovative use cases in the eCommerce sector include:

- **Predicting and managing customer churn:** AI can analyse customer behaviour and identify which customers are most likely to make repeat purchases This enables sales and marketing teams to target and better allocate resources to more profitable customers.
- **Dynamic pricing:** eCommerce giants such as Amazon and eBay are long-term advocates of AI-powered dynamic pricing. They use the technology to analyse market demand, competitor pricing and other factors so they can adjust their prices in response.
- **Enabling visual and voice search techniques:** Retailers such as ASOS, Forever 21 and Home Depot are using AI to free their customers from their keyboards, find the products they really want and accelerate their path to purchase.

It's interesting to note that financial services and ecommerce, as well as many other sectors, are exploring how AI can drastically improve online customer interactions. Thanks to generative AI (GenAI), clunky chatbots will soon be a thing of the past. Instead, GenAI-powered solutions are particularly good at finding the best answers to customer questions and sharing that information in a human-like way. Deutsche Bank, American Express and Wells Fargo are among the banks that are starting to go live with such GenAI-powered solutions.

Using AI to optimise the data centre

There is clearly huge scope for using GPU-powered AI to improve products and services within the eCommerce and financial services sectors, but the benefits don't end there. The technology is also transforming how data centres critical to these sectors are managed and optimized, how uptime is ensured, and higher levels of sustainability are achieved.

AI is not only helping data centre managers identify, troubleshoot and mitigate outages in a reactive way, it is also automating this process, predicting faults and triggering self-healing mechanisms. For instance, AI can be trained to identify unusually slow traffic within a particular node, and then re-boot a process or the entire node fix the issue. Other tasks such as energy management (including cooling and power management), inventory management and systems update management can also be automated in a similar way using AI.

AI is also being used to boost data centre visibility and decision making, helping managers to identify opportunities to optimize resource allocation and improve both workload management and capacity planning. Such processes can reveal golden opportunities to right-size data centre infrastructure, cut power consumption and reduce environmental impact.

Research by Schroders3 suggests that AI-related data centre power consumption is likely to increase sevenfold to 7GW by 2026. However, this spike in energy consumption and associated carbon emissions can be reduced, at least in part, through the careful use of GPUs. That's because GPUs are more powerful, fewer servers are needed, and data centre physical footprint and cooling requirements are reduced.

Simplifying and accelerating financial transactions with blockchain

Blockchain is another GPU-reliant technology helping to disrupt both the financial services and ecommerce sectors. Fintech firms such as OpenZeppelin are harnessing the power of smart contracts to simplify complex financial transactions. Smart contracts, which are powered by blockchain, automatically execute when certain conditions are met. They remove layers of intermediaries, reduce cost, and speed up contract execution. For example, a smart contract can be programmed to make a payment when a product or service has been successfully delivered.

Meanwhile Ripple, one of the best-known blockchain-based payment systems, enables banks, corporations and crypto exchanges to transfer money without the need for a third-party processor. The firm's solution has made cross-border payments significantly easier, faster, cheaper and more secure.
The data centre as the backbone of innovation
As transformative technologies such as AI and blockchain continue to evolve and become more integral to the success of the fintech and eCommerce sectors, the role of data centres becomes increasingly critical.

Data centres equipped with GPUs deliver the necessary computational power that AI applications require. As we continue to push the boundaries of what AI can achieve, the role of data centres will only grow in importance. They are not just a supporting infrastructure, but a vital component in the journey towards a more AI-driven future.

FURTHER READING / REFERENCE

- 1. https://www.forbes.com/advisor/business/ecommerce-statistics/
- 2. https://blogs.nvidia.com/blog/ai-accelerating-financial-services/
- 3 How AI is set to accelerate demand for data centres (schroders.com)

Data centre sector 2024 market outlook

As the world adapts to the digital transformation of almost every aspect of everyday life, the data centre sector, which plays such a pivotal role in digitalisation, is constantly evolving. The rapid evolution of data centre markets and shifts in the architecture and design of data centres means that as we start a new year, the data centre industry is navigating a complex maze of challenges and opportunities that promise to realign its trajectory.

BY SIOBAHN MIEKLE, MANAGING DIRECTOR UK AND IRELAND OF EATON CORPORATION

IN FACT, it is probably true to say that the data centre industry is undergoing its own transformative journey, one which involves leveraging cutting-edge technologies to optimise operations, streamline resource allocation, and enhance security protocols. From improving sustainability and energy efficiency while simultaneously managing rapid growth in demand – effectively doing more with less, to complying with regulatory changes, it is fair to say that events in 2024 will be far from 'business as usual' in the data centre sector.

Factors impacting growth

Several factors are impacting the growth of the data centre market. These include the continuous development of data-intensive industries, accelerated by the widespread continuous adoption of the cloud by organisations across the globe. Innovative new Internet of Things (IoT) and Artificial Intelligence (AI) technologies are another significant factor.

AI and other digital solutions have an increasingly important role to play in data centre lifecycle management. Not only do such capabilities monitor the current state of the data centre they also provide deep operational insights and perform predictive analysis, too, allowing data centre operators to manage performance, as well as understand where their assets need to be, and build the path they need to get there.

But these capabilities also represent a challenge. Much AI software is trained on vast quantities of data, so the emergence of technologies like AI has led to an enormous surge in computing demand. According to TIRIAS Research, the data demands of Generative AI applications such as ChatGPT could result in a 50x increase in the number of workloads processed worldwide by 2028. Such exponential growth in demand, along with ongoing advancements in technology, could see a change in customer requirements around data centre design and architecture. This, in turn, may result in the need for new and different solutions. At the very least, data centre operators will face an urgent need for more data centre space and increased density per rack as they maximise or expand their existing investments. They will need to find means of meeting that increased power demand that can be deployed quickly and managed effectively.

To boost availability fast, prefabrication is already emerging as one of the answers to the challenge. A modular data centre is a portable method of deploying data centre capacity that can be placed anywhere it is needed, so expect to see this trend accelerate in 2024. Modular data centres typically consist of standardised components that can be added, integrated, or retrofitted into an existing data centre or combined into a system of modules.

Power constraints

In many countries there are now restrictions on adding major new loads to the grid. Permission to build or expand is becoming more difficult to obtain, prompting questions over where data centre clusters are located. Until now, the conventional hubs for data centres have been major cities like Frankfurt, London, Amsterdam, Paris, and Dublin. This is due either to the presence of company headquarters or the natural economic advantages stemming from robust telecom connectivity and ideal client demographics. However, there is

a shifting trend towards moving away from these traditional hubs and establishing data centres in smaller cities across 'Tier Two' countries where power constraint may be less of problem and relevant permits easier to obtain. The drawback of this, of course, is the potential lack of skilled workers. Nevertheless, expect to see this trend develop in 2024 as cities less well-known for data centre clusters, such as Warsaw, Vienna, Istanbul, Nairobi, Lagos, and Dubai appear on the radar to avoid power constraints.

Sustainability challenge

With data centres being central to almost everything we do, public awareness of their presence and power requirements is inevitable and, with this, comes environmental responsibility. Data centres are the backbone of the digital world and therefore impact how society reduces emissions to combat climate change. In 2024, achieving sustainability goals will continue to be vital, and the data centre industry will bear scrutiny from regulators such as the European Union (EU) and US Security and Exchange Commission to ensure it makes progress. Balancing growth and sustainability will undoubtedly be challenging, but as the industry steers toward a more sustainable future, data centres will need to adopt innovative strategies to minimise their overall environmental footprint. This imperative aligns with a broader industry-wide commitment to a more holistic approach to system performance which recognises the interplay between technological advancements and environmental impact. Energy efficiency will play a significant role. Achieving more, with the same amount of energy or less, will be a recurrent theme throughout the year ahead that will continue into the foreseeable future.

There will be growing emphasis on addressing what are known as 'Scope 3 emissions' under the Greenhouse Gas Protocol, the most common standard for carbon accounting. Currently, much of the market is focused on Scope 1 and 2 responsibilities, which include carbon usage effectiveness, power usage efficiency, water usage efficiency, amongst others. Scope 3, however, involves looking at a data centre from a total life-cycle perspective in terms of sustainability, shifting the emphasis from metrics to entire system performance. Digital will be the key method for measuring and managing this performance, using software that takes account of aspects such as availability, scalability, flexibility, and commercial efficiency for the most thorough and detailed sustainability reporting.

Significantly, this heightened awareness is propelling the sector towards adopting eco-friendly practices. Expect to see diesel generator replacements, battery energy storage systems, on-site generation, advanced cooling, and other technologies taking centre stage. Seeking out renewable energy sources to replace or bolster existing supplies and exploring the options to introduce or increase on-site power generation from assets such as solar panels and wind-turbines, will be part of this. So, too, will be a renewed emphasis on how data centres can work in tandem with the grid. We are all likely to continue to hear more about the grid interactive data centre in 2024 and data centre operators will doubtless want to find out more about this option which brings with it the benefit of not only helping to decarbonise the grid but also contributes to their own sustainability goals.

Regulatory pressure

Unsurprisingly, the data centre industry has attracted the attention of the regulatory bodies charged with delivering the United Nations goal of carbon neutrality by 2050. In Europe, the EU is leading the way, rapidly reframing the conversation around data centres via a root and branch review of some of its core directives including the recast Energy Efficiency Directive.

The revised directive, published in September 2023, introduces an obligation for EU Member States to monitor the energy performance of data centres, and states that an EU-level database will collect and publish information relating to the energy performance and water footprint of data centres, making this far more than a cosmetic exercise.

In fact, the European Commission has ordered a study on the energy performance and sustainability of data centres meaning that the data centre industry should be ready for policymakers and regulators to take a much tougher stance on environmental integrity. The message to the industry must be that it is better to act now, in anticipation of tighter rules, than respond later. Taking responsibility into its own hands will serve the industry well, certainly in reputational terms and probably financially, too, because reduced energy and water use is likely to save money in the long term even when the costs of improvements are factored in.

A bright future

Building a robust digital strategy may seem daunting, but in today's landscape it is an absolute necessity for data centres to evolve in line with all the trends touched upon in this article. With increased demand for computing driving the need for speed and optimisation of operations, it is key to ensure that the data-rich environment is managed efficiently.

Digital transformation is affecting every sphere of life, and the way we do business is changing. Clearly, digitalisation is good for the data centre industry and AI can help the industry to become more predictive, rather than simply reactive. The future looks bright for the data centre sector in 2024. Certainly, it will be characterised by a dynamic interplay of technological advancements, sustainability imperatives, and a relentless pursuit of operational excellence, but data centres have scope to grow responsibly and sustainably to become true hubs of digital innovation – a worthwhile new trajectory for the industry and for us all.

Embracing prefabricated modular data centres for scalable growth

Prefabricated modular data centres have emerged as a powerful solution in the digital age, offering unparalleled benefits for rapid and scalable growth in the data centre industry.

BY ALEX BREW, REGIONAL DIRECTOR, NORTHERN EUROPE AT VERTIV

BOOSTED BY the rise of artificial intelligence (AI), the digital era is ushering in a new age of technological advancements and unprecedented demand for data processing and storage. In this rapidly evolving landscape, data centres play a pivotal role as the nerve centres of our interconnected world. To meet the evolving needs of hyperscalers, enterprises and digital services, data centres are undergoing a transformative shift towards optimising efficiency and adaptability in both construction and operation.

One key driving force behind this transformation is the widespread adoption of prefabricated modular data centres. These innovative solutions offer a host of benefits, revolutionising the way data centres are designed, built and operated.

The power of this approach

Prefabricated modular data centres, commonly referred to as PFM data centres or integrated solutions, offer a multitude of advantages that revolutionise the way we approach data centre infrastructure. These benefits extend across various aspects of design, construction, performance, and geographic deployment.

One of the most prominent advantages of PFM data centres lies in their ability to expedite the design and deployment process. This is achieved through a unique prefabrication process, involving the off-site manufacturing and assembly of capacity units. Unlike conventional on-site construction, this approach allows for parallel activities, making it significantly more efficient. As a result, these modular data centres can be designed, tested and made operational in a remarkably shorter time frame compared to traditional construction methods. This rapid design capability is crucial in addressing the pressing need for data centre capacity, especially in the face of unpredictable and growing demand.

Furthermore, the modular nature of these data centres allows for a more agile response to demand. Instead of planning for unforeseen growth, organisations can build capacity to precisely match their current requirements. The modular approach enables them to scale their data centre infrastructure in a modular fashion, effectively aligning capacity with business demand. This minimises the risks associated with either overbuilding or underbuilding, ensuring a more efficient allocation of resources.

Another notable benefit of is the optimisation of component performance within a holistic system. These data centres take an integrated approach, enhancing the performance and efficiency of individual components by tightly integrating various systems, including power, thermal management, and IT components. By designing, configuring, and fabricating these components off-site, they work seamlessly together, contributing to the reliability and overall performance of the data centre.

Quality control is a crucial aspect of data centre construction, and prefabricated data centres excel in this area. The manufacturing and assembly of units in controlled environments result in higher quality controls. This not only increases the reliability of the components but also extends the geographic reach of data centre deployment. Prefabricated units can be transported and assembled in various locations, even in regions where traditional construction methods might face challenges. This flexibility in deployment enhances the adaptability and scalability of prefabricated modular data centres.

Delving into the distinction between standardisation and localisation

While the approach of standardisation is widely recognised for its role in streamlining and enhancing data centre deployment, it is essential to appreciate the subtle yet significant disparities between standardisation and localisation. These two methodologies each bring their own set of considerations and implications to the table.

Standardisation is rooted in the principles of uniformity and consistency. It involves the adoption of pre-established designs, components, and practices that are designed for replication across diverse deployments. Embracing standardisation empowers data centre operators to swiftly deliver a consistent and seamless experience in terms of infrastructure, functionality, and operational procedures. This approach not only facilitates the harmonious integration with existing systems but also paves the way for efficient scalability.

Conversely, localisation places a strong emphasis on customisation and adaptability to meet regional requirements and preferences. This approach is particularly pertinent in regions where regulations exhibit significant variations. It acknowledges that different geographical locations may be subject to distinct building codes, regulations, standards, and environmental considerations that wield a substantial influence over data centre operations.

Unlocking long-term success

So, while there are certainly subtleties to consider, as evidenced by the difference between standardisation and localisation, we can clearly see that as the data centre industry continues to evolve, the benefits of prefabricated modular data centres become increasingly evident. Accelerated deployment timelines, reduced costs, improved operational efficiency, compliance with regional requirements all contribute to the long-term success and establishment of this trend.

Prefabricated modular data centres have emerged as a powerful solution in the digital age, offering unparalleled benefits for rapid and scalable growth in the data centre industry. By embracing prefabricated modular data centres, or standardised elements of the approach, companies can harness their potential to unlock new opportunities, enable efficient and reliable data processing and storage, and drive economic growth.

The risks of data centre theft

Some of the key operational elements to ensure physical security extends beyond the walls of the data centre.

CHRIS WELLFAIR, PROJECTS DIRECTOR AT SECURE I.T. ENVIRONMENTS

DATA BREACHES are nothing new, whether caused by the accidental loss of staff devices, inappropriately shared data, or a cyber attack. The latter is the most common form of theft that we hear about impacting the data centre, and many readers will have experienced the mayhem and destruction that can follow the realisation that you have become a victim. 29% of businesses reported experiencing some kind of attack or breach in its last annual Cyber security breaches survey, released in April 2023.

Physical attacks

Today's data centres are far from 'just' application and data repositories, they can control entire production lines, communication systems, enable remote working for large numbers of employees, as well as be responsible for security features such as access control and other building control systems including air conditioning. Some data centres even operate at unstaffed locations, where they may be responsible for utilities such as electricity distribution, telecommunications, or flow control on a gas pipeline.

Cyber attacks and remote data breaches are not the only type of theft experienced in small and large data centres. The most common type of crime facing businesses in the latest government Commercial Victimisation Survey (2022, released October 2023) is theft at 15%, followed by burglary (including attempted) and vandalism at (9%). Medium-sized business (up to 49) employees were the most likely to experience crime (38%), but the figure across the UK was still over 1 in 4 businesses experienced some sort of crime.

The price of data centre crime

There are of course valuable items in data centres. Assets that criminals can choose to sell on the black market, but criminals are not always looking to sell complete PCs or UPS equipment. For many it is the raw materials where the value exists and they do not care how much vandalism is caused in the process of getting it. UK metal theft offences had been declining since 2012, reaching a low of 13,033 recorded offences in 2016/2017, but have been on an upward trend since then, more than doubling to 29,920 offences in 2021/2022, and only a marginal

5% dip in the last 12 months. The open market price of copper, for example, has risen 39% in the last 4 years.

Whilst most companies will have some kind of back-up system in place, it is difficult with physical data centre theft and vandalism to be sure those systems will kick-in, particularly if those servers are located in the same data centre. It can be hard to prepare for every failure mode that could exist, and just as there is a cost for downtime, there is a cost with having different levels of failover and redundancy in place.

Downtime invariably leads to lost productivity and reputational damage, particularly if the downtime is prolonged and has an impact on customers, or other parts of the supply chain a company sits in. But physical theft can also be a data breach and a long conversation with the Information Commissioners Office can be a difficult experience to go through, never mind the explanations required to those whose data has been breached. The ICO may not always fine, but its penalties and findings are always public, though always with the intention to help you make sure the chance of breaches is reduced in the future.

Get your design right

When designing a data centre, security has to be at the heart of the discussion. It is easy to find yourself focused on the design 'within the walls', which whilst important, will have little impact from a security perspective. Ultimately, if an intruder can get as far as the data centre, they are already in a position to cause vandalism and downtime at a minimum.

That is not to say that the data centre structure itself shouldn't be considered, and all those involved in the design process would do well to get very familiar with the Loss Prevention Certification Board's LPS1175 standard. The aim of this standard is to assess the physical resistance of security products when various types of unauthorised access tools are used against them. Depending on how a product performs it is given one of five different grades, according to the time and tools likely to be used by somebody wanting to subvert those products to get at whatever they are protecting. Essentially the standard provides a buyer's guide that those designing a data centre (or anything else that needs protecting) can use to ensure the selected products meet the level of protection they require.

When considering physical security, it is important to remember that the walls of the data centre are not where your physical security should start. Yes there should be CCTV, security lighting, alarms and access controls at the data centre itself, but if an intruder has reached that point, the risks are already higher than they need to be. Think about the key areas on site where power or fibre pairs arrive at buildings? Is there a zone around the data centre where only certain staff should have access, and how is that controlled or policed? These zones too should be fitted with access controls, CCTV and alarms. These barrier zones create an opportunity for security teams or the police to be alerted to a threat, and gain crucial time to act. They also act as a deterrent.

Operationalise your security

Your cyber security is constantly monitoring your network, receiving updates, controlling access, analysing, and encrypting data to ensure customers, suppliers, applications, and the business are protected. So often, with physical data centre security, once the initial design and test is done, many companies fail to ensure that regular tests are carried out, alerts work correctly and keyholder lists are correctly maintained. One of the key reasons we have seen this occur, is because physical security is often seen to fall between facilities/operations teams, and IT staff. Each thinks the other 'has it covered'. This can lead to 'black holes' areas with faulty CCTV devices, or access lists not being tightly controlled.

Create clear roles and responsibilities for staff and teams, maintenance and testing regimes, and audit your security architecture at regular interviews so that processes and equipment can be replaced as needed.

Stay one step ahead

The essential thing to remember when considering security, is that you are not just trying to stop criminals breaking in, you are trying to make it harder and create a deterrent. In the same way that your software and hardware requirements evolve to meet the demands of user needs and the security threats that you may face, physical security must adapt to. The physical assessment of a data centre and its surroundings are as important as the way you assess and specify the hardware that sits in it.

AI Focus by DCA CEO Steve Hone

ANYONE who was at DCW 2024 recently will know that AI and Sustainability is pretty much all the data centre industry is talking about at the moment!

In The DCA's feature in the last issue of Data Centre Solutions our feature focused on Sustainability, in this issue it is time to find out more about AI. I have three informative and thoughtful articles for you; I hope you find them interesting.

Alan Brown, Strategy Advisor, Entrepreneur, and Professor in Digital Economy at University of Exeter article sets the scene by 'Demystifying AI for the Data Centre' – Alan explains some of the hype around AI and goes onto describe what AI means for DC's.

Paul Morrison, CMO Stelia, thoughtfully outlines how implementation of AI in DC's will provide changes to how DCs operate, with AI effecting functionality, sustainability (thought I would drop that word in) and much more!

Asim Rasheed, Senior Product Manager – Spirent Communications tells us more about how AI workloads are going to be a real test of Data Centre performance and provides his advice.

I'd very much like to thank Alan, Paul and Asim for taking the time to write these articles which I hope you will find interesting.

To find out more information about The DCA click here or email us mss@dca-global.org

Demystifying AI for the Data Centre: Opportunities and challenges of the next generation

By Alan Brown, Strategy Advisor, Entrepreneur, and Professor in Digital Economy at University of Exeter

THE DATA CENTRE INDUSTRY is at a pivotal point. As the world generates an ever-increasing volume of data, driven by the Internet of Things (IoT), big data, and cloud computing, the demand for efficient and scalable data centre infrastructure continues to soar. Artificial intelligence (AI) presents itself as a powerful tool in this evolving landscape, on the one hand driving increased data centre demand and on the other offering a plethora of opportunities for optimization and improvement of data centre services.

However, understanding AI's impact on business and harnessing the full potential of AI requires a clear understanding of its core characteristics and the challenges associated with its integration within data centre operations. What are the key elements of AI that are driving digital disruption in this sector? And how can AI best be applied to improve data centre performance?

Understanding AI: Hype vs. Reality for Industry Leaders

We live in a confusing world. Technologically speaking, we have entered a period where advances in so-called "smart" digital products and services are all around us. Often, they are out in the open where they can be seen, such as your bank's mobile app or the devices you buy for the home to play music, turn on the lights, and control the heating. However, more and more we see these capabilities buried inside almost all the products and services we've been using for some time: From TV and washing machines to the clothes we wear.

However, beneath the surface of these product enhancements, a more substantial shift is taking place. Today's digital transformation involves data-driven innovation bringing advanced predictive insights and broad automation of many of the tasks we now do for ourselves. Inevitably, this is opening up new business models, bringing improvements in performance, and driving product quality. However, more fundamentally it is also changing our relationship with the world around us and challenging our understanding of what is based on human judgement and decision making, and what is not.

No surprise that the hype surrounding these digital solutions has continued to grow. The expectations raised by increasing digitization over the past 20 years are now being realized by a convergence of advances in data analysis, access to new digital sources of data, high speed connectivity, and raw computing power. Seen together, they are enabling

rapid advances in how digital solutions are designed and delivered, and driving significant changes across many domains in both the public and private sectors.

While stricter and more robust definitions exist, most people now describe these broad capabilities to be part of the new wave of Artificial Intelligence (AI) systems. Generally, AI refers to any system capable of mimicking tasks traditionally requiring human intelligence. To achieve this, most AI systems rely on machine learning (ML), a technique that utilizes vast amounts of data and computing power to build and validate decision-making logic. This logic forms the core of an AI model. The AI system then feeds new data into this model, which generates human-like decisions based on the learned patterns.

In one typical scenario, more and more data is gathered from a variety of sensors contained in internet-connected devices. This could be from many sources including from domestic devices in the home, environmental sensors on the street, or physical monitoring equipment in a factory. By collecting this data, analysis is possible that explores the data to look for patterns. That is, the creation of algorithms to recognize common situations or anomalies, and solve problems by learning from earlier experiences to apply that knowledge in unfamiliar contexts.

To achieve that feat, what we're experiencing today is largely based on knowledge management techniques that use brute force application of very large computing resources to process large data sets and examine them against extreme numbers of possibilities and variations. By "training" systems with a lot of data about known situations, it is possible to create a form of intelligence that compares the new situation to what has been seen before and come to a set of likely conclusions guided by applying advanced statistical techniques and assessing large numbers of probabilities. While this may all seem rather mundane, by using these basic techniques, AI can deliver a wide set of valuable solutions across various applications. Kathleen Walch has classified these into 7 styles of AI solutions that take advantage of these capabilities:

Hyperpersonalization — using AI to develop a profile of individuals, and then having that profile evolve and adapt over time based on activities being monitored.

Autonomous systems – combining hardware and software to accomplish a task, reach a goal, interact with their surroundings, and achieve an objective with minimal human involvement.

Predictive analytics and decision support – deploying AI to understand how past or existing behaviours can help predict future outcomes or help humans make decisions about future outcomes.

Conversational AI – supporting interaction between machines and humans across a variety of media including voice, text, and images.

Exception management – applying AI to seek patterns in data sources, learn about connections between data points to match known patterns, and searching for anomalies in data.

Recognition – training AI to identify objects and features in images, video, audio, text, or other unstructured data.

Goal-driven activity – learning rules and applying AI to apply those rules to find ways to achieve stated goals in areas such as strategy, role playing, gaming, and other activities.

In isolation and in combination, these patterns of AI use enable us to address many different problem areas in a large number of domains. Much of what we see used in organizations today are attempts to redefine the challenges they face to be more amenable to being addressed by these AI patterns.

AI in the Data Centre

With these capabilities being provided through robust AI tools offered by each of the main digital technology infrastructure vendors, all industries, organizations, and domains are feeling the pressure to introduce AI capabilities. In the world of the data centre, this kind of AI is already common, including machine learning for optimized resource allocation and reduced energy consumption, predictive maintenance to prevent equipment failures, anomaly detection to identify security threats and malfunctions, and self-optimization to maintain optimal temperature, humidity, and other operational factors.

The growing adoption of these capabilities offers a compelling vision for the future of data centres. By continuing AI adoption and looking for additional ways to leverage AI, data centre operators can expect to:

Enhance efficiency: AI-driven optimization can lead to significant reductions in energy consumption and operational costs.

Improve reliability: Predictive maintenance and anomaly detection minimize downtime and ensure uninterrupted service delivery.
Increase scalability: AI can automate resource allocation, enabling data centres to seamlessly adapt to fluctuating workloads.
Gain deeper insights: AI can analyze vast amounts of data to uncover hidden patterns and trends, providing valuable insights for data centre management.

However, the integration of AI into data centre operations is not without its challenges. These benefits will only be realized if several important issues can be addressed. Three issues are particularly critical.

At the head of this lists is data security and privacy. AI relies on vast quantities of data, raising concerns about data security and privacy breaches. In practice, overcoming issues with data accuracy, quality, and bias must also be addressed. As a result, robust data governance policies and stringent security protocols are essential.

In addition, successful AI adoption also requires that any AI use offers a clear approach to explainability and bias. AI algorithms can be complex and non-transparent, making it difficult to understand their decision-making processes. This raises concerns about potential bias and the need for explainable AI models whenever AI is used in mission-critical situations.

Finally, progress in use of AI will be severely limited without appropriate investment and talent development. Implementing AI solutions requires significant investments in infrastructure, software, and skilled personnel with expertise in both data science and data centre operations. These are expensive, in high demand, and require continual refreshment.

The path to an AI future

The past few years have seen rapid progress in digital transformation across every domain. As digital technologies have evolved, AI has brought together many of these advances to drive a new wave of progress.

The large datasets and complex algorithms that underlie AI require massive storage and compute power. Consequently, demand for data centre capabilities and capacity will continue to climb.

In addition, AI presents a transformative opportunity for the data centre industry, promising to enhance efficiency, reliability, and scalability. However, navigating the challenges associated with data security, explainability, human oversight, and investment is crucial to ensure its successful and responsible integration into everyday use.

To drive success, leaders and decision makers must focus their efforts:
Develop a comprehensive AI strategy: This strategy should clearly define goals, identify potential applications, and outline ethical considerations for AI implementation within your data centre.
Invest in data governance and security: Implement robust data governance policies and invest in advanced security solutions to safeguard sensitive data and mitigate potential privacy risks.
Foster a culture of continuous learning: Bridge the skills gap by upskilling and reskilling your workforce to embrace AI and ensure the successful collaboration between human expertise and AI capabilities.

By embracing AI thoughtfully and strategically, data centre leaders can unlock its transformative potential, ensuring continued growth and efficient operation in the face of ever-evolving technological and data demands.

AI workloads put data center performance to the test

By Asim Rasheed, Senior Product Manager, High Speed Ethernet Products, Spirent Communications

AI WORKLOADS are driving an unprecedented demand for low latency and high bandwidth connectivity between servers, storage, and GPUs. To successfully support AI's rapid growth, data centers and the high-speed networks they rely on must transform intelligently along a path filled with uncertainties.

Artificial Intelligence (AI) is the hottest topic of the day. It screams from headlines. Companies are scrambling to establish an AI position. AI's massive potential is increasingly understood and appreciated, from the arts to the sciences, and beyond, thanks to Generative AI applications like ChatGPT.

As countless predictions are made about AI's value, a growing chorus of concern demands a cautious approach.

One thing is clear: we're only seeing the tip of the AI iceberg. As it evolves, just like the iPhone before

it, AI is poised to become exponentially more than anything that can be imagined today.

The communications industry is already being impacted by AI, which is having a massive effect on the data center workloads that are also catering to edge computing and cloud-based 5G networks traffic.

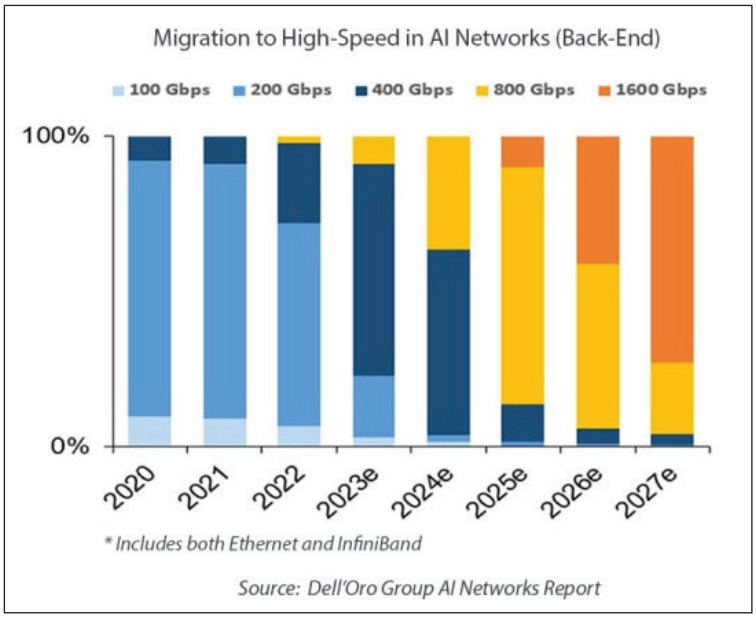
Large cloud service providers (CSPs) are seeing the earliest impact from massive AI workloads. Data center operators are right behind, already grappling with terabit networking thresholds to handle the projected AI demand for bandwidth and compute. Achieving these terabit thresholds can't be solved by adding more server racks or fiber runs. Data centers need to be rearchitected to meet the explosive growth of AI workloads.

In this blog, we recap insights from a joint Spirent and Dell'Oro webinar on "The Impact of AI Workloads on Modern Data Center Networks".
AI demands data center network transformation
AI models are growing in complexity by 1,000 times every three years, requiring low latency, high bandwidth connectivity between servers, storage, and xPUs (a device abstraction that can be mapped to CPU, GPU, FPGA, and other accelerators) [BP1] for AI training and inferencing (the generation of AI intelligence).

There's no way around it: data centers and the high-speed networks they rely on must transform to efficiently and sustainably support AI's rapid uptake. The complexity and size of AI applications dictate the number of xPUs needed to run the apps, the amount of memory, and the type and scale of the network fabric needed to connect all the xPUs. As the scale of AI applications is skyrocketing, requiring thousands to tens of thousands of xPUs and trillions of dense parameters in the near future would not be surprising.

With that kind of scale, a data center can't just keep adding racks. To handle large AI workloads, a separate, scalable, routable back-end network infrastructure is needed for xPU-to-xPU connectivity. AI apps have much less of an impact on the front-end Ethernet networks that provide data ingestion related to the training process of AI algorithms. The requirements for this separate back-end network – which relates to AI inference – differ considerably from the traditional data center front-end access network. In addition to five times more traffic and increased network bandwidth per accelerator, the back-end network needs to support thousands of synchronized parallel jobs and data- and compute-intensive workloads.

Since the progression of all nodes can be held back by any delayed flow, network latency is a critical issue for AI workload performance. Even before the anticipated massive AI workloads, latency is a problem. According to Meta, on average, 33% of AI



elapsed time is spent waiting for the network. Such delays incur timeouts that impact customer service, are costly, and impede scalability.

AI workloads are driving an unprecedented need for back-end network low latency and high bandwidth connectivity between servers, storage, and the xPUs that are essential for AI training and inferencing.
Adoption of high-speed networking
We are at the early stages of data center design evolving to start catering to AI workloads.

Dell'Oro Group provided a forecast for 2023-2027 that addresses questions many companies are asking about the timing and adoption rate of high-speed networks.

Front-end network ports are expected to remain Ethernet. Initial adoption of next generation speeds will be initially driven by front-end connectivity to AI clusters for data ingestion. By 2027, Dell'Oro expects one third of overall Ethernet ports in the front-end network will be 800 Gbps speeds or higher.

In contrast, back-end AI networks are projected to migrate quickly to nearly all port speeds being at 800 Gbps and above by 2027, with triple-digit CAGR for bandwidth growth. Back-end networks will include both Ethernet and InfiniBand, which are expected to co-exist for the foreseeable future.

One size doesn't fit all deployments

AI data center network back-end deployment approaches for AI applications are already quite variable, with hyperscalers Google, Microsoft, and Amazon taking different paths. Deployments like AI training require a lossless back-end network such as InfiniBand. Other implementations prefer standardized, well-understood Ethernet and some use both InfiniBand and Ethernet.

One solution doesn't fit all needs, and convergence on a single path is not expected any time soon. Factors and tradeoffs that influence modern data center architectures include:

- Size of deployments and number of clusters
- Complexity of applications and workloads
- The relative importance of low and deterministic latency, as well as high bandwidth, for the AI applications
- Bandwidth and load balancing application requirements, e.g., the number of lanes in an 800 Gbps channel
- Whether compute- and time-intensive AI training will be outsourced or performed in-house
- Standardized versus proprietary technologies and their anticipated evolutions to meet the needs of AI
- Desire for a diversified, multivendor supply chain
- Pricing

The AI data center journey is just beginning and will change dramatically as AI evolves. Even the hyperscalers are trying to determine the best fabric for their AI workloads for today and for the near future, recognizing that data centers being built today which are not properly planned may be obsolete in two years.

Validating high-speed, low latency AI networks

As AI technology innovations continue their rapid evolution, the networks they rely on must be validated and tested to ensure they meet the needs of growing AI workloads.

As a leading provider of test and assurance solutions for next-generation devices and networks, Spirent provides test solutions to support AI's rapid growth.

Learn more about AI networking challenges and solutions in this joint Spirent and Dell'Oro webinar on the impact of AI workloads on modern data center networks.

The AI effect: Redefining data centres and telecom infrastructure for the next generation

By Paul Morrison, Chief Marketing Officer, Stelia

AS DATA CENTRE operations and telecommunications leaders navigate the demands of operational and business digital transformation with colocation, on-premise, and multi-cloud models, Artificial Intelligence (AI) has appeared as both a formidable challenge and a transformation accelerator. For obvious reasons, the convergence of data centre modernization with telecom innovation requires a strategic approach and rigorous governance. This dual evolution demands careful integration with existing network tools and infrastructure, low level thinking on addressing cybersecurity vulnerabilities, and at least some contemplation of the ethical dimensions of AI deployment.

However, the potential benefits of AI integration extend far beyond operational enhancements, promising a revolution in both data centre management and telecom service delivery. This article outlines AI's intricate path through the technological story of our times, catalysing human ingenuity, and the potential for novel insights going beyond traditional limitations.

AI is poised to redefine sustainability in the sector, from the optimization of power and cooling systems to the timely forecasting of network workloads, addressing the soaring energy demands of interconnected global communications.

Careful implementation could see AI redefining the rigid frameworks of data centres and telecom networks alike, ushering in a new epoch marked by agile infrastructures capable of real-time,

adaptive responses. While AI promises this new era of enhanced insight and efficiency, it is the responsibility of the visionaries behind the technology to guide this journey. As AI reshapes the architecture of connectivity and computation, industry leaders are presented with a strategic decision: to utilize AI as a mere tool for operational refinement or to embrace it fully, forging a partnership that extends the boundaries of human potential and redraws the landscape of telecommunications.

AI promises to revolutionize efficiency and resilience

To date, process-driven approaches to reduce data centre outages have not reduced downtime incidents or severe impacts as much as expected. In fact, the stats are headed in the wrong direction. The Uptime Institute recently noted that over 60% of outages now cost over $100,000, up from 39% in 2019. Outages costing $1M+ also increased from 11% to 15%. Rather than merely removing humans from the loop, AI presents an opportunity to augment our best capabilities, putting people back in control with enhanced insight and reduced complexity. With proper governance and strategy, AI could succeed where policy-led efforts have fallen short.

For example, machine learning algorithms could analyse historical telemetry, infrastructure topology, and documented failure scenarios to identify risk patterns difficult for human data centre operators to discern in siloed data sets. Operators tapping into these AI-generated insights could then take data-driven thoughtful actions to strengthen vulnerabilities before outages occur.

AI presents unprecedented opportunities to enhance data centre operations through automation. By assimilating vast analytical capabilities, AI can optimize workloads, infrastructure, and staff augmentation at new scales. Machine learning will enable predictive maintenance and management, versus reactive approaches. Data centres stand to gain dramatically improved resilience, efficiency, and responsiveness from AI.

Specifically, AI could enable advances like:

- Predictive diagnostics prescribed for assets using telemetry analysis, reducing downtime through repairs made before failures, not after.
- Workload balancing adapting to live needs, rather than static models, preventing overprovisioning of power and computing.
- Intelligent utility grid integration to act as a supply and demand partner for power and excess heat
- Automated regulatory compliance via rapid data processing and documentation, reducing audit preparation time and costs.
- Local optimization via distributed learning algorithms, improving resilience through increased autonomy at the edge.
- Virtual assistants enhancing human team collaboration, amplifying technician productivity, and reducing burnout.
- Autonomous infrastructure calibration adjusting dynamically, optimizing cooling, power, networking, storage, and chip-level computing in real-time

How could AI-driven predictive maintenance and workload balancing transform your data centre operations? What specific challenges in your organization could these AI applications help address?

AI-driven revolution in telco operations: automating for the future

Telco Ops are at the cusp of a radical AI-led transformation, fundamentally altering network

management from static, manual configurations to dynamic, intelligent automation. AI will augment telco capabilities, shifting from conventional manual Quality of Service (QoS) and Class of Service (CoS) settings to real-time adaptive network services.

These advancements promise significant improvements in traffic shaping, load balancing, and bandwidth management, addressing the unpredictable nature of AI workloads and using AI toolsets to do so.

AI-driven systems are set to further revolutionize NetOps by automating complex tasks like permissions management, reducing the need for human intervention. With the advent of smart algorithms, network threat detection and response will also become more proactive and efficient.

Security operations like Network Detection and Response (NDR) and Extended Detection and Response (XDR) will increasingly rely on AI to identify and mitigate threats in real-time, capitalizing on AI's ability to analyse vast datasets rapidly.

The integration of AI into network infrastructures promises a new paradigm for traffic management, where systems can dynamically adjust to varying loads with minimal latency. This will be critical as telcos prepare to meet the demands of port speeds expected to exceed 800 Gbps by 2027 according to Dell'Oro analysis.

AI's predictive capabilities will enable telcos to scale resources efficiently, ensuring seamless service delivery in an era where the volume and velocity of data traffic are reaching unprecedented levels. It will be the challenge of legacy telcos to adapt or wither in the harsh sunlight of nimble AI-first competition.

As a telecom leader, how do you envision AI reshaping your network management and service delivery? What steps can you take now to prepare your organization for this AI-driven future?

Risks require diligent governance

Integrating AI also presents challenges requiring diligent governance:

- Address ethical risks around bias, transparency, and oversight through accountability and impact analysis.
- Manage rapidly evolving cybersecurity vulnerabilities through continuous detection-response adaptation.
- Make sizeable investments in technology, tools, and training to develop in-house AI capabilities responsibly.
- Carefully integrate AI with legacy infrastructure, given interdependencies that are often opaque.
- Pace adoption reasonably to build operational maturity in phases, focusing first on constrained use cases.

For example, conversational AI assistants could be vulnerable to manipulation or exhibit unintended bias if governance does not account for ethical risks early. Leaders must mandate rigorous testing and oversight regimes tailored to AI's complexity.

Jeptha Allen Head of Digital Advisory Service at CBRE noted "AI is a new frontier full of opportunity and uncertainty for data centre leaders. My advice? Embrace AI as a collaborator, not just a tactical tool. Start small, think big. Lay foundations in governance and training first to expand responsibly. Above all, put people first - AI should enhance human potential, not replace it.

What ethical considerations and potential biases should you be aware of when implementing AI in your data centre or telecom network? How can you ensure responsible AI adoption and mitigate these risks?

Advanced applications hold promise

Sophisticated AI techniques could provide additional transformational advantages:

- Natural language processing (NLP) to extract compliance insights from dense regulations and contracts.
- Predictive telemetry analysis using statistical models tailored to specific asset configurations and failure distributions.
- Cybersecurity augmentation simulating evolving threats to continuously harden defences.

The transition towards AI-centric data centre operations also requires a significant overhaul of traditional telecommunications protocols and infrastructure. It necessitates a forward-looking approach, where the flexible, usage-based billing models, augmented NetOps and enhanced security measures become the norm.

As we pivot to this new computational era, the telco industry must not only match but also anticipate the rapid evolution of AI demands to remain at the forefront of innovation and service delivery.

Which advanced AI techniques, such as NLP or predictive telemetry analysis, could have the most significant impact on your operations? How can you prioritize these applications in your AI adoption roadmap? Do you have required staff in your organisation?

Nobody really knows how the future of AI in the data centre will play out. Whilst the path ahead remains shrouded leaders must chart a course with care and vision.

AI may yet transform rigid data centres and legacy telcos into adaptive, resilient ecosystems - if organizations and people evolve alongside it responsibly. With patient governance and strategy, operations leaders can pioneer a new era where AI elevates rather than replaces.