Issue 55 • December 2024
datacenterdynamics.com

Poles apart

The quest to lay the Arctic's first subsea cable

NTT GDC's CEO

The crypto to AI pivot

Weather prediction

YORK®

Uptime without the downsides.

CRAH Computer Room Air Handler

YVAM Air-Cooled Magnetic Bearing Centrifugal Chiller

From magnetic-bearing chillers to purpose-built air handlers, the full line of proven data center solutions from YORK® delivers performance optimized to meet the uptime requirements of today and the sustainability goals of tomorrow. After all, we're not waiting for the future: **we're engineering it.**

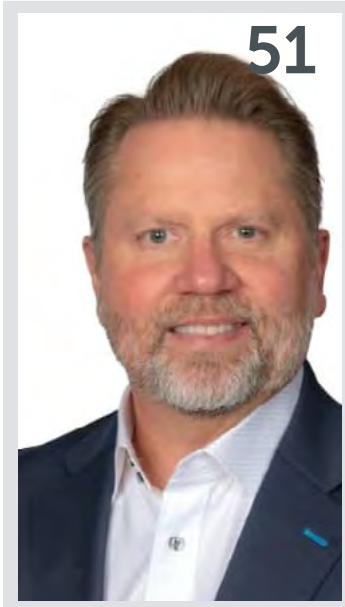
Learn more about the full line of YORK® data center solutions:

YORK.com/data-centers

Johnson Controls

Contents

December 2024



ANORD MARDIX
A Flex Company

COMPREHENSIVE ENGINEERING EXCELLENCE

www.anordmardix.com

From the Editor

When life gives you lemons, make cables

The melting Arctic is undeniably a terrible event that will cause a cascading series of much worse ones. But a reshaped global map also means a chance of reworking how the world is connected.

Head north
For the cover, we profile an ambitious plan to brave the cold climes at the ends of the Earth and lay a cable linking Europe, Asia, and North America in one fell swoop. But harsh conditions and fears of Russia could scupper their plans.

Breaking the ice, one cable at the time

When the wind blows
The ability to predict the weather has far-reaching implications beyond just knowing what to wear tomorrow. Understanding weather patterns can save lives, increase crop yields, and protect property.

But doing so is still a deeply challenging affair, one that requires increasingly powerful supercomputers.

The crypto pivot to AI
Cryptominers are tired of trying to profit off of the ups and downs of an opaque digital currency.

Instead, they're after the definitely, totally, undoubtedly more stable profits to be found in the world of AI.

NTT GDC's CEO
The boss of NTT's data center operations wants you to know they have deep pockets.

Doug Adams is done merging the company's disparate data center arms, and is willing to spend what it takes to become a data center contender.

Here comes 6G
Are you still getting used to 5G? For telco operators who are still paying off their debts, the idea of a new networks protocol may fill them with dread.

But the relentless march of technology waits for no one. Here's what we know about what's coming.

They're listening
GCHQ may be rifling through your data, but it doesn't want you to see theirs.

We chat to their former CTO about what it takes to run the IT systems of the surveillance state.

Stolen from the gods
Prometheus Hyperscale hopes to do things differently.

Despite recently hiring BP's ex-CEO, the company has a vision of greener data centers, built for the AI age.

Divine right
Google started the generative AI boom, but has seen others reap the rewards.

We talk to Google Cloud's infrastructure GM about the company's plan to become the cloud king and regain the AI crown.

Plus more
A cooling supplement, space politics, nukes & more!

750bn

Tons of ice that are melting each year due to global warming.

Sebastian Moss
Editor-in-Chief

Meet the team

Publisher & Editor-in-Chief
Sebastian Mosltoe
Senior Editor
Dan Swinho-ho

Features Editor
Matthew Goodtidings
Telecoms Editor
Paul 'The Elf' Lipscombe

CSN Editor
Charlotte Snowman
C&H Senior Reporter
Georgia But-la-la-la
E&S Senior Reporter
Zachary Sledmore
Junior Reporter
Niva Yavidad

Head of Partner Content
Claire Frosty
Partner Content Editor
Christmas Merry-man
Copywriter
Farah Johnson-December
Designer
Eleni Zevgaridou

Media Marketing
Stephen Scott
Head of Sales
Erica Baeta
Conference Director, Global
Rebecca Davison
Content & Project Manager - Live Events
Gabriella Gillett-Perez
Matthew Welch
Audrey Pascual
Channel Management Team Lead
Alex Dickins
Channel Manager
Kat Sullivan
Emma Brooks
Zoe Turner
Director of Marketing Services
Nina Bernard
CEO
Dan Loosemore

Head Office

DatasantaDynamics
32-38 Saffron Hill,
London, EC1N 8FH

Dive even deeper

Follow the story and find out more about DCD products that can further expand your knowledge. Each product is represented with a different icon and color, shown below.

Events | Intelligence | Debates | Training | Awards | CEEDA

News

Microsoft signs nuclear PPA to restart Three Mile Island

Microsoft plans to take up 100 percent of a revived Three Mile Island nuclear power plant to fuel its AI data center ambitions.

Should regulators approve the project, owner Constellation hopes to open the 837MW Pennsylvania facility in 2028. 100 percent of the power will go to Microsoft, to match the power used by its data centers in the state as well as Chicago, Virginia, and Ohio.

The PPA will last 20 years, significantly longer than Microsoft's traditional solar and wind PPAs. Constellation said that the power use was equivalent to that of 800,000 US households.

"This agreement is a major milestone in Microsoft's efforts to help decarbonize the grid in support of our commitment to become carbon negative," Bobby Hollis, VP of energy at Microsoft, said.

"Microsoft continues to collaborate with energy providers to develop carbon-free energy sources to help meet the grids' capacity and reliability needs."

While the terms of the PPA were not revealed, Constellation will invest $1.6 billion to restart the dormant reactor, which was shut down in 2019 because it was too expensive.

The Unit 1 reactor was not impacted by the infamous Three Mile Island accident, when Unit 2 suffered a partial nuclear meltdown in the worst nuclear power accident in US history back in 1979. That site is still being decommissioned by owner Energy Solutions.

Constellation purchased Unit 1 in 1999, and will pursue a license renewal that will extend plant operations to at least 2054. The newly-named Crane Clean Energy Center is expected to be online in 2028.

At the same time it seeks huge amounts of capacity to power its AI infrastructure, Microsoft is increasingly looking to offset its historical carbon footprint through the use of carbon removal.

Already the largest investor in carbon removal globally, Microsoft has invested in a number of initiatives in recent months.

The company signed a deal with Deep Sky to remove 10,000 tons of carbon through direct air-capture (DAC) - a technology that literally sucks carbon out of the air.

Microsoft has also signed a deal with Ebb Carbon to remove 1,300 tons of carbon through Ocean Alkalinity Enhancement (OAE) technology, adding alkaline substances to seawater to accelerate the ocean's natural carbon sink.

The company also signed deals with Lithos Carbon, Undo, and Eion. All three use enhanced rock weathering (ERW) to remove carbon - which traps carbon in soil and rock.

NEWS IN BRIEF

Hitachi CEO warns of transformer shortage

Hitachi Energy CEO Andreas Schierenbeck has warned that the transformer sector is "overwhelmed" and can no longer meet the demand for grid equipment required for crucial infrastructure.

Rigetti & D-Wave avoid stock exchange delisting

Quantum computing firms D-Wave and Rigetti have avoided being delisted from the NYSE and Nasdaq stock exchanges, respectively. Both faced delisting due to their low share prices.

Lumen Orbit hopes to build 5GW space data center

Space startup Lumen Orbit has raised $10 million to deploy data centers in space for training AI models.

The company aims to one day launch a 5GW station equipped with a four sq km solar array into orbit.

Microsoft announces close loop, zero-water design

Microsoft plans to deploy a zero-water evaporation design in its upcoming data centers. The closed loop system will first be piloted at its under-construction data centers in Phoenix, Arizona, and Mount Pleasant, Wisconsin, in 2026 and used in all newly-planned data centers.

OVH appoints Benjamin Revcolevschi as CEO

European cloud and data center firm OVHcloud has appointed Benjamin Revcolevschi as CEO. He replaces Michel Paulin, who has resigned. Revcolevschi joined OVH in May 2024 as deputy CEO. He previously held roles at DXC, SFR, Fujitsu, and Boston Consulting Group.

Meta accused of union-busting in Ireland

Strikes at Meta's Irish data center in Clonee have been called off, and the case is going to court.
Workers were due to strike over changes to worker shift patterns in the wake of redundancies. The company, which does not enage with trade unions, was accused of union-busting after it was suggested striking staff could be outsourced.

Amazon invests in nuclear SMRs in two states

Amazon Web Services (AWS) has signed three nuclear power deals in the US.

The company has followed recent announcements from the likes of Google and Microsoft, and signed several deals around deploying nuclear small modular reactors (SMRs) to power its data centers.

In Washington, AWS has signed an agreement with Energy Northwest, a consortium of state public utilities, that will enable the development of four SMRs.

The reactors will be constructed, owned, and operated by Energy Northwest, and are expected to generate roughly 320MW of capacity for the first phase of the project, with the option to increase to 960MW total. These projects are due to come online beginning in the early 2030s.

Through the agreement, Amazon will fund the initial feasibility phase of an SMR project, which is planned to be sited near Energy Northwest's Columbia Generating Station nuclear facility in Washington.

Under the agreement, Amazon will have the right to purchase electricity from the first project, comprising four modules and expected to generate 320MW. Energy Northwest has the option to further build out the site by adding up to eight additional modules (totaling 640 MWs) resulting in a total project generating capacity of up to 960MWs. This additional power will be available to Amazon and Northwest utilities to power homes and businesses.

Amazon is also making an investment in X-energy, a developer of SMR and fuel. X-energy's reactor design will be used in the Energy Northwest projects, with the SMRs deployed there using X-energy's Xe-100 design, a high-temperature gas-cooled reactor that can [provide 80MW each. Energy Northwest said it has engaged extensively with X-energy on plans for an Xe-100 facility since 2020.

The investment includes manufacturing capacity to develop the SMR equipment to support more than 5GW of new nuclear energy projects utilizing X-energy's technology by 2039.

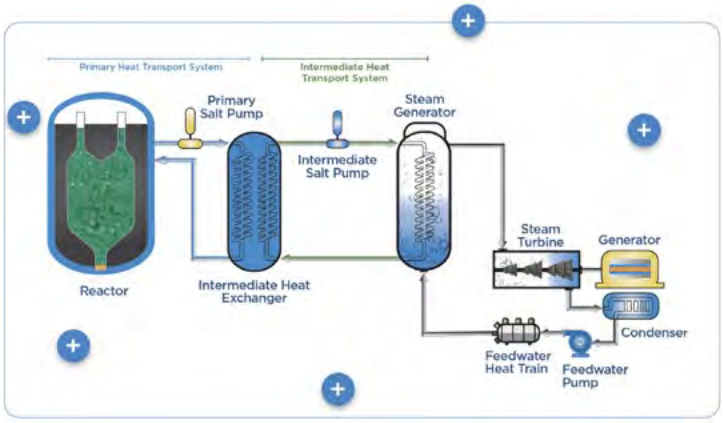
In Virginia, Amazon has signed an agreement with utility company Dominion Energy to explore the development of an SMR project near Dominion's existing North Anna nuclear power station. The company said this will bring at least 300MW of power to the Virginia region.

Though there are few details, Dominion said the Memorandum of Understanding documents the companies' efforts to "jointly explore innovative ways" to advance SMR development and financing while also mitigating potential cost and development risks for customers and capital providers.

Amazon acquired a nuclear-powered data center in Pennsylvania from Talen Energy earlier this year. The company aims to develop up to 15 buildings across 1,600 acres at the 960MW site.

Amazon and Talen have come up against resistance from other utilities in the region, who have claimed the arrangement is unfair.

Google signs nuclear SMR deal with Kairos



Google has signed a corporate agreement to purchase nuclear energy from multiple small modular reactors (SMRs) from California-based Kairos Power.

Kairos expects to deploy the first SMR by 2030, followed by further deployments through to 2035. In total, the US-only deal covers up to 500MW across six to seven reactors.

The SMR company uses a molten-salt cooling system, combined with a ceramic, pebble-type fuel, to transport heat to a steam turbine and generate power. In July, Kairos began construction on its Hermes non-powered demonstration reactor in Tennessee.

Further details about the deal, including pricing and location, were not disclosed.

Google CEO Sundar Pichai previously said the company was exploring a number of nuclear options, including SMRs, and planned to develop 1GW data centers.

Oklo, another SMR provider that has previously announced deals with Equinix and Prometheus Hyperscale (previously Wyoming Hyperscale), has signed on more data center customers.

The company in November said it has partnered with two undisclosed data center providers to deliver up to 750MW of nuclear power.

Oklo now has letters of intent totaling 2.1GW of capacity.

Meta launches nuclear power RFP, targets 1.4GW for data centers

Meta is launching a request for proposals (RFP) to identify potential nuclear energy developers to support 1.4GW of new nuclear generation capacity across the US.

The tech giant is seeking developers who can help accelerate the availability of new nuclear generators, create sufficient scale to deploy multiple units, and achieve material cost reductions.

Meta will prioritize developers who permit, design, engineer, finance, construct, and operate these power plants.

Organizations interested in participating in the RFP are invited to complete an application form by 3 January 2025. Initial RFP proposals are due on 7 February 2025.

This is Meta's second attempt to gain a foothold in the nuclear energy sector. Early last month, a potential nuclear power supply deal for an artificial intelligence data center was reportedly obstructed by the discovery of a rare bee species on the land where the project was planned.

According to a *Financial Times* report, Meta had planned to partner with an existing nuclear power operator to provide energy for a planned AI specific data center.

CEO Mark Zuckerberg told Meta during an all-hands meeting that the discovery of the bee species would have complicated the project, compounding other issues in the environmental and regulatory process.

Further details of the planned data center's location or the nuclear partner were not shared.

After a global pause on its data center build-out amid a design "rescoping" to better accommodate AI hardware and liquid cooling, Meta is again rapidly expanding its data center footprint.

December saw the company officially announce plans for a 4 million sq ft, 2GW data center campus in Louisiana.

The $10 billion development, located in Richland Parish, northeast Louisiana, will be used to train the company's Llama AI models, according to Zuckerberg.

Set on the 2,250 acres former Franklin Farm megasite between the municipalities of Rayville and Delhi, about 30 miles east of Monroe, the campus will total more than 4 million sq ft (371,610 sqm). Renderings suggest up to nine buildings are planned. Meta expects site work to begin in December, with construction to continue through 2030.

Entergy Louisiana is to support the data center with the combustion turbines with a combined capacity of 2.2GW, eight substations, and 100 miles of new transmission lines.

DigitalBridge acquires Yondr amid collapse of ISG

DigitalBridge is to acquire data center developer Yondr through one of its managed investment funds.

The deal is expected to close in early 2025. Terms were not disclosed.

Yondr will continue to operate as an independent company within DigitalBridge's portfolio, which includes Vantage, Switch, DataBank, Scala, and others.

Headquartered in London and owned by single-family investment office Cathexis, Apollo Global Management, and Mubadala, Yondr is a developer, owner, and operator of data centers. The company currently has a contracted capacity of 878MW, with 58MW currently operational. The company has projects across the US, Europe, and Asia - in Virginia, UK, Malaysia, Japan, Germany, and India.

Yondr has more than 420MW of capacity committed to hyperscalers, with "significant" additional land to support a total potential capacity of over 1GW, according to DigitalBridge.

The news arrived weeks after UK construction firm ISG, in which Cathexis as a major investor, collapsed. The company had more than £1 billion in government contracts and several data center customers.

Tract plans 2,000-acre data center campus in Austin, Texas

Data center park developer Tract has plans for a new campus outside Austin, Texas.

The company is under contract for up to around 2,000 acres along Farm to Market Road 2720 in Uhland City.

Chipmaker Micron was also previously considering the area as the home of a new chip fab site, but instead settled on a site in New York. The site is close to land that Prime is set to develop into data centers.

Founded by Cologix founder Grant van Rooyen, Tract develops master-planned data center parks on which other companies can develop facilities. Its projects often span thousands of acres, with the company filing for permission for dozens of buildings ready for other companies to build.

The company has large landholdings in Reno, Nevada; Richmond, Virginia; Phoenix, Arizona; Eagle Mountain, Utah; and Minneapolis, Minnesota.

The company is in an ongoing legal dispute with Switch in Nevada. Switch is attempting to prevent Tract developing in the Tahoe-Reno industrial site outside Reno, Storey County.

Delivering value for 60 years and beyond.

kirby
engineering & construction

DELIVERING VALUE

60

People • Safety • Quality • Delivery • Value

Since 1964

Blue Owl buys Stack owner IPI Partners

Asset manager Blue Owl Capital is set to acquire IPI Partners for approximately $1 billion.

The deal will see Blue purchase IPI from its owners, funding the deal with 80 percent of equity and 20 percent in cash.

IPI has a portfolio of 82 data centers with 2.2GW of capacity across the world.

The investment firm owns Stack Infrastructure, having formed the company in 2019 through the merging of several T5 data centers and three Infomart facilities. It also launched US regional Edge firm Radius DC in 2022.

The deal is set to bring $10.5bn in assets under Blue Owl.

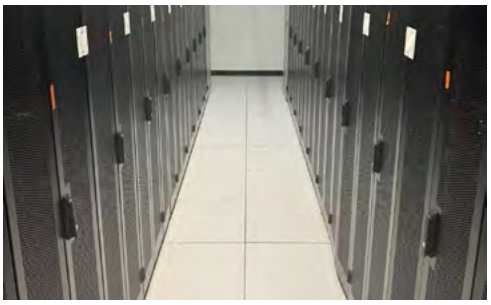
IPI's managing partner, Matt A'Hearn, will become head of Blue Owl's digital infrastructure strategy and report to co-president Marc Zahr.

Blue Owl has approximately $192 billion in assets under management. The company recently formed a $5bn joint venture with Chirisa and PowerHouse to develop data centers across the US – mostly for AI cloud firm CoreWeave.

Blue Owl has also formed a $3.4bn joint venture with Crusoe.

The money will be used to develop a new data center at Lancium's Clean Campus in Abilene, Texas, that will be leased to Oracle for use by OpenAI.

TECfusions signs 1GW deal with AI cloud firm TensorWave



US data center firm TECfusions has has signed an AI infrastructure commitment with TensorWave.

The agreement will see TensorWave lease 1GW of AI capacity across TECfusions' data center portfolio.

TECfusion will initiate the deployment in phases, with a large portion of the 1GW due to be available by early 2025.

Founded in 2023 and led by former QTS CTO Simon Tusha, TECfusions specializes in designing, building, and managing data centers for AI and HPC workloads. The company has three operational data centers in Arizona, Virginia, and Pennsylvania.

Tecfusion will utilize its on-site microgrid generation, powered primarily by natural gas, to maintain a stable supply of energy and avoid price shocks that could lead to high energy costs. It operates a 200MW onsite microgrid at its New Kensington data center in Pennsylvania and a 220MW microgrid at its Clarksville site in Virginia.

Founded in 2023, TensorWave provides companies with access to AI compute availability via AMD Instinct GPUs.

The deal is one of many made by a GPU-as-a-service companies in recent months.

CoreWeave, which is reportedly targeting an IPO next year, has continued its rapid build out. The company is converting a lab building in New Jersey into a data center, and is partnering with Cohere to develop a "multi-billion-dollar" data center in Canada. It also broke ground on a 100MW site in Muskogee, Oklahoma, in partnership with Core Scientific.

Core Scientific is also converting its cryptomine site in Denton, Texas, into an AI-focused site. The company is set to invest $6.1 billion in the project.

Sharon AI, another GPU cloud provider, is planning a 90MW campus in New Mexico powered by natural gas.

Sharon AI and New Era Helium Corp., an industrial gas business that produces helium and natural gas, announced in November that they have executed a non-binding letter of intent to form a joint venture for the design, development, and operation of a 90MW data center in the Permian Basin.

Under the terms of the 50/50 joint venture, the parties will jointly design, build, and operate an initial 90MW power plant and Tier III-quality liquid-cooled data center. Timelines for development weren't shared.

Sharon was founded this year and currently offers services from a NextDC site in Melbourne.

Dan's Data Point

A new floating data center in France is offering 200kW of capacity. Local startup Denv-R launched its debut facility on the river Loire in Nantes in October. First announced in 2022, the facility is based on Geps Technos floating platform, and is covered in solar panels.

Equinix sets up $15bn JV to build xScale data centers in the US

Equinix has set up a new joint venture to raise $15 billion to fund the growth of its xScale hyperscale data center portfolio.

The colo provider has set up the JV with GIC, Singapore's sovereign wealth fund, and the Canada Pension Plan.

Equinix says the JV will aim to purchase land for several 100MW+ data center campuses in the US, eventually adding more than 1.5GW of new capacity.

The news comes amid a shift in Equinix's strategy. During the company's most recent quarterly earnings results, CEO Adaire Fox-Martin said the company was pivoting "to build fewer and larger campuses."

She added: "Essentially, this means moving from many smaller builds with phased capacity delivery to fewer larger builds, balancing location with access to power on campuses that can service the full range of our customer's needs from SMEs to hyperscalers."

The company recently announced 400 people would be laid off by the company - around three percent of its workforce - as part of its strategy.

PRACTICAL SOLUTIONS TO AI INFRASTRUCTURE CHALLENGES

With customizable solutions and collaborative engineering, see how Legrand's approach to AI infrastructure can help your data center address:

- Rising power supply and thermal density
- Heavier, larger rack loads
- Challenges with cable management and connectivity
- Increasingly critical management and monitoring

Click on the URL below to find out more about our solutions for resilient, agile AI data centers.

www.legrand.us/ai-data-center

#Legrand**ImprovingLives**

legrand®

HMC Capital acquires Global Switch and iseek

HMC Capital has acquired Global Switch's Australian operations, as well as Aussie operator iseek, and has launched a new data center real estate investment trust.

October saw Global Switch announce a deal to sell its sell its Australian assets to investment manager HMC Capital for AU$2.12 billion (US$1.41bn).

The company operated two adjacent conjoined data center buildings in the Ultimo area of Sydney on an 11,090 sqm (119,370 sq ft) site. The 392-422 Harris Street property is known as Sydney West, while the neighboring 273 Pyrmont Street is known as Sydney East.

The seven-story West currently offers 20.5MW, with the newer East site boasting 22MW. The company recently filed to expand the site with additional floors on the existing buildings.

November saw HMC acquire iseek for AU$400 million ($264m).

Founded in 1998 and previously owned by Amber Infrastructure, iseek runs seven data centers across Brisbane, Northern Queensland, Sydney, and Adelaide with a total IT capacity of 6MW. In February, it appointed Scott Hicks as CEO and said it intended to grow its portfolio.

HMC manages more than AU$10bn on behalf of institutional, high net worth, and retail investors. In July it purchased North American digital infrastructure investor StratCap for US$28m.

The company has launched its new digital infrastructure investment vehicle, having raised AU$153m ($100m) more than anticipated in its initial public offering (IPO), and says it is in talks to buy three additional hyperscale data centers in the US.

The Digital Infrastructure Real Estate Investment Trust, which will be known as DigiCo REIT, has floated on the Australian Stock Exchange with an initial market cap of AU$4.2bn ($2.74bn). It will manage 13 data centers serving 586 customers, the filing said.

The company's pipeline is set to grow, with HMC announcing that it has agreed deals for the acquisition of three enterprise and hyperscale data centers in North America for AU$2.29bn ($1.5bn), which will value the REIT's portfolio at AU$6.1bn ($4bn).

The sale of Global Switch has been mooted for several years, since Chinese steel giant Jiangsu Shagang Group took control of the company in 2016. It was talking to potential acquirers back in January 2021 with an eye to a possible $10-11bn sale of the entire firm.

More than a dozen companies were named as potential buyers, and last year investment funds EQT, KKR, and PAG were reportedly shortlisted for a final round of bidding, but talks were said to have 'ground to a halt' in January 2023 over a gap in company valuations.

Then, in July, Global Switch reportedly rejected several bids for its Australia business because they did not meet its valuation. At the time, AFR reported that Stonepeak, Queensland Investment Corporation, and the Canada Pension Plan Investment Board were among those vying for control of the business, but they have been beaten to the punch by HMC Capital.

ECL to build 1GW hydrogen-powered data center campus

Data center startup ECL claims that it will build a 1GW AI data center on a more than 600-acre site east of Houston, Texas, powered by hydrogen.

The TerraSite-TX1 will initially have a capacity of 50MW, coming next summer, at a cost of around $450 million. AI cloud provider Lambda will be the first customer, but is not taking the whole 50MW.

ECL launched its first modular hydrogen data center back in May, with a small deployment at its site in Mountain View, California. Each module supports 1MW, and can cool up to 75kW per rack.

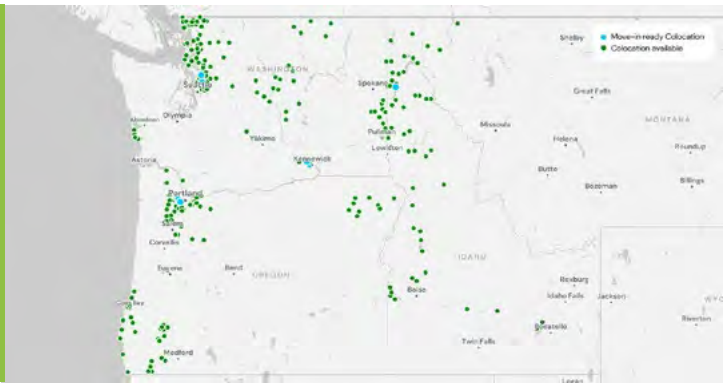
For the TerraSite, ECL said that it has three pipelines of hydrogen that will feed the facility, with an energy cost of 0.08-0.12/kWh. ECL is currently seeking funding for the entire $8 billion project, following its initial deployment. It raised $10m from Hyperwise Ventures earlier this year. In the meantime, interest in natural gas from data center developers has surged in recent months.

A multi-gigawatt off-grid campus is being proposed in Alberta, Canada, that could host dozens of natural gas-powered data centers. Former *Shark Tank* investor Kevin O'Leary is leading the project.

Meta's new 2GW campus in Louisiana is set to be supported via new natural gas generators from Entergy.

Sailfish Investors recently told *DCD* it was in the early stages of developing 13 data center campus campuses across the US, with the majority of its developments close to natural gas and nuclear plants.

Ziply launches colocation services from 200 telco central offices in Northwest US



Ziply, a fiber firm serving the northwestern US, has launched data center colocation services.

October saw the company announced the launch of its colocation services at more than 200 facilities across Washington, Oregon, Idaho, and Montana.

The sites utilize spare space in its old central office facilities that previously held large amounts of equipment for the company's old copper network.

On the company's site, it says it has around eight facilities that are currently "move-in ready" with the rest noting colocation is "available."

The firm said its "primary" facilities are equipped with a combination of UPS power systems and DC plants, back-up systems, and N+1 (or greater) redundancy.

Ziply Fiber is a local Internet service provider (ISP) dedicated to bringing fiber Internet to Washington, Oregon, Idaho, and Montana. A subsidiary of private investment company WaveDivision Capital, the company was formed by former Wave Broadband executive Steven Weed in May of 2020 when it acquired the Northwest operations of Frontier Communications.

The facilities are largely centered around Seattle, Portland, Kennewick, Boise, and Spokane, but stretch to locations across Idaho, Oregon, and Washington.

Ziply's network has its roots back in the General Telephone Company of the Northwest, Inc. which was founded in 1964. Later known as GTE Northwest, the company was acquired by Bell Atlantic and became Verizon Northwest. After being acquired by Frontier in 2010, it became Frontier Northwest.

Following the acquisition and rebranding to Ziply, the company said it would invest $500 million in improving the network and upgrading from copper to fiber, before raising another $450 million for further network expansion.

Ziply has had a busy quarter. A few weeks after the announcement, the company was involved in two acquisition deals.

November saw Canadian telco Bell Canada acquire Ziply in a deal expected to be worth as much as C$7 billion (US$5bn).

The acquisition is part of Bell's plans to expand its fiber footprint across North America and over the border into the US.

"This acquisition marks a bold milestone in Bell's history as we lean into our fiber expertise and expand our reach beyond our Canadian borders," said Mirko Bibic, president & CEO, BCE and Bell Canada.

Ziple is also set to acquire the Pacific Northwest assets of fiber-optic provider Unite Private Networks (UPN). Ziply is to take over the fiber assets, network, and customers of Cox Communications-owned UPI in Washington, Idaho, Wyoming, and Montana. Terms were not shared.

It will add more than 7,000 fiber miles across key cities, including Puyallup and Silverdale, Washington; Homedale, Idaho; Columbia Falls, Havre and Miles City, Montana; and Sheridan, Wyoming.

Four dead as helicopter crashes into 1,000ft radio tower in Houston, Texas

A helicopter crashed into a 1,000ft radio tower in Houston, Texas, killing all four people on board.

The helicopter hit the structure at around 7:54pm on October 20, with *CNN* reporting it was flying at around 600ft when it collided with the tower.

CNN reported that lighting on the tower failed days before it was hit by the helicopter.

CCTV footage appears to show the top of the tower had some lighting, however, there was no other visible lighting on the structure.

According to Houston Fire Department, an R44 helicopter hit the tower, east of the city's downtown, after it had reportedly set off from Ellington Field about 15 miles away.

The tower's lights were said to be "unserviceable" until the end of the month, as per a Federal Aviation Administration (FAA) notice to pilots in the days prior to the accident.

The tower itself is owned by SBA Communications, which according to *Wireless Estimator*, only took ownership of the tower on October 10 after acquiring the tower from Univision, which had owned the structure since it was built in 1987.

In a statement today, SBA spokesperson Lynne Hopkins said the company is cooperating with authorities on a full investigation.

StratusPower™

Future-proof your data centre with Centiel's UPS.

Green UPS

DARA TECH

SWISS MADE

Experience tomorrow's power protection technology today.

Centiel's **StratusPower™** UPS is the ultimate power protection solution for today's dynamic data centre environment.
With unmatched availability, reliability, and efficiency, StratusPower ensures seamless operations and business continuity, minimizing the risk of downtime.

Our innovative DARA design delivers unparalleled scalability, eliminates single points of failure, and provides a fault-tolerant architecture. From compact 10 kW modules to robust 62.5 kW options, the UPS meets a range of power requirements with the ability to scale up to an impressive 3.75 MW.

centiel
continuous power availability

www.centiel.com

Poles apart

Matthew Gooding
Features Editor

Could going directly through the North Pole finally deliver the Arctic Circle its first subsea Internet cable?

Lapland's reindeer herders face plenty of problems, but getting a mobile phone signal is apparently not one of them.

"I remember going across Northern Finland on an educational program with a professor who gives remote classes on reindeer herding," says Mia Bennett, an associate professor at the University of Washington and an expert on Arctic infrastructure.

"She would go out on the tundra, connect to 5G, and zoom in with her phone to show her students back in Alaska what she was doing. Connectivity in most of Northern Scandinavia is pretty good - I spoke to another reindeer herder at a workshop who was able to pull up Snapchat to chat to the other herders and see what they were up to."

Reliable 5G and social media access may be a small consolation for the herders battling against the growing industrialization of the Arctic Circle and the impact logging and deforestation have had on the land they use to graze their animals. But it does reflect the fact that virtual infrastructure in many parts of the Far North is actually in good shape.

But despite the presence of 5G in one of the Earth's most remote locations, fiber connections in the region are much rarer. Subsea Internet cables, which play a vital role in connecting the world, are few and far between in the oceans of the Arctic Circle, with climatic and geopolitical factors limiting their development.

The prize for anyone who can successfully operate a cable in the Arctic is significant, though, with the region potentially able to provide a shortcut to connect the US, Europe, and Asia, relieving pressure on other, oversubscribed, routes.

> *"We all know that Russia is up to no good when it comes to disrupting critical underwater infrastructure as part of gray zone operations"*
>
> *>>Mathieu Boulègue*

A new consortium is taking up the challenge with an ambitious plan to lay a cable, Polar Connect, right through the North Pole itself. This is easier said than done, though, with significant technological and environmental barriers to overcome if the project is to become a reality.

A brief history of cables in the Arctic

The Arctic Circle is characterized as the geographical area where the sun does not rise all day on the Northern Hemisphere's winter solstice, the shortest day of the year. It takes in areas of countries including the US, Canada, Russia, Finland, Norway, Sweden, and even a small part of Iceland. Greenland, the world's largest island and a territory of Denmark, sits almost entirely within its borders.

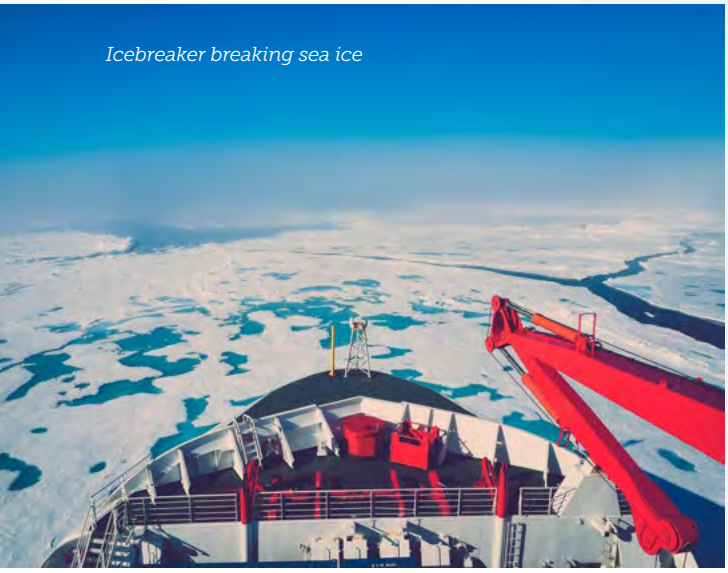
Despite covering an area of some 20 million sq km (7.7 million sq mi), most of which is occupied by ocean, the Arctic Circle is an almost entirely subsea cable-free zone. Per Submarine Cable Map, the only cables traversing the seas

of the far north currently are Russia's domestic Polar Express cable, the first section of which was completed in 2022, and the Svalbard cable system, a 1,400km (~840 mile) link that connects the island of Svalbard, the world's northernmost civilian settlement, with the Norweigan mainland.

Enterprising engineers have been trying to lay a larger network of cables around the Arctic, utilizing the Northern Sea Route, since the 1990s, says Howard Kidorf, managing partner at Pioneer Consulting, a firm that advises its clients on the planning and implementation of undersea cables. "I remember a project called Polarnet, which was originally supposed to start construction in 2002," Kidorf says. "It was a vision to connect Europe, Asia, and the US."

Polarnet was a Russian-owned company, and seemed to be making progress towards its goal of linking East and West when, in 2012, it signed a contract with US cable vendor Tyco (now Subcom) to construct the Russian Optical Trans-Arctic Submarine Cable System, or ROTACS. But work on the project was put on hold when Russia invaded Crimea in 2014, and it was eventually canceled.

Attempts were made to revive the ROTACS route in 2015 via Arctic Connect, a project led by Finland which would have seen a 13,800km (8,575 mi) cable laid connecting the Nordic nation and Japan via the Northern Sea Route. This would have been the shortest cable linking Europe with Asia, and Finnish telco Cinia joined forces with Russian telecoms firm MegaFon to deliver the project, which



Icebreaker breaking sea ice

Icebreaker at Svalbard, the world's most northern civilian settlement where Polar Connect will run

CONQUERING THE NORTHERN SEA ROUTE

Polar Connect will traverse the Northern Sea Route (NSR), the shortest connection between Europe and Asia, measuring some 5,600km (3,500 miles).

Lying entirely in Russian territory, the route was first conquered by Swedish-Finnish explorer Nils Adolf Erik Nordenskiöld in 1878 and 1879. Nordenskiöld and the crew of his ship, the SS Vega, were the first to complete the journey along the northern coasts of Europe and Asia, though the party was forced to a halt in September 1878 after the Vega became frozen in icy waters near the Bering Strait.

These icy conditions have limited the use of the NSR to date, with parts of the route only thawing out for two months a year, meaning any voyage has traditionally needed an icebreaker to accompany it. Russia uses the NSR to transport goods between six major ports found along the route, and increasingly for voyages to countries such as China and South Korea.

Climate change has led to the route becoming far more accessible, and in 2017 a specially constructed Russian oil tanker, the Christophe de Margerie became the first ship to complete the route without being accompanied by an icebreaker, making the journey in six days.

Since then, traffic on the route has increased, giving weight to the opinions of those who believe it can offer a realistic alternative to the Suez Canal for cargo ships. In 2023, 2.1 million tons of cargo were transported along the NSR in 79 voyages, according to figures from the Centre of High North Logistics.

> *"All the oceans in the world have been mapped in detail apart from the Arctic"*
>
> *>>Eric-Jan Bos*

was due to be up and running in 2021/22. However, sanctions imposed on Russia after it waged war on Ukraine, plus spiraling costs, saw Arctic Connect sink without a trace.

The Russians instead pursued Polar Express, which conjures images of a cosy Christmas tale, but is in fact a state-funded cable connecting different parts of Russia via the Arctic Ocean. The first section, linking Teriberka and Amderma, was launched in October 2022, with a second section due to come online in 2025. For obvious reasons, Polar Express is not connected to any global telecoms networks.

Elsewhere, Alaskan fiber company Quintillion proposed a northern cable from London to Tokyo via the US state, and even built part of the link, with a subsea cable in the Arctic Sea running from Alaska's Nome to Prudhoe Bay. But plans for a more ambitious network collapsed in spectacular fashion when Quintillion CEO Elizabeth Pierce was sentenced to 60 months in jail for defrauding investors out of more than $270 million. Pierce was convicted in 2019 after being found guilty of fabricating future revenue contracts to convince investors to back the scheme.

"Quintillion has not connected the big markets, so you might look on that project as a failure, but in many ways it serves as a useful proof of concept," Kidorf says. "It shows that you can lay and service cables in Arctic waters, so it's a step in the right direction."

> *"I think having more infrastructure projects in the Arctic is a net negative for the environment. It's a very sensitive ecosystem"*
>
> *>>Mia Bennett*

While it may seem like all attempts to drop cables in the Arctic Ocean are cursed, enthusiasm about the region's potential as a future fiber route remains high. There are two practical reasons for this, Kidorf says. One is that the region itself is increasingly busy with shipping traffic and industrial activity as global warming causes the polar ice cap to recede. The other is that it could open up a faster way to connect Europe with Asia and the US, avoiding the crowded and problematic routes through the Middle East.

"When the northern passages started opening, more ships started to traverse those routes," Kidorf says. "Beyond that, there was already a need for connectivity because there oil and gas facilities on the northern coast of Alaska, First Nations on Canada's coast and many extractive and military facilities on Russia's.

"There are also security concerns to consider - the US, Canada, and Russia have always had radar posts in the north, and Russia also has its submarine fleet anchored there."

As the Arctic ice thaws, things are also hotting up further south in the Red Sea, through which 90 percent of all Internet traffic between Europe and Asia flows. "All the cables available right now between Europe and Asia are expensive," Kidorf explains. "They wiggle through the Mediterranean and the Middle East, and one of the motivations for setting up new routes is to avoid the Egypt bottleneck."

Not only is the route through the Middle East crowded, but it comes with considerable peril. DCD has previously reported on the threat posed by Houthi rebels in Yemen, who have declared they will sabotage cables as part of the ongoing conflict in the African nation. Attacks on ships in the region could also cause indirect issues for the 15 cables that run adjacent to Yemeni waters, with a stray anchor capable of causing an outage that can last for weeks at a time.

Connecting the dots

The latest attempt at laying an Arctic cable is Polar Connect - a project with similar ambitions (and a confusingly similar name) to its predecessors, but one that is taking a slightly different direction.

It intends to lay a cable directly through the polar ice sheet, passing west of the geographical North Pole. The route

would run from Sweden via Norway and Svalbard through the North Pole to Japan, South Korea, and the Asia-Pacific region via the Bering Strait between Alaska and Russia, with landing points in the US, too. The 10,000km cable will pass under some 2,000km of thick ice, with a target launch date of 2030.

Polar Connect is being planned by NORDUnet, the Nordic regional research and education network serving academics in Norway, Finland, Sweden, Denmark, and Iceland.

Erik-Jan Bos, senior advisor at NORDUnet, is overseeing the project, and says it is designed to enable more efficient and resilient data sharing between Europe and the rest of the world.

It is hoped this will help bolster Europe's R&D efforts and the continent's economy. On a political level, Bos says there is buy-in at both ends of the route. "Japan and South Korea are definitely interested because we need a direct path from Europe to Asia for digital sovereignty and autonomy reasons," he says.

Bos says Polar Connect was first conceived a decade ago under the name Borealis, but was considered infeasible at the time. NORDUnet decided to revisit the idea after the other Arctic cable projects sunk, and has spent the last 18 months carrying out studies to determine whether they can make their plan a reality.

So far, the answer seems to be... maybe. A Northern EU Gateways project report from the Swedish Polar Research Secretariat (SPRS) looking into the plans suggests deployment of Polar Connect will require three ships; two heavy icebreakers (at Polar Class 2) and an ice-strengthened cable laying ship. "We need the two icebreakers so that the cable-laying ship can go about its business safely," Bos says.

Icebreakers are required because of the sheer thickness of the ice that Polar Connect intends to slice through. Some arctic icebergs reach the ocean floor, the SPRS report says: "Ice conditions in this area are among the toughest in the world," it notes. While fresh ice can be relatively easy to cut, multi-year ice, which has thawed and refrozen several times, can be harder than concrete, the report says.

The problem is, the ships needed by Polar Connect for this sizeable task do not all exist yet. When it comes to icebreakers, the Swedish Government currently owns a vessel, the Oden, that would be suitable for assisting a cable-laying ship in the Arctic. It is planning to commission a second imminently, known as the Swedish Heavy Polar Research Vessel (SHPRV), which it hopes to receive by 2028. A third icebreaker may even be needed as a backup, the Northern Gateways report says.

The oil and gas industry in the Arctic Circle would be one big beneficiary of Polar Connect

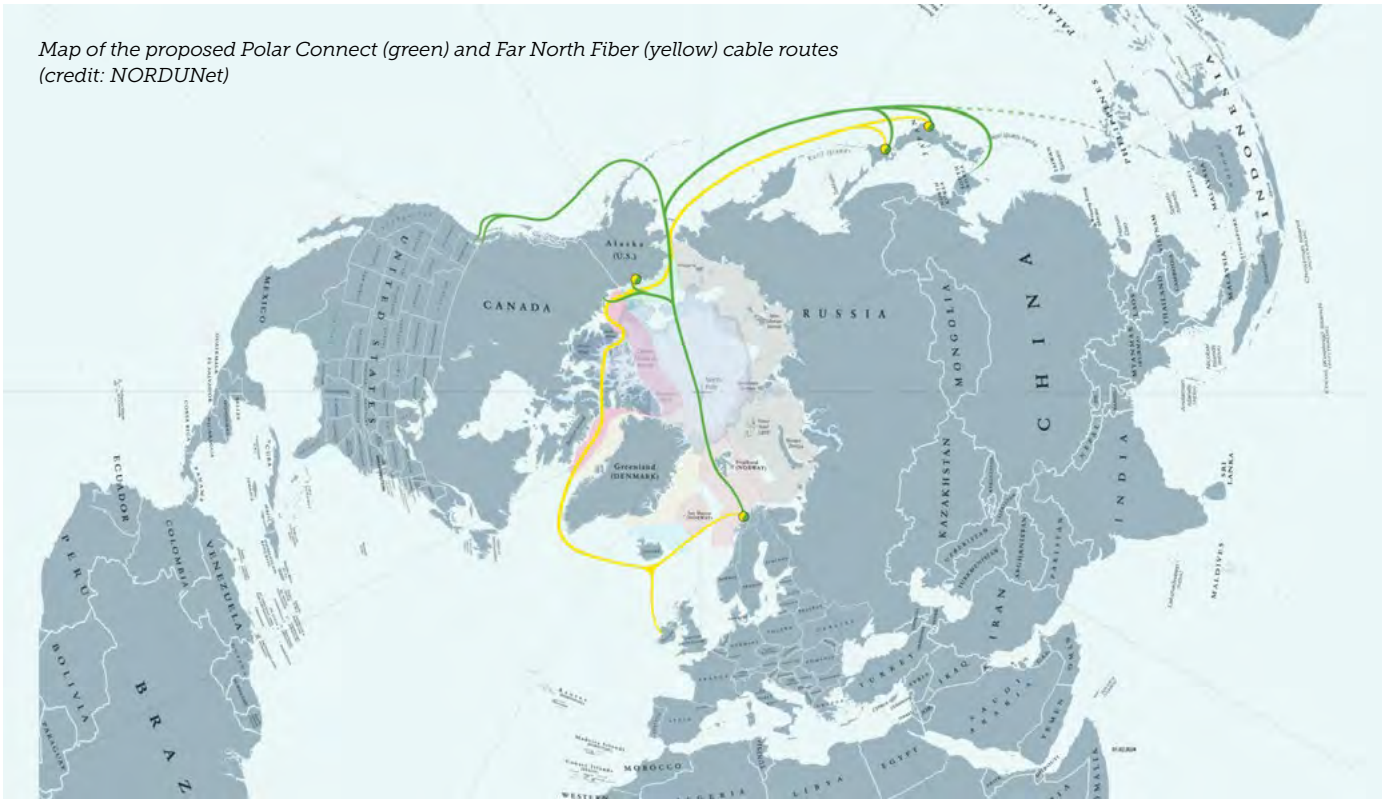
At present, there is also no cable-laying ship suitable for the extreme conditions of Arctic waters. The report suggests converting an existing icebreaker, as this could be "significantly more cost-efficient" than building a new ship. Finnish crafts Fennica and Nordica have been identified as possible candidates for conversion. The SPRS report explains that in the cable laying operation, the SHPRV would lead the way, performing reconnaissance and initial cutting of the ice. The Odin would follow, ensuring the way was clear for the cable-laying ship, which would travel at the back of the convoy.

How this complicated-sounding operation will be paid for remains to be seen, but the European Union is certainly throwing its weight behind Polar Connect, and in December 2024, announced it had awarded the project €4 million ($4.22m) to carry out mapping of the seabed in the Arctic Ocean.

This is something, Bos says, that has never been done before on the required scale. "All the oceans in the world have been mapped in detail apart from the Arctic," he says. "It remains largely unknown because it's extremely difficult to do. We are planning to go there on expeditions over the next two years and collect bathymetry information."

Bathymetry information is that relating to the depth of an ocean, and thanks to the EU funding, the Polar Connect team will have access to the Oden to carry out its mission. Ieva Muraškienė, strategy, and policy officer at NORDUnet, says the bathymetry detail will be key to finalizing the route for Polar Connect.

"It's extremely important to get this data to have an understanding of where the safe path for the cable is," Muraškienė says. "We already have some information about ridges and seismic activity, and there's a whole continent underwater, which is relatively safe because it's flat. But there might also be things like



Map of the proposed Polar Connect (green) and Far North Fiber (yellow) cable routes (credit: NORDUNet)

mountain slopes where you don't really want to be putting a cable."

Russian roulette

While the Polar Connect team grapples with the physical challenges of laying cables through one of the most remote and unknown parts of the world, it will also be mindful of the geopolitical situation in the region.

Russia's status as an international pariah has put the kybosh on many pan-Arctic schemes, the University of Washington's Bennett says. "There's been a lot of excitement around infrastructure projects in the Arctic, driven by climate change making the region more accessible," she says. "In the early 2010s, things like roads and pipelines came to fruition, and Russia made significant investments."

But, she says, a lot of projects fell through, partly because of cost overruns making the Arctic unviable, but also because "Russia becoming separated made things more challenging and volatile."

Mathieu Boulègue concurs. A researcher on Eurasian security and defense issues, and a visiting scholar at New York University, Boulègue recently authored a paper entitled Arctic seabed warfare against data cables: Risks and impact for US critical undersea infrastructure, and says the current situation in the region is something of a return to normality.

> *"The potential benefits are there, it just remains to be seen if the benefits will outweigh the costs."*
>
> *>> Howard Kidorf*

"In the 1990s, Gorbachev floated the Murmansk Initiative, which aimed to promote Arctic exceptionalism and cooperation," Boulègue says. "After that, we had what now looks like a 20-year 'lull' of low tension, but we are back in an era of operating in a confrontational geopolitical environment."

No longer a reliable partner for other countries in the Arctic region, Russia now squats as a hostile state next to Finland, and Polar Connect will pass close to its waters. This is a potential problem because the Kremlin - and its allies in China - have shown they are not averse to interfering with subsea infrastructure as a way of disrupting their enemies in the West.

In November 2024 two Internet cables in the Baltic Sea were severed by a Chinese vessel, which investigators believe set out to deliberately cause damage. The Yi Peng 3, which had a Chinese captain and at least one Russian crew member, sailed over both the C-Lion1 Helsinki-Rostock cable and the BCS East-West link cable between Lithuania and Sweden around the time they were cut. Russia has denied any involvement, and such incidents can happen accidentally, but other damage to cables in the region has aroused suspicion, notably in 2022 when the Svalbard cable was hit and had to be repaired. Investigators said at the time the cause was likely "human-made but unintentional."

"We all know that Russia is up to no good when it comes to disrupting critical underwater infrastructure as part of gray

Regular cable ships like this Blue Marine Telecom craft will not be suitable for laying Polar Connect (Credit: Blue Marine Telecom)

THE FAR NORTH OPTION

Polar Connect is not the only live cable project with designs to link Europe and Asia.

Far North Fiber wants to build a 14,000km cable connecting Nordics to Japan, via Greenland, Canada, and Alaska. Its route is similar to Polar Connect, other than that it will loop around the south of Greenland to connect the US and Europe, rather than laying cable directly through the North Pole.

The project is a joint venture between Alaska-based Far North Digital and Finnish telco Cinia. Ethan Berkowitz of Far North Digital says: "We're Alaska-based, and we recognize how fragile the connectivity that we have is to the rest of the world.

"We also know that the Arctic is an underutilized piece of real estate, and that it can be used to offer direct connections that are safer, faster, and more open."

While the main purpose of Far North Fiber is to provide a high-speed trunk route, the cable could also have spurs linking communities in the Arctic Circle, which struggle for connectivity.

On top of this, it is hoped the cable will boost Internet access in Alaska and make it a more viable proposition for data center operators. Governor Mike Dunleavy is keen to attract the hyperscalers, and believes the cool climate and abundant natural resources in Alaska make it an ideal location for large-scale campuses. He has apparently already held talks with Amazon, Meta, Microsoft, and Switch about investing in the state.

"You can't have data centers without cables," Berkowitz says. "We want to make sure that cables are available so that you can do data center development, not just in Alaska, but across the North."

Berkowitz adds that some preliminary funding is in place for Far North Fiber. "I would describe us as being shovel-ready in process," he says. "We have a contract with Alcatel Submarine Networks to design and install the system, and now we're seeking the additional funding that's needed to pay for a marine survey and build and install the cable.

"We want to begin the marine survey in 2025, and we are optimistic we will find the support we need to move this project forward, because it's going to be globally significant." ■

zone operations," Boulègue says. "They consider themselves to be in a form of conflict with the West. It's really low-intensity, simmering, warfare, and this kind of sub-threshold operation is what Russia knows best."

Boulègue argues that Russia is likely to continue targeting subsea cables because the pushback from Nato nations has, so far, been limited. "They disrespect us so much that they think it's ok to wage subsea warfare because the deterrents have been minimal," he says. "This is an area where they perceive they are not vulnerable."

What does this mean for Polar Connect? "There's always going to be a risk of attack," Boulègue says. "They won't be able to protect the entire length of the cable, so they will need to look at choke points around landing stations, or areas where the topography of the seabed means the cable is more vulnerable to being attacked."

More generally, he believes a coordinated response is required from the Nato nations, both in terms of using more monitoring technology to detect and respond to incidents quickly, and in calling out vessels that are spotted behaving unusually near cable routes. "If there is a 200,000-ton Chinese freighter doing weird movements near a cable route, then we need to say so publicly," he says.

"We need to be bold, and I think we are learning from the cybersecurity world about this. Twenty years ago, we were still on training wheels when it came to the legal and technical attribution of cyberattacks, but now governments and industry leaders are not afraid to say if they think an attack is state-sponsored. We need to hear similar rebukes about cable attacks."

Bringing Polar Connect to life

There's plenty of work to do before Polar Connect reaches the stage where it needs to worry about being sabotaged by a massive Russian anchor.

Pioneer Consulting's Kidorf says laying and maintaining cables in Arctic waters is likely to be fraught with difficulties, the biggest of which - in a very literal sense - is ice scour, where moving icebergs scrape the bottom of the ocean floor.

"Icebergs reach a long way down, and

if there's a cable there, they don't really care," he says. "They'll just rip it up."

Sea ice, the kind that occurs on the surface of the ocean as the water freezes, is also likely to be problematic for laying and maintaining cables, Kidorf says. He explains: "You can probably dodge a bullet during installation by building it during the half of the year when there isn't any ice, but that's going to be costly for a big project. Navigating weather windows is something the subsea cable industry is used to - you don't want to be in the Northern Atlantic in the middle of winter, either."

There's less flexibility when it comes to maintenance. Kidorf says: "If your cable gets cut in the wrong time of year, what are you going to do? Are you just going to have an icebreaker on standby? That's still an open issue."

Polar Connect believes that the laying of the cable could take place over 80 days during the Arctic summer, in August and September, when the ice has receded. That said, the SPRS report notes that this may prove unrealistic when the time comes. "One must bear in mind that nature rules in this part of the world," it says. "What seems to be possible one year can be totally impossible the next."

The laying process is expected to cost some $142.5-237.5 million, but the overall bill for the project is likely to be far higher, Kidorf says. "The market for these cables is totally unproven," he says. "All the proposals so far have been priced in the region of $600 million - $1.2 billion, and that's a lot of money to put down for an unproven market. Of course, this is often the case with virgin markets, but you need investors with a healthy appetite for risk."

Polar Connect will also have to lay its cables without doing further damage to an environment already under severe stress. The Arctic faces multiple threats from climate change, with global warming shrinking the size of the summer ice cap by 13 percent each decade, according to figures from the World Wildlife Fund. Sea levels are rising, permafrost is thawing, and extreme and unexpected weather events like wildfires are becoming more commonplace.

Against this backdrop, is building more infrastructure in the Arctic a wise idea? Kidorf believes the cables can be laid without doing too much additional damage. "The cables themselves don't pose much risk because they're only the size of your thumb," he says.

"The bigger risk is what it does to further open up the polar region. The more we facilitate the oil and gas industry and the extractive industries in general - there's also a lot of mining going on in northern Russia - the more land we can lay waste to."

The wider impact of the project is also a cause for concern for Bennett, who says that communities in the Arctic are still "paying the price" for botched infrastructure projects of the past. She says: "I think having more infrastructure projects in the Arctic is a net negative for the environment. It's a very sensitive ecosystem, and the Arctic amplification effect feedback loops causes climatic factors to impact elsewhere."

The Arctic amplification effect refers to the fact that the polar region is warming two or three times faster than other areas. And because the Arctic plays such a vital role cooling the rest of the planet, a feedback loop is created which worsens climate change elsewhere.

Polar Connect's Bos believes the project can be beneficial to the environment thanks to sensors that will be placed on the cable. The plan is to install two types of sensor, one to monitor the cable for breaks and other damage, and another that can assist scientists. "Not only is bathymetry information about the Arctic Ocean missing, but scientists lack many other types of data that could help them study climate change," he says. "The cable can help provide that."

Attaching sensors to subsea cables is something that is often talked about within the industry but has yet to become a reality. Bos says products are already on the market that can be incorporated into cables, citing a range of sensors from Alcatel Submarine Networks that can deliver "a constant stream of data" on things like water temperature and salinity. "Because we have six years until deployment, we're talking to sensor manufacturers about what else might be possible," he says. "We're exploring a lot of different options."

Time will tell whether Polar Connect succeeds, and whether the reality of the cable route matches the current vision of Bos and his team. But Pioneer Consulting's Kidorf says, regardless of the fate of the project, it is only a matter of time before someone makes a cable connection through the Arctic Circle. "I think it will inevitably happen at some point," he says. "The potential benefits are there; it just remains to be seen if the benefits will outweigh the costs."

CONNECTING THE SOUTH POLE

Like its Northern Hemisphere counterpart, the area around the South Pole is currently devoid of subsea cables.

When it comes to commercial routes, Antarctica is likely to remain a cable-free zone, with no obvious economic benefit to developing long-distance connections through the region.

However, it may be about to get its first subsea cable to support scientific research. The US government's National Science Foundation (NSF) is investigating the possibility of building a cable connecting the largest US Antarctic Program research facility, McMurdo Station, with either New Zealand or Australia.

Though the main scope of the installation "is to provide advanced high-speed, low delay telecommunications" to McMurdo Station, the cable "will contain additional point sensors and/or distributed sensing infrastructure, enabling for the first time myriad investigations across a broad range of scientific disciplines," the NSF says.

Having already carried out feasibility studies to determine that building a cable to the Antarctic is possible, in December 2024 the NSF published a call for information that can help move the project forward. No timescales have been published as to when it might come into service.

DCD Awards>2024 Winners

Following months of deliberations with an independent panel of expert judges, DCD Awards is proud to celebrate the industry's best data center projects and most talented people.

With thanks to our headline sponsor Mercury Engineering.

Edge Data Center Project of the Year

Winner: **Scale Computing**

Royal Farms wanted an Edge solution that optimized operations and enhanced customer experience. Scale Computing platform accomplished this with a scalable platform.

Category Sponsor VIAVI Solutions

Asia Pacific Data Center Project of the Year

Sponsor MITSUBISHI ELECTRIC

Winner: **Firmus Metal International**

Firmus' Sustainable Metal Cloud's HyperCube deployment at STT GDC in Singapore, introduced innovative immersion cooling technology, reducing energy consumption by 50 percent and achieving a PUE of 1.02.

North American Data Center Project of the Year

Winner: **ECL MV1 Off-Grid Data Center**

ECL's MV1 facility in Mountain View, California, is the world's first off-grid, hydrogen-powered data center. Designed for AI with up to 75kW per rack and a PUE of 1.1, it sets a new standard for sustainable, high-density infrastructure.

Sponsor ZincFive

Latin America Data Center Project of the Year

Sponsor Hyphen

Winner: **Scala Data Centers**

Scala's 24MW IT capacity data center combines innovation, energy-efficiency and AI-ready design to deliver resilience and sustainability for next-generation data center projects across Latin America.

Middle East & Africa Data Center Project of the Year

Sponsor: STULZ

Winner:

iXAfrica Data Centres, in collaboration with Schneider Electric

iX Africa Data Centre has transformed data center services in Kenya and East Africa, creating jobs and driving regional growth. With over 35 Kenyan staff employed directly and a strong focus on local suppliers, the project delivers a lasting impact on both the industry and the community.

European Data Center Project of the Year

Sponsor: GLOBAL COMMISSIONING

Winner: **Green Mountain**

Green Mountain's OSL-Hamar data center in Norway is a 90MW facility and Europe's largest colocation data center. It was built at record speed for social media giant TikTok and features 100 percent renewable energy from hydropower.

Mission Critical Tech Innovation Award

Sponsor: CBRE

Winner: **Centersquare**

Centersquare's Liquid Cooling Showcase is a modular facility that redefines data center cooling for the AI age.

Energy Impact Award

Sponsor: RED

Winner: **T.Loop**

T.Loop's vision is to place data centers in every commercial property—office blocks, industrial sites, or shopping centers. By tapping into unused grid capacity, they are driving a true circular economy and uncovering power potential in key city markets.

Environmental Impact Award

Sponsor: EQUANS DATA CENTERS

Winner: **Intel**

Intel's D2 P1 data center achieved LEED platinum data center status without the need for major capital improvements, demonstrating that high levels of environmental performance, energy savings, and low embodied carbon can be achieved even for a complex facility.

Community Impact Award

Sponsor: ELAND CABLES

Winner: **Applied Digital**

Applied Digital has transformed its Ellendale, North Dakota, data center campus by combining artificial intelligence-ready data centers with community projects, from housing initiatives to partnerships with local organizations, proving IT innovation thrives with community enrichment.

Editor's Choice Award

Winner: **K2 Strategic**

The DCD editorial team were impressed with how K2 Strategic deployed two 40MW data centers within just nine months. A prefabricated modular approach was combined with an innovative non-water, air-cooled fan-wall, delivered high flexibility for AI.

Young Mission Critical Engineer of the Year

Sponsor: kirby engineering & construction

Winner: **Laura Munoz Becerra, Connectivity Engineer, Meta**

With a degree in mechanical engineering, Laura joined the highly competitive Meta DEC Gold program, where she discovered her passion and talent for telecommunications. In just a few years, Laura has grown from a talented engineering graduate into a key leader.

Data Center Construction Team of the Year

Winner: **Scala Data Centers**

Scala's remarkable construction team, responsible for building the largest data center campus in Latin America, managed six major projects simultaneously, including a new substation, while maintaining high safety standards.

Sponsor: CLEVELAND CABLE COMPANY

Data Center Operations Team of the Year

Winner: **Elea Data Centers**

Elea Data Centers demonstrated operational excellence when faced with the flooding in Porto Alegre in May 2024, ensuring service continuity and responding to the urgent demands of clients & community.

Sponsor: MITSUBISHI ELECTRIC

Data Center Woman of the Year

Winner: **Dame Dawn Childs, CEO, Pure Data Centres Group**

This award celebrates a visionary who has not only pushed the boundaries of technology but also paved the way for inclusivity and innovation.

This year's winner, Dame Dawn Childs is the CEO of Pure Data Centres.

She is known for her dedication to promoting STEM education for young people across the UK and addressing the gender imbalance within the world of engineering. She is also a champion of driving collaboration across stakeholders within the data centre industry to meet the growing challenges of rapidly increasing data consumption, the subsequent demand for delivering more data centers and the need to do this sustainably.

Sponsor

ANORD MARDIX
A Flex Company

Outstanding Contribution to the Data Center Industry

Winner: **Andrew Jay, CBRE**

Every industry has its champions, thought leaders and pioneers who will be revered for their achievements for years to come. The previous winners of this Award are all distinguished by their extensive service to the data center industry, and have in some way changed the way in which the industry looks at key challenges or key opportunities.

This year's winner, Andrew Jay, is an Executive Director within the Global Corporate Services division based in London. He is head of the EMEA Data Centres team which is the world's largest data center real estate advisory group.

He is recognized as one of Europe's leading advisors to the data center industry.

Sponsor

MERCURY

Data Center Workforce Initiative of the Year

Winner: **Microsoft**

The Microsoft Datacenter Academy (DCA) initiative, launched in 2019, focuses on workforce development by offering training, certification, and job placement in tech. Through partnerships with educational institutions and local organizations, the program has expanded its reach, supporting 5,482 students, by creating pathways to employment and significantly enhancing community engagement and economic development.

Sponsor

DCD Academy


Headline Sponsor
MERCURY

Sponsored by

Life Is On | Schneider Electric

The Cooling Supplement

Cooling for generative AI

INSIDE

Liquid Cooling Made Easy

Liquid cooling has become essential for high-performance accelerated computing. Air cooling was practical when chip densities were lower. Yet, these densities have skyrocketed, placing more pressure on traditional air cooling until it become increasingly unmanageable. So, new approaches for heat removal are needed to avoid the risk of hot spots that lead to equipment failure and downtime.

Direct liquid cooling is not a product – it is an architecture supported by critical systems including coolant distribution units. Direct liquid cooling however is a bit more than the name implies, as it includes the air cooling and heat rejection units (e.g., chillers) you are already familiar with and prevalent within data centers today.

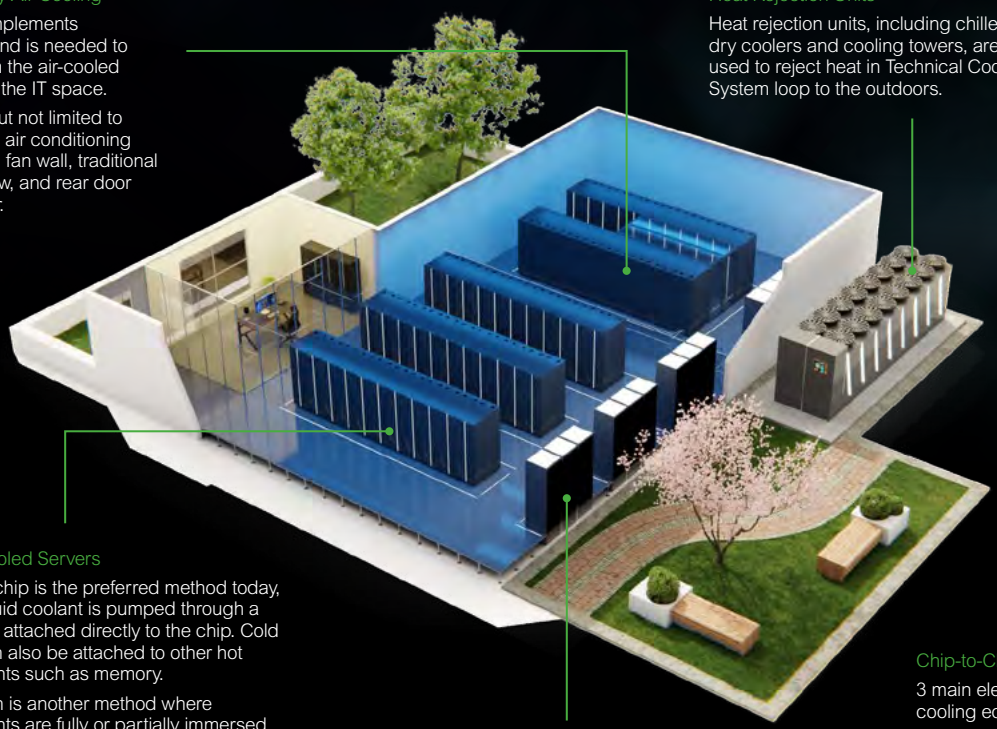
Complementary Air Cooling

Air cooling complements liquid cooling and is needed to reject heat from the air-cooled components in the IT space.

This includes but not limited to computer room air conditioning and air handler, fan wall, traditional perimeter, InRow, and rear door heat exchanger.

Heat Rejection Units

Heat rejection units, including chillers, dry coolers and cooling towers, are used to reject heat in Technical Cooling System loop to the outdoors.



Liquid Cooled Servers

Direct-to-chip is the preferred method today, where liquid coolant is pumped through a cold plate attached directly to the chip. Cold plates can also be attached to other hot components such as memory.

Immersion is another method where components are fully or partially immersed in liquid coolant.

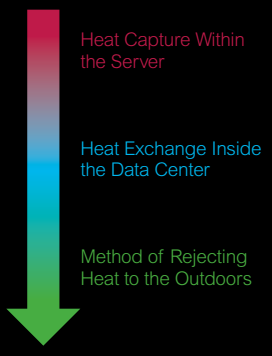
Coolant Distribution Unit (CDU)

CDU isolates the Technical Cooling System loop from the rest of the cooling system and controls temperature, flow, pressure, fluid treatment, and heat exchange.

CDUs vary in type of heat exchange, capacity, and form factor (rack- vs. floor-mounted).

Chip-to-Chiller

3 main elements in a liquid cooling ecosystem



Navigate liquid cooling

Life Is On Schneider Electric

Contents

Liquid dreams

As 2024 draws to a close, those working in the data center industry may reflect on it as the year that liquid cooling went fully mainstream.

For so long talked about as a coming technology for the sector, liquid cooling has become embedded in the data center design process to the extent that it is now the default option for many developers, particularly those looking to cater for AI workloads. At the same time, the number of flavors of liquid cooling on offer is also on the rise, with more direct-to-chip and immersion options coming to the market than ever before.

In this cooling supplement, we take a look at various implications of having more and more liquid flowing through data centers.

Vlad-Gabriel Anghel brings us up-to-date with developments in the liquid cooling market in the generative AI era, looking at how systems are developing, new releases from some of the leading vendors, and the factors that data center companies should consider when plumping for a liquid cooling system. With the upfront investment required for such systems being significant, the importance of making the appropriate choice for a particular data center, and the workloads it will be running, has never been greater.

Elsewhere, Dan Swinhoe looks at how data center operators serving hyperscalers, GPU cloud providers, and leading enterprises are adapting their approach to resilience, redundancy, and uptime to ensure they are appropriate for liquid cooling. Though for many companies a liquid cooling set-up will see a reduction in the number of fans and other mechanical parts that need to be maintained compared to a traditional air cooling system, factors such as thermal inertia come to the fore. What's certain is that the stakes are high - even a brief period of liquid cooling system downtime can cause big damage to some very high-value components.

Part of resiliency planning is considering what to do if a water leak occurs at a data center. These can be damaging events, and in the past have caused fires and environmental damage. Because of this, operators are putting more emphasis on water leak detection systems, and vendors that spoke to Matthew Gooding for his article said they have seen the average spend on such systems increase significantly in recent years. The technology at the heart of water leak detection is tried and tested, but it is important to plan implementation carefully to avoid potentially costly gaps in coverage.

An often-forgotten factor in the cooling equation is humidity. Getting the balance right in this area can often be a challenge - too much humidity can lead to the erosion and damage of key metal components across the white space. But data centers do need some moisture to avoid the build-up of the sort of electrostatic discharge that can cause sparks to fly. Niva Yadev investigates just how much humidity is too much.

How liquid cooling impacts uptime thinking

Dan Swinhoe
Senior Editor

A dive into the murky waters of resilience and redundancy of liquid cooling

With the launch of Nvidia's latest generation of GPUs, liquid cooling has moved from a nice-to-have to a necessity for many firms. Chips that are simply too potent to cool through air-based methods are driving the adoption of direct-to-chip and rear-door heat exchangers.

As data center operators serving hyperscalers, GPU cloud providers, and leading enterprises look to adapt their offerings to cater to increasingly dense workloads, changes will need to be made around how companies approach resilience, redundancy, and uptime.

Is liquid more or less resilient than air?

Data center operators across the board are rolling out liquid-cooled enabled data halls, and many hyperscalers are developing bespoke liquid-cooled systems in their own data centers.

The need to adopt liquid for today's high-end hardware is simply one of physics. Air as a cooling medium has a physical limit and, after a certain point, no amount of cold air can be passed over chips quickly enough to keep the hardware within the required temperature thresholds.

"Data center operators have done everything they can with air-based cooling systems," says Ben Coughlin, chairman and co-founder of liquid-cooled data center operator Colovore. "It reaches a tipping point where the physics are too impossible to ignore, and you just need to have liquid as the cooling medium."

While moving to liquid might be a necessity, it comes with new difficulties to ensure that the same 'five-nines' level of uptime customers have come to expect.

Vlad-Gabriel Anghel, director of solutions engineering at DCD's training unit, DCD>Academy, notes that while there will be fewer fans and moving parts to maintain, operators are now relying on a continuously moving liquid to maintain the environmental conditions that the hardware needs in order to function correctly.

"Robustness will not come from one system being used over another but through it being engineered into the wider system," he says.

Colovore's Coughlin adds: "There isn't a playbook of what liquid cooling is. There are lots of different ways to deliver it; rear door heat exchangers, direct liquid connections into server chassis, or water delivered by a cooling distribution unit (CDU) either in the cabinet or a floor-mounted CDU."

How liquid changes resilience and redundancy

NTT GDC has more than 200MW of liquid-cooled capacity (mostly direct-to-chip) rolled out at half a dozen data centers across the US – one of the largest such footprints globally. Bruno Berti, SVP of global product management at NTT Global Data Centers, tells DCD that, to date, its deployments have not encountered any major failures.

"We have a lot less time to accommodate any type of stoppage of that liquid, so we need to ensure that that liquid continues to flow to continually extract that heat," he says. "Resiliency in liquid distribution systems is more important than redundancy in an air-based system. So it's definitely something that's been in the forefront of our design and engineering teams' minds."

DCD>Academy's Anghel tells us that thermal inertia is considerably smaller in liquid cooling systems compared to air cooling which means even a very brief disruption in water flow can lead to massive temperature spikes, which can cause hardware to throttle or potentially lead to permanent damage.

"The speed at which heat builds up in liquid cooling system makes cooling failures much more critical than with air cooling," he says. "Without adequate redundancy and proactive monitoring, the risk of an outage is significantly greater in this new era of high-performance computing."

"While the core goals of reliability and uptime remain the same, liquid cooling necessitates a more integrated approach that combines physical redundancy, advanced monitoring, and automation," adds Anghel. "While the principles of 2N and N+1 redundancy still apply, they must be adapted to address the unique vulnerabilities of liquid cooling, such as rapid thermal buildup, coolant leaks, and flow disruptions."

To do this, he says, operators need to prioritize redundant cooling loops, pumps, and heat exchangers, ensuring these systems are fully independent to avoid cascading failures. Real-time monitoring of flow rates, pressure, and temperatures, coupled with automated failover mechanisms, is critical to maintaining uptime in the event of a fault.

> *"From a resiliency perspective, I think liquid is more complicated, and has more failure scenarios."*
>
> *>> Bruno Berti*
> *NTT GDC*

Designing liquid systems for uptime

California-based Colovore is a long-standing advocate of liquid cooling. The company launched its first all-liquid facility in Santa Clara more than a decade ago. The company is about to launch a second data center next door and is planning more in Reno and Chicago.

On its site, the company claims the original facility has maintained 100 percent uptime throughout its operation, offering densities up to 250kW through rear door heat exchangers and direct-to-chip cooling.

"I don't think there's a different way of thinking about it," says Colovore's Coughlin. "You still have the base fundamental design considerations of delivering power and cooling to a cabinet."

"Traditional air-based systems have to be able to be concurrently maintained, and if you've got an issue with a certain part of the floor you want to isolate that," he adds. "You still have to think those issues through in a liquid-cooled system too; like the water distribution and having redundant pathways and being able to close off certain valves and taps in certain situations."

While the fundamental goals might be the same, there are still practical aspects to consider. One of the major considerations around the resiliency of water loops is whether to build a system with high-pressure water flow rates or low-pressure water flow rates, and very large piping or narrow pipes. There are factors around materials to consider; should a rubber hose or a silver metal braided hose be used?

Coughlin says the company has always erred on the side of using" very industrial very, very thick components, but there are a lot of providers who are using plastic hosing in rubber tubes to distribute water."

"Generally, we design our system with larger pipes and lower flow rates, because we believe that's a safer approach," says Colovore's Coughlin. "High-pressure water flows, to the extent you ever had a leak or a break, can escalate into bigger issues."

"We've always erred on the side of using very industrial very, very thick components, but there are a lot of providers who are using plastic hosing in rubber tubes to distribute water," he adds. "Broadly speaking, it does seem like the industry is trending towards warmer inlet temperatures of water and lower flow rates, which lines up well with how we've designed our system."

Berti says NTT has a flexible design that can mix and match air and liquid cooling, allowing it to deploy CDUs where fans would otherwise be. The company's experience with primary water loops for the air-cooled chillers meant NTT was comfortable with the idea of a secondary loop for the liquid-based systems.

"From a resiliency perspective, I think liquid is more complicated, and has more failure scenarios," he says. "You've got a lot of piping inside the data center; a lot of joints, a lot of moving parts, a lot of potential for leaks. There are also a lot of valves that people are plugging in and out.

"We've got multiple CDUs in an N+1 type configuration. If any pump goes down, there are always going to be more CDUs to handle any failure and fault scenario, and we have a dual loop system that allows us to isolate that loop at any point."

Isolation is important. Berti notes some customers and other colo providers might be happy just running a water loop to the racks; but this means if there's a leak, you have to shut down that whole thing.

"If there's a leak somewhere, we have multiple places where we can intercept that loop and then basically put it on the secondary loop," Berti says. "And because we can isolate that loop. We'll be able to add CDUs later on in the process as well."

"That's important in the resiliency piece; how do you do maintenance on these inherently less resilient pieces of equipment."

Liquid SLAs: do the guarantees need to change?

While standards are emerging, the industry is still settling on industry standards and best practices for many aspects of liquid cooling.

"There is still quite a bit of work to be done to determine what the industry standards would be," says Coughlin. "There isn't that playbook of 'here's how I design a liquid-cooled data center.' We're working through this in real-time."

"It is ultimately driven by those server platforms and what their requirements are. Until we get some consistency from the hardware side, there's always going to be some degree of customization involved, because we don't have uniformity in those standards."

Beyond standards, questions over what guarantees operators need to offer around liquid systems are still yet to be fully answered.

NTT's Berti says the company is updating and redefining a number of SLAs for liquid-cooled deployments and has updated some contracts with clients.

"We're coming up with a whole new set of SLAs," he explains. "Flow rate and differential pressure between the supply and return of the CDUs are SLAs that are being defined. The water temperatures, both primary and secondary, are being updated. Secondary liquid temperature, secondary flow rate, secondary return, and supply differential pressure. Those are the ones that we're working on right now with our clients, redefining, and putting into standard SLAs."

"We never used to have to SLA the primary water loop temperatures because it was the air that dictated what temperature we could run the liquid at. But now it matters because of the heat extraction requirements that are on the primary side."

SLAs for liquid-cooled systems "must go beyond the typical guarantees of uptime and equipment reliability seen with air-cooled systems," says DCD>Academy's Anghel. These should include, he says, requirements for flow rates, temperature ranges, redundancy, monitoring, and maintenance specific to liquid cooling.

> *"Resiliency in liquid distribution systems is more important than redundancy in an air-based system. It's definitely something that's been in the forefront of our minds."*
>
> *>> Bruno Berti, NTT GDC*

For retail colocation provider Colovore, however, things may be different.

"The critical variables are ultimately what's required by the hardware platforms," Coughlin says. "To some degree, it's already built into the equation. If you can't meet those [hardware] specs, there's no point having an SLA [because] you can't do it."

He acknowledges that may change in the future. But for now, he says, the company's customer base "doesn't ultimately care too much" about the medium.

"We have to deliver power and cooling to our customer environments. Those end metrics and requirements from an SLA perspective, haven't changed. The cooling medium might be slightly different, but, from the customer's perspective, that doesn't really matter. "

There are also questions about who owns that water infrastructure, especially the CDUs. Some customers may want that infrastructure managed by the data center operator (or perhaps the CDU provider), but some of that may be located in areas where facilities staff might traditionally not be authorized to enter.

Does it matter?

Most liquid cooling technology has matured in the realm of high-performance computing. Jacqueline Davis, research analyst at the Uptime Institute, suggests

that because of this, the liquid cooling landscape is still speaking that language of uptime and mean time between failures, rather than redundancy.

"At present, liquid cooling is going where needed to solve thermal challenges where operators have few other choices. It's not being chosen primarily for reasons of efficiency," says Davis. "They can choose very durable pumps, fluid connections, etc, to try and get the best uptime and the best average mean time between failures. But it's harder to build things redundant."

That so much GPU hardware and therefore liquid cooling is still currently focused on training - not unlike the batch computing mindset of HPC deployments - means workload outages don't usually mean an immediate loss of service or revenue. Because of that, the money from vendors isn't going into products that can offer the highest level of redundancy.

"I don't think they're targeting trying to meet conventional business IT at the redundancy standards that they've had in the past," says Davis. "Because liquid cooling is still primarily in the HPC and AI world, it's not yet being asked to meet that need of redundancy."

She suggests that workloads that require both high availability and liquid cooling might be better off focusing on resiliency in the software layer rather than in the cooling hardware. Better to fail over to another server or rack than try to fail over to another pump or pipe.

Despite the high price of GPUs and the difficulty sourcing them, Davis suggests that hyperscalers are well versed in this kind of failover thinking and will have the funds to take the hit because the target is speed to a finished product i.e. a trained model. Likewise, she suggests the smaller GPU cloud providers might simply be willing to risk running GPUs hotter and leaner in the near term to achieve similar goals.

And after we move on from training models to day-to-day inferencing? Microsoft recently suggested it is seeing huge demand for its GPUs for inferencing, but the market is yet to fully shake out on what hardware will do much of the day-to-day AI legwork - and therefore what kind of cooling will be required.

"Some people will choose to run their inferencing on CPU hardware," says Davis. "Some will run it on some more application-specific silicon. There's going to be a diverse set of hardware serving inferencing. Some of that will be liquid-cooled, some of it not."

What about the rest?

Beyond the high-end GPUs and AI-specific chips, questions remain about whether other types of compute will require liquid cooling.

Uptime's Davis predicts the industry will see "significant" segmentation, where extreme densities are reserved for AI training for "quite a while" yet. For non-AI workloads, the jury is still out.

Colovore's Coughlin suggests there are plenty of data-intensive applications and services that customers are deploying and utilizing that don't require GPUs and don't require AI.

"We've got a number of healthcare data providers that are using CPU servers, not even GPUs, and a fully-packed cabinet of most modern CPU server systems hits 25kW."

He notes a rack full of some of the highest-density storage systems could also reach north of 20kW; not quite at the tipping point of absolutely needing liquid, but not far off. Networking, while getting denser, is still some way off needing liquid.

"We're not there yet," Coughlin says. "Will we get to a point where they need to be liquid-cooled? I don't know, all these pieces of underlying hardware have gotten more and more dense."

Davis suggests conventional business IT currently lacks the overall incentive to densify. And while some hardware OEMs are offering increasingly dense storage and networking devices, she thinks they're likely to remain the preserve of the small minority at the top for a time to come.

"It's going to be several years before those super high-powered CPUs and really high-density products really take up in large volume."

If liquid cooling becomes standard across more facilities, it's entirely possible that all-liquid halls like Colovore's might become more common – and drag non-GPU workloads into being liquid-cooled by default. Other providers, however, remain set on offering flexibility, which means air-cooling will remain viable for a while to come. Resilience and redundancy, however, will remain top of mind, whatever the medium.

How to react to AI from the architectural standpoint of hybrid cooling

Should artificial intelligence require natural coolant?

The speed at which data centers have been required to adapt to the new realities of artificial intelligence (AI) has caught many on the hop. There's little doubt that liquid cooling is part of the picture, but should it be the whole picture? The short answer is 'no' – air-based solutions can cope with most traditional loads and many liquid cooled servers require at least some air cooling. So what's the solution?

Depending on site density, a hybrid approach could be the way forward, combining the best facets of air and liquid to the workloads that need it most. However, the denser the workload, the less viable air cooling becomes, as demonstrated in *Figure 2*, creating quite the conundrum for the frazzled data center operator.

This is why Schneider Electric has committed to an end-to-end approach, encompassing knowledge and advice from chip designer to end user. We caught up with Maurizio Frizziero, director, cooling innovation and strategy at Schneider Electric, to look at the challenges, and how this end-to-end mindset can begin to solve them. But first, how does liquid cooling work?

"Between servers and heat rejection units, cooling distribution units (CDUs) are the backbone of liquid cooling architecture," Frizziero tells us. "CDUs ensure optimized flow, temperature and pressure control to the technology cooling system (TCS), while manifolds distribute fluid throughout the servers." *(Figure 1)*

"Chip heat is then rejected into the fluid and distributed via cold plates, maintaining the device's optimal operating temperature. Economization or free-cooling are used as the primary means of heat rejection, designed with flexibility and efficiency in mind. If required, chilled water and direct expansion solutions can be utilized should liquid cooling systems or auxiliary rooms need it."

Having explained the basics, Frizziero emphasizes that while AI has put liquid

Maurizio Frizziero, Schneider Electric

front and center of the industry mindset, it has been available for more than a decade, and its popularity is a result of densification, generating even more heat. Rollout is costly, and to date, there are no perfect solutions at these scales.

Since the pandemic, he points out, the data center industry has invested a lot into facilities that are either newly online or in development where a hybrid approach seems the most appealing because it's a lot easier to add liquid cooling from the outset than a later retrofit.

Although Schneider's approach to finding a solution is "agnostic," they cite single-phase direct-to-chip as providing the best balance for the widest number of use cases. However, every data center, every client, and every situation is different.

The lack of standardization represents a significant challenge for the industry which stakeholders are now rushing to solve. With local, industry, and federal regulations, the additional vector of cooling, with its countless variables, can leave a data center designer blinded by choice, a situation Frizziero believes is not tenable:

"In the United States, regulation – particularly in relation to refrigerants for example – can be similar to Europe, but completely different to Asia. The industry needs to converge to a subset of standards because otherwise, people

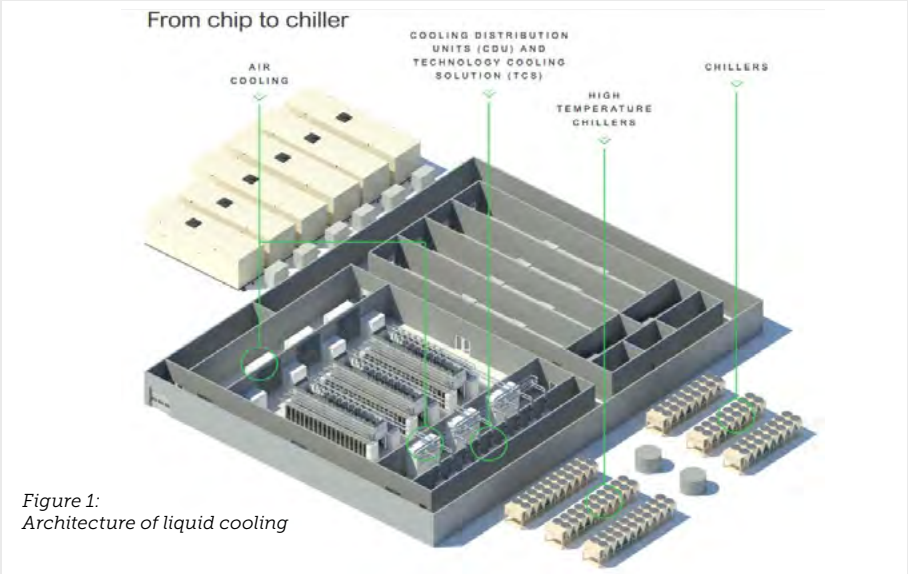
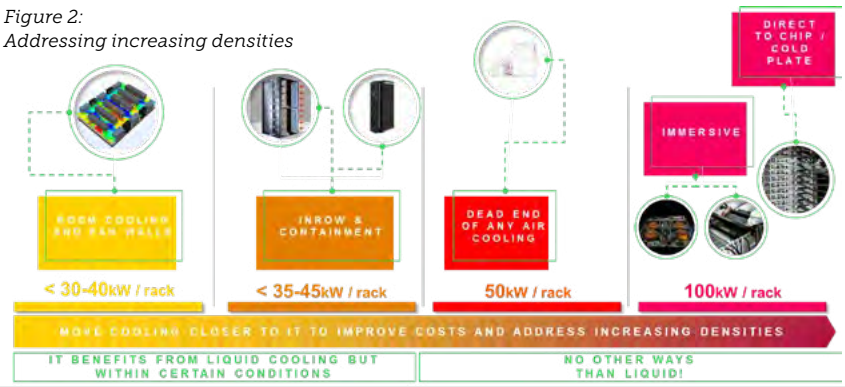


Figure 1:
Architecture of liquid cooling

Figure 2:
Addressing increasing densities



get crazy. You cannot have a different solution in every different location, in every data center."

Frizziero is keen to emphasize that at a time when sustainability is front and center, there is a fine line between action and greenwashing, especially as many of the initiatives that we may see as positive can be counterproductive once additional densities are brought into the equation.

A *recent report* from the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) pointed out that once densification increases load averages, energy to reach the required temperatures for liquid cooling will be more akin to the inefficient air-cooled data centers of the early 2000s, thus negating any net gain in carbon footprint reduction.

"We need to be very clear when we combine sustainability, decarbonization, and liquid", he warns, "It is easy to start greenwashing liquid cooling as sustainable. That's only a portion of the sentence. If you don't mention density you're going to grow rather than reduce footprint."

Frizziero has already alluded to his concerns regarding refrigerant standards, but here, he ties together its twin problems of sustainability and regulation:

"Schneider has been leading the transition to low global warming potential for years. The next challenge is migrating from 'synthetic' to 'natural' fluids. In general, people think something natural is better, forgetting that there are plenty of natural poisons in the world, so saying that natural is better than synthetic isn't progress, it's pure greenwashing."

He cites carbon dioxide (CO2) coolant which, when used in heating systems, is an excellent coolant, but its efficiency is so low that in a data center environment it can actually increase carbon footprint. So is natural always better? It's a constant tug-o-war:

"The point is not to look with just one angle," says Frizziero, "you have to take our end-to-end approach to have a full picture of sustainability. It's not just 'natural' instead of 'synthetic' because we can use 'natural wood' for many things. That doesn't mean it's better than 'synthetic wood' if we need to destroy the forest to build with it!"

Heat reuse is another example. Great on paper, but even if the target heating system is within the radius of the data center, the waste heat generated is not hot enough to pump straight into nearby homes, requiring more energy to be expended to bring it to a usable state.

All this need not be a crisis, but rather an opportunity, best leveraged as a cooperative industry, not single players. There is strength in numbers, and the fittest will survive as a pack:

"We are shooting at a moving target and it's pure Darwinism. If you don't evolve, you get extinction. The good thing is that the industry is running in the same direction, thanks to AI creating a widespread challenge."

Lest we forget that even if every data center were to switch to all-liquid tomorrow, they would still require air cooling because something needs to keep the power supplies and switch gears running at optimum temperatures, and indeed, keep the engineers warm.

This speaks to Schneider's desire to see end-to-end engagement across the entire ecosystem to solve these issues. Frequently data centers look down the chain, but Frizziero reminds us:

"We shouldn't forget that there is an opposite journey. At Schneider, we are working as a solution provider, from grid to chip, from chip to chiller, and on design, which is downstream, from us to the community, but also upstream from us to the server manufacturers."

Want to talk to the expert?

Reach out to Maurizio Frizziero at: *maurizio.frizziero@se.com*.

Life Is On Schneider Electric

The leak busters

Matthew Gooding
Features Editor

Water leak detection systems have often been an afterthought for data center operators. In the age of AI and liquid cooling, they have assumed much greater importance

Credit: Diamond Controls

In April 2023, Google services went offline across Europe as the company's europe-west9-a cloud zone was powered down. Customers, including mobile network Orange and video game developer Ubisoft, were reportedly impacted by the 24-hour outage.

The root of the problem was a fire caused by a cooling system water pipe leak at a data center in Paris used by the company and operated by Global Switch. According to Google's incident report, the leak "originated in a non-Google portion of the facility, entered an associated uninterruptible power supply room, and led to a fire.

"The fire required evacuation of the facility, engagement from the local fire department, and a power shutdown of the entire data center building for several hours."

Global Switch's other customers also lost access to their servers as a result of the blaze, and though it was swiftly brought under control, the incident highlights the damage that even a small water leak can do if not detected swiftly.

Luckily, help is at hand in the form of water leak detection systems. Though these systems have often been an afterthought in the data center construction process, the increasing amounts of liquid flowing through data halls in the AI era mean they are more important than ever before.

Taking a leak

Water leaks can emanate from a variety of sources. Computer room air conditioning units, of the type found in most older data centers, have liquid flowing through, so any damage to these systems can lead to flooding.

For smaller data centers, particularly those located in larger, general-use buildings such as office blocks, corroded pipework elsewhere can be a major issue, says Henry Ettinger, European service manager at Infiniti, a vendor that provides a range of data center solutions including leak detection.

"If you're an SME company you might be in a high-rise building with a kitchen or a toilet directly above your data center," Ettinger says. "There is not much in terms of protection between these facility rooms and data centers that facilitate critical hardware and supporting systems."

As well as pipes, water can enter through leaky roofs, or due to human error, be that from accidental spills or improper maintenance of equipment, Ettinger adds.

For larger data centers, the challenges around water leak detection are somewhat different. "With bespoke buildings, a lot of the problems are engineered out pretty well," says Iain Ames, director at Diamond Controls, a company that specializes in leak detection for data centers and other industrial and commercial settings. "Though you can get condensation in the cooling corridors in the summer."

External factors, such as extreme weather, can also be a source of leaks, and as well as the threat such incidents pose to IT equipment, data center operators have to beware of the impact that leaks can have on the world around them. In November, a data center at an industrial park in Offenbach, Germany, suffered a leak that saw cooling water seep into the soil. The leak originated in the pipe system on the roof, and entered the ground beneath the building via a rainwater infiltration system.

The cooling water reportedly contained a "low concentration of additives for corrosion protection and preservation" and two of the substances in it were considered hazardous. Fortunately for the unnamed developer of the data center, the nearest drinking water wells were located 1.5km away, so the chances of contamination were low, but since the leak occurred, environmental health authorities in the region have been continuously monitoring the groundwater for signs of contamination.

Not only are the environmental stakes getting higher, but the amount of water in data centers is also on the rise. While cooling fluid is present in traditional air-cooled systems, many developers, particularly those operating high-density AI-focused environments, are plumping for liquid cooling solutions. These can involve either direct-to-chip cooling, where cooling fluid is pumped directly onto cooling plates which keep components running at a steady temperature, or immersion cooling, where servers are dunked into vats of fluid.

Because of this, and the increasingly large investments being made in data centers, operators are ratcheting up their spend on leak detection technology. *DCD* spoke to several vendors that have seen

> *"We're starting to see customers demanding higher levels of coverage, showing more respect to leak detection systems"*
>
> *>> Iain Ames*
> *Diamond Controls*

the size of the orders received for systems mushroom over the last 18 months.

Splashing out

Diamond Controls is one company which has found its services in high demand. It works with a range of data center companies including major colo providers.

Ames says the need to mitigate the risk of water damage has never been greater. "There's more money in data centers than ever before with all the spending on AI," he says. "Because of this, people want a foolproof investment and don't want to take risks - they need to be able to certify and prove to their clients they've got safeguards in place."

This hasn't always been the case, Ames says. He founded his business in 2004 as an electrical contractor, carrying out a variety of jobs for clients including the installation of building management systems (BMS) at data centers and other industrial buildings. The BMS provides a central point of control for all electrical systems in a property, and through this work, Ames and his team became acquainted with water leak detection.

Spotting an opportunity, in 2008 Ames refocused his company to concentrate primarily on leak detection. "We'd previously just been doing installations," he recalls. "But we started offering services across the whole life cycle of a system, including design, maintenance, and training. That has allowed us to become a specialist business and do it really well."

Specialists were needed at the time, according to Ames, because water leak detection had been considered an afterthought when installing a data center BMS. "Specifications for these systems are often not project specific, consultants just copy and paste them in.

"We can bring an on-site perspective to the design process, and we're starting to see customers demanding higher levels of coverage, showing more respect to leak detection systems, and having more engagement at the design stage. This is how you get a better solution for the end user, and it's gone from an area where people would look to save money to something that's an important part of the design and build process."

Tale of the tape

While many parts of the data center have undergone rapid technological change in recent years, water leak detection technology has remained relatively unchanged.

Detection systems can broadly be split into two types: those that use water sensing cables, known as "leak detection tape," and spot sensors.

Infinity's Ettinger says his company recommends tape-based set-ups for most data centers. "These are physical sensors that provide a secure perimeter around the racks, potential water leak sources such as pipework, or even the perimeter of the data center," he says. Detection tape involves installing meters of cable that snake around a data hall's servers, either running along the top of racks with a drip tray underneath to catch any water that falls from above, or installed below the data center's raised floor.

"Water leak detection cables consist of multiple conductive wires encased in a protective, flexible material," Ettinger says. "These wires are usually spaced apart by non-conductive materials to prevent contact under normal conditions. When they come into contact with water, an alarm will be triggered."

Indeed, the detection tape is based on the principle of a simple electrical circuit. The tape will usually contain two stainless steel wires with an insulating material between them. When the wires come into contact with water or any other conductive liquid, it reduces resistance, completing the circuit and triggering the alarm or alert on the facility's BMS.

Lengths of tape can be up to 50 meters, and a data center will typically be split into different zones, meaning a leak can be pinpointed to a particular area of the server room. Diamond Controls' Ames says this can be as small as a single square meter, making it easy to find and address a leak. "The important thing is what the BMS does with the information it receives from the leak detection system," he says.

"Many come with an integrated digital leak map."

Ames says tape is "very sensitive," meaning it can pick up signs of a leak very early and potentially spare a data center from major damage. But there are downsides to this, too, he explains.

"A common problem with tapes is that if they're in areas of high-volume foot traffic, or places where plant equipment is being moved around, they can easily get contaminated," Ames says. "This means they give alarms frequently and start to be ignored."

For such areas, spot detectors are available which can be placed around the room, or under specific pieces of equipment. These sound the alarm when they come directly into contact with water, but don't offer the same range of coverage as tape.

Liquid cooling has added an additional dimension to the leak detection conundrum, but Ames says developers are taking "more precautions" when it comes to moving coolant around new data centers. "They're being a lot more careful about how they do it," he says. However, liquid can pose problems in the design phase, he adds. "I've seen contractors occasionally try to run pipework through electrical switch rooms with a drip tray underneath," Ames says. "In that situation, I would always try to get in early and ask them to reroute those pipes."

> *"My overall message when it comes to leak detection is to 'keep it simple.' These systems are there to detect water, so you don't need to overengineer them"*
>
> *>> Iain Ames*
> *Diamond Controls*

Keeping it open

When it comes to running water leak detection systems, Ettinger says data center operators must stay on top of maintenance to ensure their equipment is protected.

"The general rule of thumb with detection tape is that it needs to be replaced every ten years," he says. "Our advice is to carry out annual maintenance checks, and test each individual zone and the emergency backup batteries."

Ames says that when installing new systems, opting for an open protocol can make life easier. Vendors such as nVent build their technology using open standards, meaning it is easier to connect and maintain.

"With an open protocol, multiple installation companies can work on a system," Ames says. "That way, the end user gets a system that's manageable and offers better long-term value. It also makes it easier for a company to train its own engineers so that they can keep it maintained throughout the year."

When it comes to training, Ames says a lot of data center companies are still neglecting leak detection systems. "Very few end users get us in to train their staff," he says. "We often do some after installation, but then people move on and accurate records aren't kept, so no one knows what has been done and by whom."

He adds: "My overall message when it comes to leak detection is to 'keep it simple.' These systems are there to detect water, so you don't need to overengineer them.

"But once you've got coverage, it is important to consider how you will react when you get an alert.

"How do you respond? Which valves get shut down? Users often buy a leak detection system, arrange coverage, but don't start looking up the chain. It's important to take a holistic view."

Consistency is key: Data center humidity

Niva Yadav
Junior Reporter

Just how important is humidity in data centers

Like lock and key, cooling and humidity come as a pair. Controlling the humidity of your racks is just as important as controlling the temperature. However, humidity is often the forgotten half of the equation.

Not too moist and not too dry, balancing the humidity can often be a challenge. Too much humidity can lead to the erosion and damage of key metal components across the white space. However, you still need some moisture to avoid the build-up of electrostatic discharge, capable of creating sparks that interrupt operations.

Finding the balance between the two is a challenge, says Ken Fulk, vice president of the American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE).

ASHRAE now recommends a humidity range of between 20 and 80 percent relative humidity. It's called relative humidity because the figure takes into account the outside conditions. Fulk explains there is "no magic number" and, in reality, most components nowadays can withstand humidity ranges of between 10 and 90 percent before experiencing any undesirable outcomes. Because of this, fixating on a number or a range is often unhelpful, he says.

Steve Skill, senior application engineer at Vertiv, adds that typically it is legacy data centers that are most sensitive to humidity fluctuations, as newer facilities are equipped with more resilient components, capable of withstanding higher and lower humidity levels.

Ultimately, Fulk says, "components don't like change," so controlling humidity is more about keeping it the same, rather than aiming for a specific number.

In 2008, ASHRAE issued new guidance, reducing the lower limit to 20 percent from the previously issued 40 percent. It also set guidance on dew point, another measure of humidity.

Outages caused by humidity issues are rare, but not unheard of. Last year, the University of Utah experienced widespread IT failures due to increased humidity outdoors causing increased humidity inside, triggering an outage.

Maurizio Frizziero, director of cooling innovation and strategy at Schneider Electric, explains the ideal dew point sits between 5°C (41°F) and 15°C (59°F). The dew point is the temperature at which any body of air can no longer hold water in gas form; essentially, the temperature at which condensation happens.

Vlad-Gabriel Anghel, director of solutions engineering at DCD's training unit, DCD>Academy, says that dew point has become a popular choice for monitoring humidity as it monitors each rack with an individual sensor. He explains that the temperature and humidity can vary across each rack in a data hall, depending on how close they are to the cooling or humidity system. Dew point allows operators to monitor the humidity levels of racks on an individual basis using monitors attached to each rack.

Getting hot and steamy over time

Fulk has been designing and deploying data center cooling systems for decades. When he first began, data center facilities had a fixed density with "heat being produced on a square foot basis." Conventional air conditioning was

more than enough to distribute the air and maintain optimal humidity levels. Since then, and more crucially since the surge in AI and cloud computing workloads, maintaining humidity is not as easy as installing one "clamshell" arrangement, says Fulk.

"You can no longer have nice rows of equipment and perfectly arranged computer rooms with air conditioning units," he explains. Selecting the ideal humidity range for a facility is complicated by the fact that different components have different parameters.

ASHRAE has set classifications for data center components: A1, A2, A3, and A4. DCD>Academy's Anghel explains that these classifications effectively measure how sensitive a data center component is to humidity and temperature changes. For example, A1, the strictest classification, states that equipment in that category can only operate between 15°C (59°F) and 32°C (89.6°F) at between 20 to 80 percent relative humidity. It is in A1 that you will find the majority of servers and the most critical infrastructure.

Anghel says the difficulty is selecting the right humidity and temperature that satisfies all components of your data center ecosystem. Between all the classifications of all the equipment, there is likely an overlap, or in other words 'the sweet spot.' Setups have to be thought of as an entire ecosystem, with considerations as to how those parts interact with each other.

The most important consideration is the cooling method. Fulk says liquid or direct-to-chip cooling typically "lessens the impact of humidity" because the chip is essentially surrounded by moisture anyway, proving less of a challenge to keep consistent. Even still, Fulk adds that an operator has to ensure that humidity is low enough to prevent condensation based on the temperature of the liquid being circulated in the system.

Frizziero agrees that liquid cooling systems require far less precise control of relative humidity. However, Fulk cautions that the building could easily grow mold in high humidity even if the servers remain unharmed.

Skill adds: "Certain chemicals in the atmosphere of a site can react with the moisture in the air to form acids." Chlorides and sulfides in coal-burning countries like China and India can produce acids that could easily corrode components. So, keeping humidity low is still important.

> *"You can no longer have nice rows of equipment and perfectly arranged computer rooms with air conditioning units"*
>
> *>>Ken Fulk, vice President at ASHRAE*

Previously, data centers also had to consider the people working inside them, says Skill. High humidities made for intolerable working conditions. He says nowadays, facilities have fewer people working inside them, so this has become less of a consideration when selecting the optimum humidity.

Free cooling can be costly cooling

Free cooling can pose a challenge for maintaining consistent humidity, but only in terms of energy, Fulk says. For example, when operating in a humid location like Dallas, the air being brought into the facility must be 'dried' before it hits the servers, and this requires more energy.

Outside air is typically dehumidified by bringing the surface of the heat exchanger to a temperature lower than the set dew point, explains Frizziero. The air passing through leaves with a fraction of the water as when it entered.

Fulk says somewhere like Phoenix might be warmer than Dallas, but is far less humid, thereby costing the operator far less in terms of energy. The opposite can be said for the Nordics. Granted, they have become a popular location for their cooler temperatures, but that is not to say the air will be dehumidified enough.

However, even if the air does need to be humidified or dehumidified, it will still require considerably less energy than "putting in refrigerants and glycol in a closed loop system," says Anghel. Plus, it will be more cost-effective in comparison to an operator using liquid cooling in the same location.

It is more likely that air would need to be dehumidified than humidified, generally speaking, says Skill. In these cases, operators would typically install a dehumidification plant separately from the facility to treat the air before it enters inside. Conapto is an example of a Nordic operator using a separate plant to treat air before it goes inside the facility.

"Every operator will be thinking about sustainability," says Fulk, because cooling, temperature, and humidity control all require energy. When designing a facility, great consideration will go into making that facility as energy-efficient as possible.

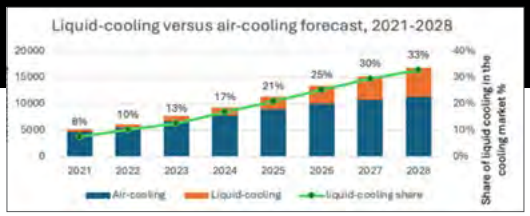
"Specifically with AI, more so than cloud computing, it is important for cooling systems to be as efficient as possible to reduce the impact on our environment," he says. Operators should be picking locations strategically to avoid the added expense and added environmental burden of dehumidifying and humidifying air.

AI might be the buzzword today, but it won't be the only thing raising densities. There will soon be something else that changes compute as we know it, adjusting ASHRAE's recommendations and the relevance of humidity. Until then, the best an operator can do is to stay consistent.

Liquid cooling in the GenAI era

Vlad-Gabriel Anghel
Head of solutions engineering, DCD>Academy

Requirements for liquid cooling in data centers are evolving fast, and vendors are reacting accordingly



With the latest H100 Nvidia chip drawing up to a whopping 700 watts when configured on a SXM socket and a hefty 400 watts when configured via PCI-E, it's no wonder that 2024 has been the year where liquid cooling has shot to the forefront of minds throughout the data center industry.

Technical Specifications

	H100 SXM	H100 NVL
FP64	34 teraFLOPS	30 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	60 teraFLOPS
FP32	67 teraFLOPS	60 teraFLOPS
TF32 Tensor Core*	989 teraFLOPS	835 teraFLOPS
BFLOAT16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS	3,341 teraFLOPS
INT8 Tensor Core*	3,958 TOPS	3,341 TOPS
GPU Memory	80GB	94GB
GPU Memory Bandwidth	3.35TB/s	3.9TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	Up to 700W (configurable)	350-400W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGS @ 12GB each
Form Factor	SXM	PCIe dual-slot air-cooled
Interconnect	NVIDIA NVLink™: 900GB/s PCIe Gen5: 128GB/s	NVIDIA NVLink: 600GB/s PCIe Gen5: 128GB/s
Server Options	NVIDIA HGX H100 Partner and NVIDIA-Certified Systems™ with	Partner and NVIDIA-Certified Systems with 1-8 GPUs

The AI boom has forced operators to look beyond the traditional air cooling solutions that the vast majority of data centers leverage to keep their IT systems running efficiently. Novel liquid cooling solutions are coming to the fore, driven by the need for owners and operators to come up with completely new designs for their new greenfield facilities, with the majority also having to balance this against the retrofit and upgrade of current brownfield sites.

These workloads, at the moment, only seem to be getting bigger across all requirements - be it the need for power, cooling, bandwidth, or data storage. For example, Nvidia's upcoming Blackwell generation takes power consumption to new heights. The B200 GPU is expected to draw up to 1,200W, while the GB200 - featuring two B200 GPUs paired with a Grace CPU - could reach an astounding 2,700W. This marks a staggering 300 percent jump in power consumption within a single GPU generation, reflecting the accelerating energy demands of AI systems.

The liquid cooling market is also experiencing a bit of a limelight moment - with analysts placing the data center cooling segment to reach a staggering $16.8 billion dollars by 2028, at a 25 percent CAGR, with liquid cooling emerging as the predominant technology and biggest driver of this.

As AI compute loads expand with increasingly wider and more intensive deployments, ultra-high-density AI racks are becoming a reality. These racks can demand 100kW of power and house equipment valued at more than $10 million per rack, often relying on direct-to-chip or immersion liquid cooling. This shift introduces significant challenges in delivering adequate power, space, and cooling to accommodate unprecedented workload levels.

Before settling on an AI compute cooling strategy, owners and operators must evaluate a broad spectrum of engineering considerations. These decisions should account not only for supply chain constraints but also for long-term corporate ESG and sustainability objectives.

From the plethora of liquid cooling solutions, it seems cold plate technology and wider direct-to-chip solutions are

leading the charge in terms of adoption. The preference for direct-to-chip and, specifically, cold plate liquid cooling is attributed to its effectiveness in handling high-density computing environments and its compatibility with existing data center infrastructures. This method offers a balance between performance and implementation complexity, especially for brownfield sites and retrofits. That being said, greenfield sites will most likely drive an uptick in immersion cooling deployments - either single-phase or two-phase.

Since late 2022, vendors have been hard at work finding the middle ground between innovation and risk management bringing new solutions to the market with some blurring the lines between direct-to-chip and immersion.

Accelsius unveiled its NeuCool two-phase direct-to-chip cooling solution in April 2024. It utilizes a dielectric refrigerant that evaporates upon absorbing heat from high-power components like CPUs and GPUs. The vapor is then condensed and recirculated, creating an efficient cooling loop. This supports up to 1,500W per socket and up to 100kW per rack, making it suitable for current and future high-performance computing needs. Furthermore, the system works well for older, air-based equipment where the heatsinks are replaced by their proprietary CPU and GPU "vaporators." These are designed to slot in the same location and form factor as the heatsinks so this can become the de-facto solution for existing facilities that need an upgrade to support the requirements of these new workloads.

This solution can work with or without the use of water and is entirely modular. As such, it allows for seamless integration into existing data center infrastructures. It accommodates standard server racks, facilitating ease of deployment across a range of facilities from Edge to hyperscale data centers. The system is compatible with various final heat rejection systems, including waterless, pumped-refrigerant options.

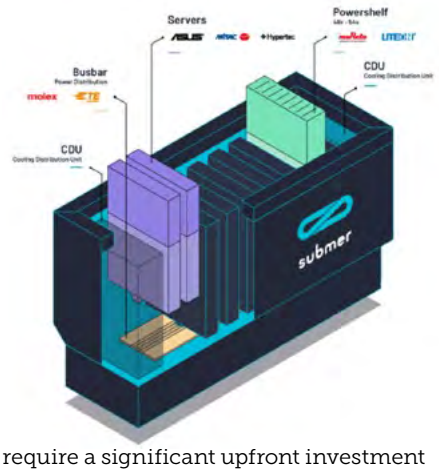
Chilldyne, on the other hand, has come up with a direct-to-chip solution that eliminates one of the main risks of liquid cooling - leaks. The solution works by creating a vacuum that draws in and circulates coolant into the system and circulates it across cold plates mounted on the CPU/GPU.

Should a rupture occur in one of the tubes carrying the liquid, the vacuum ensures no liquid is spilled and air is drawn in. The system also continuously monitors for pressure changes and alerts the owner/operator if something has gone awry. Furthermore, in certain configurations the system may isolate the affected section, minimizing the impact on other components. After releasing this, Chilldyne turned its attention to brownfield sites, and in July 2024 launched a plug-and-play liquid cooling starter kit designed to modernize data centers and support AI workloads. The kit includes two CDUs and cold plates rated up to 2,000-watt TDP, supporting up to 150kW cooling per rack.

Other vendors are betting big on the new wave of greenfield sites coming online right now with companies like Asperitas, LiquidStack, and Submer hedging their bets on a different flavor of liquid cooling - immersion.

These systems are either two-phase or single-phase and revolve around the concept of immersing the servers directly into a tub of dielectric fluid and using the high heat-carrying potential of these fluids to move the heat away from the IT equipment. These designs offer an extremely high cooling efficiency, but also pose challenges, particularly around integrating them into already existing air-cooled data centers.

Liquid cooling systems (of any flavor)



require a significant upfront investment in equipment regardless of whether it is deployed in brownfield or greenfield sites. And while liquid cooling offers long-term energy savings, owners and operators are still on the fence when it comes to the need for liquid cooling in smaller, low-density facilities.

Adoption of liquid cooling within the data center space is driven by the workload itself, so facilities not offering AI or HPC services are unlikely to see a need to upgrade their cooling infrastructure, as air cooling is more than sufficient for most use cases.

The lack of standardization across the parts making up the liquid cooling solution is another blocker towards adoption - components like manifolds, reservoirs, and even the way the CDU is connected, vary from vendor to vendor. Regulatory pressures coupled with the drive for a sustainable data center industry have made certain solutions less desirable even though they boast great energy efficiency potential. These concerns primarily revolve around water usage, chemical impacts, energy consumption, and waste management. For example, certain fluids may require special handling and disposal processes due to their chemical properties.

The liquid cooling market still has considerable room to grow, and as the industry works towards standardization and vendors keep coming up with solutions that cater to the plethora of risks that this technology brings, the breadth of adoption will only increase. After all, the laws of physics have not and will not change anytime soon: Broadly speaking, every added watt of power needed by a chip means one watt of heat through energy transfer that needs removing.

Keep it cool in the era of AI

EcoStruxure IT Design CFD by Schneider Electric helps you design more efficient, optimally-cooled data centers

Optimizing cooling and energy consumption requires an understanding of airflow patterns in the data center whitespace, which can only be predicted by and visualized with the science of computational fluid dynamics (CFD).

Now, for the first time ever, the technology Schneider Electric uses to design robust and efficient data centers is available to everyone.

Equipment Models – Range of equipment types allowing you to customize product attributes of anything from racks, to coolers, to floor tiles.

CFD Analysis Report - Quickly generate a comprehensive report of your data center with one click.

CFD Analysis – One of the fastest solvers in the industry delivering you results in seconds or minutes, not hours.

Visual Maps – Visualizes temperatures, pressure, airflow speed, velocity.

Optimize your data center's airflow using our IT Airflow Effectiveness and Cooler Airflow Efficiency metrics.

Reference Designs – Quickly develop your design from pre-built templates.

Custom Attributes – Easily enter your specific equipment attributes for accurate modeling.

Cooling Check – Shows you if your equipment is optimally cooled according to ASHRAE's thermal guidelines.

Get Free Trial

Life Is On Schneider Electric

The crypto pivot to AI

Sebastian Moss
Editor-in-Chief

Even as Bitcoin reaches record highs, miners are chasing a different digital gold

Bitcoin began in the bedroom. Initially intrigued by a mathematical curiosity, the first miners used spare compute on personal gaming rigs to unlock worthless amounts of virtual currency.

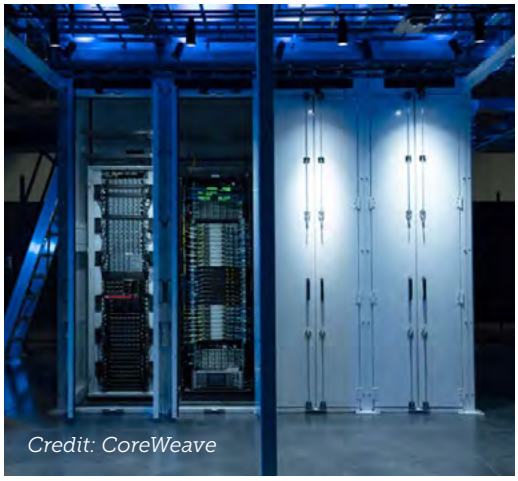
In the following years, however, as the cost of mining grew rapidly, and the value of the coin grew even faster, it would soon leave the home. An entire cottage industry sprung up, consuming as much power as nations, relying on custom silicon to chase billions.

This growth mirrored a similar explosive build-out in the traditional data center sector, but was always seen as separate from the Internet's infrastructure - a shadow data center network for a shadow financial system.

Now, however, the two worlds are converging. The rapid proliferation of artificial intelligence, and its dramatic power demands, have forced companies to reconsider what they'll accept from a data center provider, allowing historical outcasts to be remodeled as new industry titans.

"Our company is a data center business at heart," Core Scientific's CEO Adam Sullivan says. "We have the largest operational footprint of Bitcoin mining infrastructure, and that's required traditional data center people to operate

Credit: Crusoe Energy



Credit: CoreWeave

"Within a matter of seconds to minutes, you can take a miner and go from full power consumption down to a trickle and then pick right back where we left off. And it's beautiful"

it - because no one had operated nearly a gigawatt of infrastructure in the past, and you couldn't take regular Bitcoin mining operators and put them into that role."

Core Scientific is but one of a growing number of cryptomining companies that has pivoted to serve generative AI businesses. Some, such as CoreWeave, have completely abandoned their mining roots in search of new riches, while others are still trying to straddle both worlds.

Hive Digital Technologies, Northern Data, Applied Digital, Iris Energy, Mawson Infrastructure, Crusoe, and others are but just a small number of those that have made the shift.

"We're moving very quickly into building the infrastructure that's going to be powering the AI wave, and we've got a big pipeline of opportunities we put together," Crusoe co-founder, president, and COO Cully Cavness told *DCD*, adding that the company is aiming to deliver "gigawatts of new data center capacity."

The company launched in 2018 as a cryptominer deploying containerized data centers to oil wells to harness natural gas that would otherwise be "flared off" and wasted.

Unlike latency-sensitive workloads or ones beholden to regulation, crypto doesn't care where it is mined. This meant that the sector ignored most other location factors, instead giving primacy to power - focusing simply on cost and access.

AI training clusters are not too dissimilar, and as a result the data center sector is now chasing power. Inference data centers are more latency-sensitive, but the industry will still prioritize power access above all else, locating data centers in suboptimal locations simply because it's the only place with enough local grid capacity.

"We're really doing this in an energy-first way where we're going to places that have a lot of overdeveloped or excess, otherwise curtailed, power," Cavness says. "And sometimes these are in remote areas.

"We're taking that same ethos into the AI development space where we're going to some places that might have curtailed wind, hydro, or geothermal power. We still have some AI capacity being powered by our flared gas operations - such as in Montana, the Bakken oil field - and really just moving quickly to try to bring on as much capacity as we can."

The company, rather unusually, also has its own power generation infrastructure. "We own and operate 230MW of natural gas-fired power generation equipment, which includes turbines and reciprocating engines," he says.

Cavness argues that the wider industry is still trying to come around to terms with this new world. "Bitcoin mining has a pretty fast-moving, nimble, scrappy development cadence, and my sense of the way that the AI infrastructure industry is gonna move more towards that kind of rapid development and deployment," he says.

The company this November filed with the SEC to raise $818 million, on top of hundreds of millions in previous rounds. The month before, Crusoe entered into a $3.4bn joint venture with asset manager Blue Owl Capital to build a huge data center in Abilene, Texas.

The site is planned to be leased to Oracle, who will then rent it to Microsoft, for use by OpenAI.

These elaborate nesting doll deals are a function of this particular moment in time, where capacity is so constrained. Microsoft has long turned to wholesale colocation partners to supplement its data center capacity, but the demand for AI and the shortage of Nvidia GPUs forced it to loosen its control even more.

The hyperscale giant signed two blockbuster deals last year, with Oracle and CoreWeave, for access to data centers and GPU infrastructure. "The Oracle deal was a big one, we had to get Satya [Nadella, Microsoft CEO] to sign it off," one senior Microsoft executive told *DCD* under the condition of anonymity.

Microsoft's investment in OpenAI also came with a condition of cloud exclusivity, but the generative AI firm has pushed for more compute and faster capacity, causing Microsoft to have to turn to its competitors to keep up - as long as its name is at the end of the chain of companies.

"It's so hard to speculate where that type of deal is going to go," Core Scientific's Sullivan says.

"On one side, folks like CoreWeave are forming a very sticky business. They're providing a service to these hyperscalers that they really can't find elsewhere, shifting capex to opex while being able to move at a speed that is often unmatched in the data center.

"But on the other side, at some point, will there be disintermediation? Will Microsoft get to a point where their own economics require them to start going direct to the data center developers, or will their internal data center team catch up and be able to replace all that infrastructure?"

This infrastructure, Sullivan says, is not easy to "lift and shift." He explains: "These are hundreds and hundreds of megawatts of infrastructure that have been purpose-built for their use cases. So the big question is, given the stickiness, given the amount of investment in infrastructure, given that they already have the facility and the connectivity, will they be able to disintermediate in the future?"

Core Scientific is a part of that interconnected web of infrastructure. It operates its own crypto operations and HPC/AI business, but is also a major infrastructure provider for CoreWeave

Credit: Core Scientific

- such a large part, in fact, that the latter business unsuccessfully tried to acquire it earlier this year.

"We hosted them when they were an Ethereum miner, back in 2019-2022," Sullivan recalls. "Core Scientific was one of the largest hosters of GPUs in North America, so we had a close relationship with them. They knew our capabilities; they knew our sites."

That partnership abruptly ended when Ethereum went proof of stake, a different type of Blockchain technology requiring far fewer GPUs. "I think they were excited to move back into some of the sites where we previously had GPUs for them," Sullivan says. "Obviously, the infrastructure looks much different for this."

How different the infrastructure needs to be remains a point of contention. Crypto sites could often avoid much of the complexity of an enterprise endeavor, allowing them to be faster and cheaper.

Thin walls, low security, little redundancy, and rows of cheap ASIC rigs are often the hallmarks of a crypto data center - with the end result closer to a shed than critical infrastructure.

The companies that had the most ramshackle bare-bones facilities have found it the hardest to pivot, while the larger crypto companies that had more to work with have found themselves in a better position.

"If you look at our facilities and you take a look at Yahoo data centers, they look very similar," Sullivan says, referencing the traditional 'chicken coop' design that was once the rage. "So we essentially stripped back what was necessary to run Bitcoin mining from that design. And now we have a number of things to add back when we convert it to HPC. But from a structural standpoint, it is a very similar design.

> *"We're really doing this in an energy-first way where we're going to places that have a lot of overdeveloped or excess, otherwise curtailed, power. And sometimes these are in remote areas"*

"We add in chillers, batteries, gensets, UPS systems, and the whole mix of things that are necessary to do a conversion to HPC."

While this is still a lot of work and cost, Sullivan argues that other data centers are similarly facing a conversion challenge at this inflection point. "Many of the traditional data centers are having trouble converting existing data centers as they have to do full facility conversions due to the power density and the water-cooled nature of the newest generation GPUs," he says.

Crypto operators are currently reviewing their existing footprints to see which make sense for conversion - during which time they would, of course, lose any revenues from mining. At the end of it, they'll be left with less capacity.

"A crypto site that's 55MW on a piece of fixed land is not going to be 55MW Tier III," says Corey Needles, managing director of Northern Data's Ardent Data Centers. "No, it's going to be half of that, if not less, for all the infrastructure and at this density that I'm going to have to put in the ground."

The real question, however, is how much of this extra infrastructure is necessary. Security will be non-negotiable for most customers, while shifting from ASICs to GPU servers that cost as much as a house will necessitate more cleanliness, care, and temperature control.

"You have to be a lot more respectful of the AI and HPC hardware," says Applied Digital's CTO Mike Maniscalco. But, at the same time, he argues that there are lessons to be learned from mining's approach to power use and redundancy.

"Within a matter of seconds to minutes, you can take a miner and go from full power consumption down to a trickle and then pick right back where we left off," he says. "And it's beautiful. It just works really well where there's a dynamic power availability - sometimes power may be really, really cheap, so it makes sense to go full bore, but other times,

it may get expensive, and it may make sense to wind down."

For training runs, the power costs can be enormous, while adding redundant infrastructure is both costly and time-consuming. "When you talk to a lot of AI companies about how they design and engineer their training runs for resiliency, they expect some failures, bugs, OS issues and hardware issues," Maniscalco says.

"So what they tend to do is make checkpoints systematically throughout training runs. If the training was to fail because of a hardware or software failure, it just picks up from the last checkpoint."

This, Maniscalco says, "theoretically gives you the ability to wind down power quickly and pick up from the last checkpoint once that power is restored." He continues: "All you've lost is a small amount of training time, in theory. We really like that mindset of how you can change the power delivery and design to reduce costs to get things to market faster to save a lot of money on generator purchasing."

The market is still experimenting with the right way, "because they're very conditioned to having full redundancy," Maniscalco adds. However, in a stark reminder of the importance of redundancy, Applied itself posted a loss earlier this year after faulty transformers caused a lengthy outage at its North Dakota data center.

Most likely, Core Scientific's Sullivan predicts, is for the sector to adopt a mixture of redundancies.

"We're seeing a lot more multi-tier being developed for GPU clusters," he says. "If you take the traditional Tier III model, you might be spending $10 to 12 million a megawatt. You can come in a few million per megawatt below that number for standards that the customers are completely fine with.

"They understand that there are points in time where, if it's absolutely necessary, they can go back one or two minutes in terms of where the model was. That's worthwhile."

Another area he sees convergence between the two sectors is AI "moving into the world of ASICs," that is, specialized AI hardware instead of the more general GPUs. "I think we're going to start to see much more commoditized compute, similar to what we saw in Bitcoin mining."

> *"A crypto site that's 55MW on a piece of fixed land is not going to be 55MW Tier III. No, it's going to be half of that, if not less, for all the infrastructure and at this density that I'm going to have to put in the ground"*

OpenAI is believed to have contracted Broadcom to help it build its own ASICs for 2026, but whether they will be able to compete with GPUs (especially for training) is unclear.

It's also unclear what the future holds for both crypto and AI. The launch of ChatGPT in November 2022 is generally seen as the start of the current generative AI race.

At the time, a single Bitcoin cost some $17,000. For most of 2024, it has hovered around $50-60,000. The election of Donald Trump has sent it soaring to record heights however, with the virtual currency just shy of $100,000 at time of publication.

A promise of looser regulations, a national Bitcoin reserve, and economic uncertainty all seem to suggest continued high valuations.

"It has not changed our philosophy at all," Sullivan says. "Even with the more recent significant increase in the price of Bitcoin, mining economics are still very challenging."

Power costs remain a constraint, and show no sign of improving. At the same time, the ability to mine crypto has commoditized, making it harder for first movers to stay ahead. "And so that edge that you had is getting eked away," Sullivan said, although the company plans to continue its crypto business as its AI one builds.

Others *DCD* spoke to pointed to the challenge of serving a sector with wild price swings to create a digital coin with no intrinsic value. Similar things could be said about generative AI, which has yet to prove a sustainable business model, and is built on the promise of vast technological advances that are yet to come.

"You can really mitigate that risk through the credit quality of the tenants that we're building around," Crusoe's Cavness says. "So if we're taking a large, long term, that requires building a data center and a permanent location, multibillion-dollar investment, that needs to be tied to a high credit quality lease, that ultimately there's a guarantee that that's going to be paid."

Sullivan, similarly, says that Core Scientific is "focused on long-term contracts, some that are more than 12 years."

The challenge is, however, that in a speculative gold rush, only some of those building data centers actually have an anchor client.

"The part that does worry me is the amount of people that are building purely on spec right now that are going to be delivered four years from now, like 2028," Sullivan says.

"The industry could face some headwinds where we start to see some cracks in terms of demand, in terms of people being able to find clients, actually fill all that space, given the amount of capital being put into this industry right now. That is a major concern. Is the level of investment that is going to match where the demand falls?"

An interrelated concern is "the amount of debt that's being taken on to purchase the GPUs, it is probably one of the more worrisome aspects of the growth of this industry," he says.

For the AI sector, this may prove the final and most painful lesson to be learned from the crypto miners: During Bitcoin or Ethereum price crashes, the market has swept away those that timed it wrong, and were left holding debt when demand evaporated.

As both former miners and former traditionalists chase the same AI workloads, they will face a precarious gamble, one that could end in riches or disaster. Operators will be forced to ask the same question that has long plagued the Bitcoin sector: Can the line keep going up? ■


EQUANS
DATA CENTERS
Strategic data center partner
Expert delivery. Global Scale. Local precision.
Equans Data Centers is a strategic data center partner, delivering new build, fit-out, retrofit, and upgrade covering general contracting for hyperscaler and colocation requirements.
Bringing construction and engineering, global standardisation, and precision localisation, we deliver the highest quality, speed, and scalability, meeting increasing capacity needs across Europe.
Find out more
equansdatacenters.com

Grundfos Data Center Solutions

Keep your cool

Efficient water solutions for effective data flow

Meet your efficiency and redundancy goals

Smart pumps offering up to IE5 efficiency, redundant solutions that meet up to 2N+1 redundancy – whether you're planning a Tier 1 or Tier 4 data center, you can rely on Grundfos to keep your data center servers cool with a 75-year history of innovation and sustainability at the core of our corporate strategy.

Your benefits:

- High-efficiency cooling solutions saving water and energy
- Redundancy meeting up to Tier 4 requirements
- End-to-end partnership process

GRUNDFOS

Possibility in every drop

Doug's secret weapons

Dan Swinhoe
Senior Editor

NTT GDC CEO Doug Adams on taking the helm at a complex global company

NTT's data centers business is something of a Frankenstein's monster, forged from the remnants of various acquisitions made across the world over the course of 15+ years.

The company had long operated a data center business in Asia that it built itself, and the acquisitions of Dimension Data in South Africa in 2010, Gyron in the UK and NetMagic in India 2012, RagingWire in the US 2014, and e-shelter in Germany in 2015 have helped form its global portfolio, providing a foundation that has enabled the company one of the biggest players in the digital infrastructure space.

In 2015, *NTT Com CEO Tetsuya Shoji told DCD* that the company had big plans for the sector, and today it bills itself as the third-largest data center provider globally. NTT GDC's footprint globally now spans more than 1,500MW across some 150 facilities in 20 countries.

"There are really only a few data center providers that I would argue even have a truly global footprint," Doug Adams, NTT GDC's global CEO, tells *DCD*.

An *investor report from 2023* suggested NTT's data center business totaled $600 million in 2022, and would reach more than $1.4bn by 2027. Adams says the company is already beyond the $2bn mark, and has been growing at more than 20 percent CAGR.

When we last sat down with Adams in 2019, he was CEO of RagingWire, the data center firm he had helped found at the turn of the millennium with a single facility in Sacramento, California. He "stuck around" after RagingWire was sold to NTT to help run the business – which in the US today spans seven campuses and more than 900MW across California, Illinois, Oregon, Texas, Arizona, and Virginia.

Sticking around has paid off, and now Adams is running the company's whole data center business globally. After letting its disparate data center units operate under their own names for many years, NTT finally announced that all its regional brands would be rolled under the NTT Global Data Centers (GDC) moniker in 2019. Adams was appointed GDC's US CEO in 2020, and in June 2023, he was given a global remit.

While NTT's data center business has operated under the *single GDC brand* for half a decade now, behind the scenes it functioned as separate global regions.

"I've spent the last year and a quarter putting together a global team, putting together the strategy, and moving all 3,500 global GDC employees underneath my global management team," Adams tells *DCD*. "In the past, we had four connected

India, Navi Mumbai, NAV1A

but autonomous business units. And the charter, when I took over, was to ensure that we had economies of scale and we were running the business as efficiently as possible."

While NTT GDC still has four region heads, Adams now has global function heads for every discipline – HR, legal, marketing, operations, construction, engineering, product, etc – all drawn from the best candidates across the company.

"We are truly a global company now, with a global management team, one global P&L, and that frankly makes us a lot more agile," he says. "Instead of four separate units, we're now one unit. So one set of buying power, one standard design, one set of policies, procedures, and governance."

Switzerland, Zurich 1

One brand, one culture

The number one issue on Adams' mind when he started on this journey was how to "win the heart, souls, and minds" of what he describes as a diverse, large, multicultural organization.

"I wanted to make sure that we did this in a way that we took care of the employees first," he says. "If we don't make it a better environment for employees, it's a fail. So we took a year to figure out the process and how to do it."

Adams says most companies underestimate how much of a task it is to move from managing one geography to a global company.

"Making sure that we have an equitable, diverse, thoughtful, and talented employee pool I think is incredibly important," he says. "No one completely won. It was a compromise for all of us. We didn't just implement one culture and one team. Everybody had to bend a little bit, move a little bit, and change a little bit. I don't want to say it's a

Doug Adams

Frankenstein culture, but it is an amalgam of the best parts of each of the operating regions."

This is good for employees' career prospects, and Adams says NTT staff now have more opportunities to move across the business globally. He says NTT now benefits from "massive" economies of scale leading to better costing for raw commodities, more standardized products, better governance, stronger processes, all globalized. And for the clients, he says, they get a more aligned global experience.

"We're in the infancy of this. I don't want to claim victory," he says. "But the early results are very strong. We handily beat our plan and every single metric, bookings, revenue, operating income, every metric we're beating."

A change of scene

It was important, he notes, that the new-look global leadership team not just be his old confidants from the RagingWire days; a multicultural management team was "mandatory."

"I've worked with my team in the US for more than 10 years. It would have been pretty easy to move them up to the top management positions and not create a truly global company. But in the long run, you don't get the strength and the resilience and the diversity that you get from having a truly global management team."

For Adams personally, he says adjusting to working across multiple timezones has been a challenge.

"It's difficult to operate at 4am in the morning and 2am in the morning," he notes, talking at a more reasonable hour from the US. "It is more difficult, but more rewarding, to manage a global company."

On the flip side, he says he has learned to deeply appreciate the individual cultures and things people bring to the table.

"Prior to taking over this role, I didn't have a lot of exposure to people beneath the top-level management that I do now," he says. "The strength of the team that's been developed across the globe is significant. What I enjoy and what I love is working with the people, and I think that's our biggest differentiator."

As well as being global CEO, Adams has still been moonlighting as CEO of the US business for more than a year – when we spoke in October 2024, he was finally set to hand over the reins of the US biz to Joe Goldsmith, who will be stepping up from chief revenue officer of GDC's Americas business. Joining NTT in 2018, Goldsmith previously held the same role at Vantage, and spent nearly a decade at Digital Realty before that.

"It's a little bittersweet. The US will always be my baby," Adams says. "But it opens a great role for Joe. He'll do an incredible job, and it allows me to concentrate on ensuring that we are doing a better job on globalizing.

"I like to talk about how we're borderless, but at the end of the day, you can't just centrally manage an organization that's as large as ours from a central location, you have to have those regional folks that have the local client intimacy, the local employee intimacy, understand the legislations and have government, city, and county contacts."

Data centers are getting bigger

NTT bought 80 percent of RagingWire in 2014, completing the purchase in 2017. When *DCD* spoke to Adams in 2019, it was in the wake of many telcos offloading data center assets to whoever would buy them and NTT noticeably bucking the trends and buying up operators in major markets globally. The industry's biggest and latest facilities were still only offering capacities in the low tens of megawatts.

In 2024, the winds have changed and everyone is focused on AI. Hyperscalers,

AI cloud firms, and even telcos and enterprises are seeking capacity en masse to cope with the power demand of today's GPU hardware. Data center campuses are often in the hundreds of megawatts. Outside of retail colocation, NTT generally offers 6MW data halls as a minimum, but is signing much larger leases.

A major shift in recent years has been the switch to liquid cooling. Adams says the company has more than 150MW of liquid cooling capacity in production or in the latter stages of development – a more recent call with the company suggests that figure is now above 200MW.

"What we're seeing in the market is so much different than we saw just a few years ago," Adams says. "We're getting massive takedowns all across the world, especially in the US and India for liquid-to-the-chip AI deployments. I think we have one of the largest footprints of liquid of any of the competitors."

According to Adams, this equates to "a substantive lead within the marketplace because of the amount of liquid-to-the-chip we've actually deployed." He says that, while the competitors are talking about liquid cooling, "we've actually done it," and adds: "I think we have some very substantive learnings."

Most of NTT's liquid deployments are direct-to-chip, but the company has done some limited immersion deployments in Asia.

Immersion is "not easy and brings a whole new set of unique challenges," Adams says. "But one of the massive values of globalizing businesses is we just bring the best practices. When clients want us to liquid immersion in the US, we bring all those learnings with us from India."

Secret Weapon: Money, money, money

NTT, officially the Nippon Telegraph and Telephone Corporation, is undoubtedly a Japanese company. Founded as a state monopoly in August 1952 to take over the telecommunications system operated by AT&T during the occupation of Japan by US forces after World War II, the Japanese Ministry of Finance is still the largest shareholder in the telco business.

Adams says working with his bosses across the Pacific has been "probably


Malaysia, Selangor

ADAMS ON EXPANSION VS M&A

Despite being formed from a collection of acquisitions, NTT has been relatively quiet on the M&A front in recent years. There's been a number of big deals by investment firms either looking for a ready-made player to take over, or collections of smaller deals that have been combined into broader platforms.

During recent earnings calls, CEOs of other large, publicly-listed colo operators have generally been uninterested in making major buys, with little in the way of tuck-in deals by many of the established players.

Adams echoes that sentiment, suggesting multiples are "unnaturally high," with many companies overpaying because they're driven by ambitions to flip the company.

"Do I see any interesting M&As out there? Not right now. Have we looked at a couple in the last year? We have," he says. "But part of the advantage of being part of a large global company, that thinks in 30-year time frames, not one-quarter timeframes, is the ability to be very strategic."

"Do we have the capability, and will we do M&As? Yes. I just haven't seen it that makes a lot of sense. We're looking at very point source things to improve our portfolio, we're not just looking at adding EBITDA to increase our war chest so that we can flip the company and get a big payday. We're looking at being very strategic to fill specific gaps where we have barriers to entry in a market or a reason why we don't want to go build."

In the meantime, the company continues to build out in the existing major major markets, and is eyeing a number of smaller markets as demand for capacity continues unabated.

"We are a big hub and spoke believer. So we always want to make sure we have inventory in the large hub markets – Ashburn, Paris, Frankfurt, Tokyo," he says. "We're also investing in the smaller, up-and-coming markets, and looking at that [regional] Edge compute, because once we truly get 5G in place, I think the Edge markets are going to be critically important."

Adams defines his Edge as regional Tier 2 and 3 markets such as Milan, Atlanta, and Colorado. When it comes to serving hyperscalers at the Edge, he suggests customers are demanding capacity figures that would have been considered sizable in primary markets just a few years ago.

"I don't think in terms of 5MW or 10MW. 16MW is the absolute minimum. But pretty much everything we buy is 36MW or greater. And in large markets, we're looking at hundreds and hundreds of megawatts of capability and buying hundreds of acres of land."

On the divestment front, Adams says most of its facilities are "relatively new" and designed to be flexible to accommodate liquid cooling where customers might want to increase densities. Where deploying liquid isn't as feasible, the company is happy to switch those sites to offer lower-density retail colocation. Of the 150 or so data centers the company operates, he says there are five or six smaller older sites in Europe and India it is looking at exiting.

"Most of our facilities are a decent size, and we have them for the long run. I don't see retail going away; I see retail continuing, and so there's always room for smaller customers, and we just put them in those older data centers."

ADAMS ON THE NEW GPU CLOUD PLAYERS

While the long-standing public cloud players aren't going away any time and are all making major investments into AI infrastructure, there's a boom of new players. GPU-as-a-Service and AI cloud companies are rapidly investing in their own data center build-outs in all corners of the globe (see more, page 45).

While Adams says the company is well acquainted with established hyperscalers, enterprise companies, and a "fair amount" of retail, especially in Asia, NTT hasn't done a huge amount of business with this new school of AI cloud companies.

"We've been relatively cautious on who we do business with today, which is why some of the smaller or more emerging providers we don't have in our portfolio," he says. "I'm not sure I'd be comfortable with a brand new provider in the space taking on 100MW and custom buildings with a 15-year lease at this point. I think that sometimes you need a little history of the business before you start taking on large chunks of infrastructure like that."

Despite that historical caution, Adams says NTT is open to bringing more of these 'new hyperscalers' on board as they stabilize.

"A lot of people have the fear that some of these smaller providers are going to tank. I'm not convinced that's the case," he says. "We're seeing such strong demand for AI right now. We haven't done it thus far, but as this newer generation of players grows and stabilizes their business, they're becoming much more interesting. We've got a few of them already; we'd like to have all of them with our portfolio over time.

"We're in risk versus reward business. The more risk, the higher the price is for the client, and you just factor that potential risk in there. We're a highly diversified multi-billion dollar global business, and so I think we can take on some of that risk. We just have been very cautious about it and very risk averse thus far, and haven't taken a lot of it."

"I think that there's a space for everybody to play. We're in a long-term business where we signed 10-15-year contracts with clients. There are providers that specialize in working with some of these small emerging companies. I think they've done a really good job of managing that credit risk." ■

There's no working with the banks to get the money to build a business. It is a much more agile way of managing things when you just write a check out your own checkbook

the best experience of my 35-year career."

He says: "The team in Japan wants to do nothing but nurture and support the business, period. They've been incredible business partners and one of our secret weapons." He notes that he often gets asked about culture clashes and whether his superiors in Tokyo are controlling or difficult to work with, but says the opposite is true.

"They don't control us," he reveals. "I'm not on the phone every week asking for permission. I do an annual budget, I manage the business, and they want to see us flourish."

Today, the whole company has total revenues of more than $20 billion a quarter, along with a $10bn war chest from the mothership to build out data centers, which is a boon for Adams.

"We're self-funded by NTT, we're not beholden to any banks," he says. "We don't have to go ask permission from banks if we want to expand or make changes in the portfolio. That gives us long-term thinking and makes us very formidable in the marketplace."

That easy access to cash is paying dividends in the current AI gold rush. Adams says one of the big changes in recent years is hyperscalers wanting providers like NTT to do infrastructure build-out because time-to-market is so important right now.

"Historically, they've built out the infrastructure because they could save money on it," he says. "Now, they're growing so quickly that they want to have that done during the construction stages of the build – they want to have all the circuits put in place, the racks put in place, the liquid cooling in place.

"If you are beholden to a bank to make your financial decisions and to build your business, you have to have a complete design, all the building materials, all the parts list, all the pricing for all the equipment, and you have to take to the bank and then ask for permission and get the loan. The covenants that they have to deal with and the banking arrangements they have are not easy."

He adds: "I don't have to go through any of that. There's no working with the banks to get the money to build a business. It is a much more agile way of managing things versus my competitive set.

"It's much different when you just write a check out of your own checkbook." ■

MERCURY

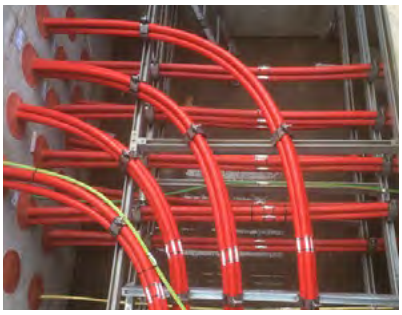
YOUR VISION, OUR DUTY

Mercury is the European leader in construction solutions.

For over 50 years, Mercury has helped the world's leading corporations deliver technologies and life-changing advancements that connect people, communities, and businesses.

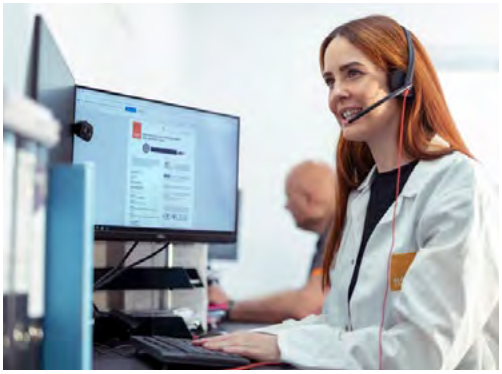
We go beyond the call of duty with a bold promise that we will always deliver, and it is this serious dedication that turns clients into partners, people into teams, and builds relationships that thrive.

To learn more visit **www.mercuryeng.com**



EUROPEAN DATA CENTRE POWER CABLE EXPERTS

ELAND® CABLES



POWERING A SUSTAINABLE FUTURE

CPD CERTIFIED
The CPD Certification Service

TALK TO US ABOUT CPD CABLE TRAINING

OUR ACCREDITATIONS AND COMMITMENTS

bsi ISO 9001 Quality Management Systems CERTIFIED FS 672069
bsi ISO 14001 Environmental Management CERTIFIED EMS 672067
bsi ISO 14064-1 Carbon Footprint Verification VERIFIED CFV 776862
bsi ISO 45001 Occupational Health and Safety Management CERTIFIED OHS 672066
bsi ISO 50001 Energy Management CERTIFIED ENMS 776260
IECEE CBTL Accredited Testing Laboratory
UKAS ISO/IEC 17025 Accredited Testing Laboratory
bsi Cable Verification KITEMARK™ KM 662609
FORS FLEET OPERATOR RECOGNITION SCHEME D:012590
SCIENCE BASED TARGETS BUSINESS AMBITION FOR 1.5°C
WE SUPPORT UN GLOBAL COMPACT
ecovadis

www.elandcables.com

The 6G future must learn from 5G's shortcomings

Paul Lipscombe
Telecoms Editor

The launch of the service isn't expected until the end of the decade, but work is well underway on 6G. Does the industry need it?

It wasn't that long ago that the telecoms industry saw 5G as the next big revolutionary technology.

"5G will have an impact similar to the introduction of electricity or the car, affecting entire economies and benefiting entire societies," former Qualcomm CEO Steve Mollenkopf said back in 2017.

Mollenkopf wasn't alone in making such a big claim about the technology. In 2018, former O2 chief operating officer Derek McManus made a similar statement, and there have been plenty of others within the telecommunications industry that hyped up the then-new generation of wireless.

Fast-forward to 2024, and GSMA figures reveal that more than 260 operators have launched 5G across more than 100 countries.

But with 5G now a mature technology in several markets, attention is turning to 6G, which will be the sixth generation of mobile connectivity.

The rate at which the topic is being talked about at telecoms industry events is growing, and while it may not yet have the buzz that accompanies discussions around AI, 6G is definitely on the agenda.

Managing expectations

Industry experts don't expect 6G to hit the airwaves until around the end of the decade, with many suggesting the first such commercial networks will launch around 2030.

At present, the technology is in the pre-standardization phase, with standards and features yet to be finalized.

The first 6G specification will appear in 3GPP's Release 21, which is set to be completed around Q4 2028. Following this, initial commercial deployments are expected to begin a year or so from then, and then on into the new decade.

For context, the telecoms industry is currently working on Release 19, which is 5G Advanced.

But vendors, including Nokia and Ericsson, are already looking to the future, and were researching what 6G might look like long before 5G was even launched.

Petter Vetter, president of Nokia Bell Labs Core Research, told *DCD* that the company's 6G research started more than five years ago. This was because, Vetter says, "you need to be at least 10 years ahead of the curve."

He explains: "The expectation is, every decade, there is a new generation of networks. So 6G is expected to come around 2030, an important reason to have a new generation is that you have new technologies, but you also have new requirements and new spectrum for new capacity."

At present, Nokia Bell Labs is in the phase of researching 6G, which it claims will "radically transform" what a network can do, unleashing new potential in the people and businesses that use these networks.

According to the company, 6G will "fuse the physical, digital, and human worlds," allowing for interaction between digital and physical realities.

Nokia's Nordic rival Ericsson has also been working on 6G. Magnus Frodigh, VP and head of Ericsson Research, tells DCD that Ericsson has been researching the technology since 2017.

Frodigh explains the current status of 6G development is very much a work in progress, as the industry works on standardizations of the technology.

He says: "We are already in this alignment phase between different

players, trying to see what they believe, what we believe, and see where we agree and to find some sort of consensus on what 6G should be."

Though we heard similar promises with 5G, Frodigh says he expects 6G to drive augmented reality use cases, along with digital twins.

Ericsson has already signed 6G deals with carriers, including a Memorandum of Understanding with UAE-based telco e& to explore 6G technology concepts.

Other big-name vendors such as Huawei and Samsung are also looking ahead to 6G.

During Huawei's Mobile Broadband Forum held in Istanbul, Turkey, in October, Turkcell's CEO suggested 5G Advanced, otherwise known as 5.5G, is an important part of the eventual transition to 6G.

Autonomous cars

"We need to be clear and very precise about this, 5G is not an upgrade," said Dr. Ali Taha Koç, the carrier's chief executive. "It's a new phase in which we will connect intelligence. With 5G we start to connect intelligence.

"The value of 5.5G technology is that it's a step towards 6G and the future of communication. So we need to use 5.5G in the ladder to 6G, while we must continue to invest in private technology, in coalition with our mobile and public institution partners."

An evolution of 5G?

While vendors have kicked off their research into 6G, delivering their early views on the technology, industry analysts are doing their best to make sense of what 6G will look like.

Chris Antlitz, principal analyst at Technology Business Research, expects to see an "evolution" of 5G.

"6G is shaping up to be an addendum to LTE and 5G, providing a new antenna overlay that supports net-new frequency bands, as well as enhanced spectral efficiency features and capabilities that provide further network performance and operational improvement," he suggests.

He adds: "The telecoms industry must also contend with supporting new use cases and how to embed AI, ML, and sustainability into the fabric of the network while covering security gaps and preparing for a post-quantum cryptography world.

"Though there is tremendous brainpower (spanning the public and private sectors as well as academia) assembled to tackle these issues, growth prospects for the telecom industry continue to look challenging."

Vetter agrees that AI adoption will be a big driver for 6G.

"The new technology that justifies a next-generation now is AI," explains Vetter. "Having an AI native approach is a very important target for 6G. The network needs to be more programmable and allow for more monetization."

5G has not been a success

To understand what 6G will look like, it's important to understand its predecessor. 5G has been available since 2018, but the race to launch was frantic and arguably rushed, and many question the technology's success.

"Some said that 5G would be a generation like no other. That has turned out to be true but not in the manner expected," wrote Professor William Webb in his book, The 6G Manifesto, which sets out to learn from what he views as the "disappointment" of 5G.

"The expectation was that 5G would lead to vast numbers of connected devices, to new 'metaverse-like' ways of communicating, to autonomous cars, and robotic surgeons; in short to a science-fiction world. Instead, 5G has led to increasingly cash-strapped MNOs and a skeptical population. The metaverse, autonomous cars, and robotic surgery remain far away."

Webb argues that 5G failed to identify a tangible solution to an existing problem, noting that 2G delivered capacity and security, before 3G added data, while 4G fixed the technical issues of 3G. He says the technology has been massively overhyped.

Instead, he suggests that "in the absence of clear needs for 5G," the industry "made some up."

Still, he's not alone in his views on 5G being a disappointment. Andy Hicks, senior principal analyst at research company Global Data, says vendors and carriers "oversold" the technology.

"They were pitching 5G as if it was some sort of amazing thing," Hicks says. "The network response is better, but what's the major 5G use case? It's fixed wireless. That's about as unexciting a use case from a new technology point of view as you can get."

Learning from 5G hiccups

When 5G was first launched, the technology was based on non-standalone architecture (NSA), meaning it was designed to be deployed on top of an existing 4G LTE network.

A few years later, 5G Standalone (5G SA) technology emerged, which was not reliant on older mobile generations and was based solely on a 5G core network. Many observers, and even telcos, refer to this as the "true 5G."

According to Antlitz, launching 5G NSA was a mistake.

"The industry is really trying to take the learning from the 5G NSA debacle, because it was not a good idea to do NSA," he explains.

"I was at an event recently and a lot of people were saying that NSA has done the industry a great disservice because it fails to live up to the promises of what 5G was supposed to be. It was a band-aid that didn't give the industry what it needed."

Antlitz noted that 5G NSA wasn't able to drive capabilities such as network slicing, touted as a key benefit of 5G. He argues that even in 2024, the majority of carriers still don't have a 5G network. "We cannot make that same mistake again with 6G," he says.

Ericsson's Frodigh agrees that the move to launch 5G NSA might not have been the right avenue for the technology.

"That's been a good learning," he says. "It was very unfortunate that we were forced into this. There were a lot of customers wanting to launch 5G before the core networks were updated, and then we got this NSA, and that was not a good move."

Yet, Nokia's Vetter maintains that 5G has been a success.

"I think NSA was a very good introduction scenario [for 5G], but it has slowed down the move to a 5G core," he explains.

Ericsson pushing 6G

Because of this, he adds, the mobile industry has not been able to leverage 5G Standalone as rapidly as hoped.

"The lesson learned with 6G is to make sure that evolution allows for an immediate use of the 6G capabilities," he says.

In his book, Webb says that despite the shortcomings of 5G, it won't stop 6G from emerging in the coming years. However, he says that there's plenty that can be learned from 5G.

He says he wants mobile carriers to provide a stronger voice in setting out what they require for 6G, which may require helping to educate politicians, he points out.

"The overriding lesson for 6G is to find solutions to problems, not the other way around," says Webb.

Are we ready for 6G?

While vendors and carriers will be keen on pushing the latest technology they have to offer, Hicks doesn't think it's necessary to rush 6G.

"I really hope that it's not going to be a repeat of the same thing where 2030 comes out with a new radio generation and nothing else, and then it's only sort of the middle of the 2030s where we actually start really deploying a 6G core," he explains.

Indeed, Hicks says he would rather the industry slowed down the rollout of 6G, in favor of making sure there are tangible use cases in place that can enable value for businesses.

Paul Rhodes, director of RAN at Edge data center firm AtlasEdge, is also cautious on the topic of 6G.

"Nobody has yet come up with a compelling use case for 6G," he says. "There are lots of things to happen with 5G and, at this point in time, there's no reason to be enthusiastic and excited for 6G."

Rhodes believes it is in the interest of vendors, rather than the carriers to push for a new technology.

"For Ericsson and Nokia, they want to push their products, and their product is hardware," he says. "So every 10 years, they want to push some new hardware, because that's how they make their money."

He believes the technology needed for 6G may be software-based, which could help carriers save some money compared to the costly hardware installations required by previous generations.

We must work towards 6G now

Many working groups have already been formed to investigate 6G and are enthusiastically pushing the technology's benefits.

One such group is the 6G Smart Networks and Services Industry Association (6G-IA), an organization that calls itself the voice of the European industry and research for next-generation networks and services.

The 6G-IA works alongside the European Commission on the Smart Networks and Services Joint Undertaking (SNS JU), which is the European Union's funding for 6G research. It's one of the largest sources of non-commercial funds for 6G research with a budget of at least €1.8 billion ($1.9bn).

The group aims to unite the telecoms and digital actors, such as operators, manufacturers, research institutes, universities, verticals, SMEs, and ICT associations. Its list of partners includes vendors Ericsson, Huawei, Nokia, and Samsung, plus telcos such as Deutsche Telekom and Telefónica, Telecom Italia, and Vodafone.

"5G is a really good network and it has delivered technically the promises that it had made. But if we want to do better during 6G, we have to take some considerations and some lessons learned from 5G," explains Dr. Alexandros Kaloxylos, executive director of 6G IA.

"One of these is that the architecture should not have too many options. It should not be too complicated, because if you go to standardize all these things, there is a lot of effort to standardize, and sometimes not all standardized features are used in operational networks. So the design of 6G networks needs to be made based on sensible decisions that will bring benefits to the end users and will offer clear monetized solutions for vendors and operators."

He notes that 6G will continue the industry's vision to be more sustainable, while AI will also support this, through the use of more intelligent networks.

"We believe that AI can be beneficial in two things," Kaloxylos says. "First of all, improved operation of the networks themselves, and at the same time offer AI as a service to service providers for vertical industries.

"5G has delivered considerable improvement on energy efficiency, and network performance but it has, so far, failed to keep its promise for engagement from vertical industries."

Who will be pushing for 6G?

Analysts expect network vendors to do a lot of the groundwork for 6G, given their interest in selling the hardware that 6G will require.

Carriers will also be expected to chip in, but telcos are fairly tight-lipped about pushing 6G just yet, justifiably so given 5G's room for growth with the launch of more 5G SA networks.

Antlitz expects the level of capex investment from carriers to be more subdued with 6G compared to 5G.

By the end of last year, *Analysys Mason* reported that mobile operators had globally invested more than $600 billion in cumulative capex on 5G networks. Similar numbers for 6G investment are tricky to find, but *GSMA Intelligence* estimates that between 2023 and 2030, carrier capex will hit $1.5 trillion.

"The telecom industry continues to struggle with realizing new revenue and deriving return on investment from 5G, even after five years of market development," says Antlitz. "TBR continues to see no solution to this persistent challenge and with no catalyst on the horizon to change the situation, communication service providers' (CSPs) appetite for and scope of investment in 6G will likely be limited."

Are carriers cautious on 6G?

One carrier that is proactively looking ahead to 6G is SK Telecom. The South Korean firm is currently carrying out R&D activities for 6G and is working with Nokia, NTT, and NTT Docomo on the development of a 6G AI-native air interface which it hopes will help advance AI telco infrastructure.

That said, the carrier is cautious about 6G, noting that it's crucial to develop

tangible use cases to push beyond what 5G could offer.

"Unlike previous generations, 6G is expected to present both opportunities and challenges for MNOs," says Minsoo Na, director and head of 6G R&D at SK Telecom. "Earlier generations of mobile communications were able to generate revenue by efficiently handling the surge in data traffic with the help of advanced wireless transmission technology.

"However, the technology to efficiently manage this increasing data traffic has already reached maturity with 5G. Therefore, the incentive to evolve to 6G, overlaid on 5G infrastructure, may not be substantial if it relies solely on revenue from data traffic."

Na says that SK Telecom is focusing on new usage scenarios "on top of what was defined in 5G for 6G's deployment."

According to Na, the capex challenge of developing a new generation of mobile network is a significant challenge for mobile operators. He says that AI will help carriers monetize 6G.

"To turn these challenges into opportunities, there is a need to enhance service intelligence and diversify revenue streams through collaborations with industries outside of telecommunications," he says.

"The integration of AI into telecommunications could be one significant method to achieve this."

As for potential use cases, Na suggests it will include autonomous driving and robots, plus a meeting of the real world with the digital world through the metaverse, XR, and digital twins. It's worth remembering these same use cases were thrown about by carriers prior to 5G's launch.

Governments might just run the 6G race

Where vendors and networks have previously been the ones to get excited about next-generation networks, now governments are also taking a keen interest in how technology can drive future innovation and generate billions of dollars for economies around the world.

An example of this is when the UK government revealed it would *award £28 million ($35.7m) to three UK universities* to help design and develop 6G network technology with selected vendors.

With this in mind, the race for 6G may be more political, argues Antlitz. "6G will happen, and the reason it's going to happen is because the major governments of the world have designated 6G a technology of national and societal importance," he says.

"TBR expects the level of government involvement in the cellular networks domain (via stimulus, R&D support, purchases of 6G solutions, and other market-influencing mechanisms) to significantly increase and broaden, as 6G has been shortlisted as a technology of national strategic importance."

He says that China's position on 5G is an example of the government heavily pushing mobile technology, and argues its position has forced the West to invest heavily as well, impacting telcos along the way.

"We've seen unprecedented government involvement in certain technologies, including but not limited to quantum computing, Open RAN, plus new energy technologies that are deemed to be of national security interest, like small modular nuclear reactors," Antlitz says.

He adds that while the economics might not always seem to make sense, the societal and national imperatives for investing in those technologies are significant. Because of this, Antlitz expects that the public sector will take the lead on seeding and facilitating market development versus the private sector for most of these strategic technologies.

For the many

While opinions on what 6G networks will look like remain varied, the feeling that the technology will arrive by 2030 appears almost universal.

Nokia Bell Labs' Vetter argues: "Every 10 years or so, you buy a new car, and you're not going to buy a smelly old diesel, you'll go for an electric vehicle or something that has the latest technology. It's the same with technology."

It's this technology that Vetter says will bring more efficient mobile networks for the carriers.

If 5G hasn't been a success - and to many, including Webb, it hasn't - then 6G needs to deliver something more.

"We have not benefited from 5G, despite being told that it will transform our lives, and we have lost faith that the existing companies and methodologies will deliver what we want rather than what some researchers believe we need," Webb writes.

"We want our voice to drive 6G in a direction that makes the world a better place. "6G should be for the many, not the few."

WHAT SPECTRUM WILL 6G USE?

Governments will also play a role in overseeing spectrum allocation, working with regulators such as the FCC and Ofcom to auction off the spectrum to carriers, telcos, and other organizations.

In his book, Webb says that the mobile industry needs to stick to frequency bands below 2GHz. Antlitz, however, anticipates that the industry may go in a different direction.

"After an initial belief several years ago that 6G would leverage millimeter wave and terahertz spectrum, the wireless technology ecosystem has settled on the upper mid bands, specifically in the 7GHz-24GHz range (also known as the Frequency Range 3 [FR3] tranche)," he says, noting that within FR3, 7GHz to 15GHz could be the "golden range for 6G."

He expects 6G will utilize a mix of spectrum tranches, with midband, upper midband, and mmWave frequencies all in play.

Earlier this year, during Mobile World Congress, during a keynote event, Doug Kirkpatrick, president and CEO at radio unit maker Eridan, argued that the best course of action should be for the existing spectrum to be repurposed.

"From a technology perspective, we don't need any more spectrum," said Kirkpatrick. "The challenge is not spectrum, the challenge is pushing the technologists to push the pieces that you need, we have probably spectrum in the wrong places.

"We need to think more about what kind of spectrum we want to use for what kind of applications. Moving applications from one spectrum to another is not a zero-sum game; you're going to have a transit period in between. So we will likely need some intermediate spectrum in order to make those movements happen, but when we're done, the answer should be zero."

Ehvert
MISSION CRITICAL
A Salas O'Brien Company

ADVANCED MEP ENGINEERING

ww.ehvert.com

The forgotten workforce:

Suicide in construction

Niva Yadav
Junior Reporter

The growth of data centers relies on a long-suffering workforce

Data center growth is reaching ever loftier new heights.

New capacities and new markets are being launched and targeted every day. But this growth relies on a forgotten industry that has failed to evolve and is saturated with age-old working practices, gender stereotypes, and poor workforce treatment.

That is the construction industry; the sector at the bottom of the supply chain that is satisfying every hyperscaler's building frenzy. What may be some of the greatest developments and milestones for hyperscalers and data center operators, is now taking its toll on construction workers.

The construction industry has devastating rates of suicide, which stand at four times the national average in both the US and the UK. US public health organization CDC (Centers for Disease Control and Prevention) said in a *recent report* that 53.3 out of every 100,000 workers commit suicide. In comparison, the average rate of suicide in the US is just 12.93 per 100,000 people.

On The Tools, an online community platform for tradespeople, said that 93 percent of UK tradespeople had been affected by mental health at some point in their career, with 73 percent still being affected.

Despite undergoing around four industrial revolutions, the construction industry is now well overdue for another overhaul for the benefit of its long-suffering workforce.

Wash, rinse, repeat

We all live in the built environment. Construction has an impact on us all. When it comes to data centers, however, they are now "at the tip of the spear when it comes to the fifth industrial revolution," says Nancy Novak, chief innovation officer at US operator Compass Datacenters.

Compass Datacenters has 19 data centers, both operational and under construction, across eight markets in three countries. And, like everyone else in the industry, has been under pressure to build out and expand as the world grows hungrier for compute, artificial intelligence, and cloud computing.

It is not a coincidence that both Novak and Alan Blanchett, group SHEQ director at McLaren, used the term 'push, push, push' to describe the pressures of productivity placed on the shoulders of construction workers. McLaren, based in the UK and the UAE, builds data center projects for hyperscalers, colocation firms, and enterprise customers that average about 200MW.

"The data center industry has this way of going wash, rinse, repeat, get better," says Novak. Ironically, at the bottom of the supply chain, the construction industry has failed to improve to meet the needs of its workforce.

That being said, Sam Downie, managing director at mental health charity Mates in Mind, explains that construction has now addressed physical safety, with the risk of physical accident more or less 'grappled with.' Blanchett points out: "We tend to shout safety and whisper health." McLaren is now actively incorporating mental health initiatives into its health and safety program.

Finally, mental health is now becoming considered as important as physical health. And it's about time. Suicide is the biggest killer in the industry, leading to more deaths than falls from height.

The perfect storm

The construction industry is "the perfect storm" for poor mental health, says Charlotte Brumpton-Childs, national officer for construction and engineering at UK workers' union GMB. She explains the sector is rife with job instability, unpredictability, and working far away from home. These contributing factors have now become intrinsically linked with the construction profession.

Rachel Neal, VP of global safety at Compass Datacenters, adds that the average worker in the US is contracted to 12 sites a year.

Whilst these are factors that could be prevalent in any profession, these triggers are exacerbated in the construction industry and those pressures often mount on individuals towards the end of the supply chain, says Downie.

Like any other industry, things go wrong. However, the sector is poorly equipped to combat these adversities. For instance, in the case of an injury, a worker is often left choosing between

paying their mortgage and taking time off, because of lack of proper sick pay and looming job insecurity, says Brumpton-Childs.

Charities such as Band of Builders have tried to alleviate some of the pressures caused by injury. Gavin Crane, CEO of the UK-based charity, explains the organization was initially set up to provide tradespeople with practical help following injury. This could mean building a stair lift, finishing housing extensions, or general house maintenance. Following Covid-19, the charity is now expanding its offerings to mental health support and support helplines, and in-person groups.

A chicken and egg conversation

The construction workforce in both the UK and the US is an aging one, with a third of the workforce in the UK being over 40 years old. Crane says in the next decade, 500,000 workers are set to retire, leaving gaps to plug and fill.

However, the aging workforce could also explain some of the macho and outdated ideals within the sector. Blanchett explains that, when he joined the industry, speaking about emotions and feelings was unheard of, particularly in a male-dominated sector like construction. "They'd say just get on with it, or 'man up,'" he says.

A recent and scathing article by right-wing British paper *The Daily Mail* accused builders of becoming "woke, sensitive souls more likely to enjoy yoga, muesli, listening to Radio 4, and sharing their feelings." Such sentiments have prevented the sector from addressing its largest and most silent killer.

Brumpton-Childs asks: "Is it a macho environment because it is dominated by men, or is it so dominated by men because it is a macho environment?"

Downie explains that dismantling a male-dominated environment is not just about 'landing women in a sector and making it better.' Experience has shown adding women into a workforce would in reality suppress women into conforming to stereotypical ideals. On top of that, women are not exempt from the pressures of construction; women in construction are 28 percent more likely to experience poor mental health, according to research by On The Tools.

At Compass Datacenters, Neal and

> *The data center industry has this way of going wash, rinse, repeat, get better" - Nancy Novak, chief innovation officer at Compass Datacenters*

Novak say that dismantling stereotypes and masculine ideals begins with language. The company has changed the language used to address its workforce, replacing 'foremen' for the gender-neutral term 'frontliner.' "It doesn't matter who you are or where you come from, if you're a frontliner, you're a frontliner,'" says Neal. Frontliner encompasses all middle-management positions and supervisors. On The Tools research has said in the UK, tradespeople in those middle-management roles are the most susceptible to poor mental health.

But, it is more than just a fancy title. Compass' initiative also equips its workforce with the language to discuss emotions and sentiments that have previously been banished by the industry. Neal explains the program encourages 'humility' and 'vulnerability' through dedicated workshops, so that frontliners not only know how to share complex feelings, but how to respond to their colleagues.

As Blanchett points out: "We spend most of our time at work," and so interactions amongst the workforce can have a large impact on our mental health. On the flipside, he explains, that it is our colleagues who have the best chance at noticing and identifying changes in behavior and who are best equipped to start conversations around mental health. He says at McLaren staff are being encouraged to check in on one another.

"The aim is not to make everyone a counselor," he says, but to spread awareness so that people at risk can be referred to professionals and helped at the earliest stage. McLaren still offers formal mental health training to its workforce and an all-over health check which includes health, hygiene, and mental well-being.

Out with the beer, in with the brew

"We're not reinventing the wheel," says Crane, speaking on the Band of Builders' 'Big Brew' initiative. Based on the Macmillan coffee morning, Big Brew creates an environment for tradespeople to talk, socialize, and be vulnerable. Crane explains the pub was once a popular space to 'get things off your chest' and decompress. However, with alcoholism amongst tradespeople on the rise and pub culture decreasing in popularity, there is a need to create another space. Brumpton-Childs adds that pub culture often fosters unhealthy relationships with addiction and creates alcohol dependence.

Mobile tradespeople working on sites away from home often have access to welfare facilities. But, as Crane points out, these facilities are inconsistent in the services they offer. Brumpton-Childs explains that workers could be walking more than half an hour to reach facilities, cutting into their respite. At McLaren, Blanchett says there has been an active effort to make these facilities a real place of relaxation and decompression, removing posters about 'falls' and health scares, and encouraging non-work related conversation.

Starting and managing the conversation is one thing, but there is an increasing need for government intervention, says Crane. He says that given construction contributes nine percent of GDP in the UK, it seems bizarre that there is no dedicated government department for the industry. Government intervention would also help to ensure mental health initiatives are successfully passed down the supply chain.

Without construction, the economy in any country is in danger of not growing. "Construction impacts us all. We all want homes, we all want hospitals, we all want construction." It's high time the forgotten industry was remembered.

If you are struggling with mental health, you can call the National Suicide Prevention Helpline UK on 0800 689 5652 or Samaritans on 116 123. In the US, the 998 Suicide & Crisis Lifeline can be reached by dialing 998.

COMMSCOPE®

Think Fiber

Think CommScope

Supporting applications:

- Artificial intelligence (AI/ML)
- Cloud computing
- Augmented reality (AR)
- Industry 4.0
- 5G cellular networks
- Data Center Infrastructure Management (DCIM)

Accelerate Your Data Center Connectivity for AI

Cabling considerations can help save cost, power and installation time:

Quicker installation and uptime

Sustainable and future-proof

Global reach and scale

Scan below to learn more or visit
commscope.com/insights/unlocking-the-future-of-ai-networks

A (small) nuclear revolution?

Zachary Skidmore
Senior Reporter, Energy and Sustainability

Hyperscale partnerships could be the catalyst for small modular reactor development

The data center sector is walking a tightrope between ever-increasing power demands and its commitment to decarbonizing. Building out and securing low-carbon power sources is becoming increasingly imperative for the sector, fuelling a renewed interest in nuclear energy, especially small modular reactors (SMRs).

For proponents, SMRs offer the perfect solution to data centers' needs, namely, consistent low-carbon baseload power. They also provide high integration potential due to their modular nature and ability to be deployed quickly in various locations, independent of any external power sources or grid connections.

However, significant concerns remain over whether the technology can be successfully scaled for use. Critics would point to the fact that SMRs remain firmly in the demonstration phase, and with countless proposals out there, it remains to be seen which of the SMR developers - if any - will produce a commercially viable product.

In 2024, several data center operators took the plunge and partnered with SMR vendors, begging the question: Are we on the brink of a nuclear revolution in the data center sector?

What are SMRs?

The International Atomic Energy Agency defines SMRs as small power reactors with lower outputs ranging from less than (up to) 10MW, known as microreactors, to a standardized capacity of 300MW.

SMRs are designed to be portable and can be shop-fabricated and transported as modules, allowing for on-site installation. Their smaller footprints and flexible deployment make them suitable for regional or industrial clusters. SMRs are designed to operate for long periods of time before refueling, with some lasting up to 30 years.

There are varying novel reactor concepts. The initial generation I and II reactors were developed by the US military in the 1950s. Of current designs, Gen III pressurized water reactors are the most common, and operate as miniaturized traditional nuclear plants.

Recently there has been a proliferation of Gen IV concepts, which hold promise of much higher efficiency through alternative cooling methods, including gas-cooled, liquid metal-cooled, and molten salt designs. However, these concepts have little to no real-world industrial experience.

Hyperscalers on board!

The commercialization of SMRs will depend on overcoming significant hurdles, including financial challenges. Construction timelines proposed for most SMRs span the mid-2030s, with considerable delivery uncertainty and cash flow risks deterring traditional debt financing. As a result, SMRs have historically relied on government funding, with the US, UK, and Canada all launching funding rounds in recent years to support domestic SMR development.

February 2024 saw the first privately funded SMR agreement between Westinghouse and Community Nuclear Power to deploy four SMRs in North Teesside, UK. This milestone was followed closely by a series of commitments from the hyperscalers, with Google, AWS, and Oracle all inking long-term agreements with SMR providers to power their operations.

For Ivan Pavlovic, executive director, energy transition at investment bank Natixis, these agreements could support the sector similarly to how renewables were backed through state subsidies.

"Just as renewable energy benefited from feed-in tariffs and green certificates, SMRs may rely on private contracts with large off-takers to support early development, mimicking the conditions of renewable financing," Pavlovic says.

The hyperscaler agreements all represent long-term commitments to the sector into the 2040s. Google, for example, has penned a 20-year power purchase agreement with Kairos Power, a molten salt-cooled Gen IV SMR.

The long-term commitment was crucial for Kairos, said Mike Laufer, the firm's CEO, as it met the two "main challenges" for financing nuclear projects - "the long time scales involved and the need for financial backing to cover the development period, even under an aggressive timeline like 2030–2035 for our initial deployment."

The commitment of these companies marks a significant step toward the commercialization of SMRs. Acting as early adopters, and in AWS's case, investing directly into the SMR firm in the form of X-Energy, developers are provided with the financial stability and long-term agreements necessary to substantially de-risk SMR projects, which could pave the way for broader market acceptance. The agreements also serve to temper concerns about rushed deployment, which can inflate costs and cause delays.

A mutually beneficial arrangement

For Pavlovic, data centers are likely to be the "best possible off-takers" for nuclear energy. "They need low-carbon, 24/7 electricity, and strong financial balance sheets to support long-term contracts," he says.

SMRs have exceptionally high capacity factors, which is a measure of how often a power plant operates at maximum power, and how consistently it produces energy over time. This is shown in both Nuscale's and Rolls Royce's SMRs, which have both registered a capacity factor of 95 percent or more, increasing their attractiveness to data centers. In addition, with most projects ranging between 60MW and 300MW, they offer large amounts of clean, consistent power, free of issues of intermittency and curtailment, as seen in solar and wind.

> *"SMRs can be placed almost anywhere and offer an exceptional capacity factor, a key measure of energy consistency, which surpasses even that of gas or coal"*
>
> *>>James Walker, CEO of Nano Nuclear*

"SMRs can be placed almost anywhere and offer an exceptional capacity factor, a key measure of energy consistency, which surpasses even that of gas or coal," states James Walker, CEO of microreactor company Nano Nuclear. These attributes explain why the tech giants have begun to view nuclear as the preferred solution for reliable, low-carbon energy.

The relationship is likely to be mutually beneficial, with data center developers being one of the few industries with the capital and forward-thinking to take a risk on a yet-to-be-proven technology. Kairos Power's Laufer emphasizes that partnerships with hyperscalers not only provide financial security, but also facilitate iterative learning and cost reductions. "The partnership with Google provides a strong alignment for both parties, enabling cost reductions and learning through the deployment of multiple reactors of the same kind," he says.

This allows companies such as Kairos to build "something that is either exactly or very close" to their solution as part of a demonstration project, as seen in its Hermes Demonstration Reactor. Laufer adds that the more measured approach ensures a "true learner effect" that reduces costs before entering the capital-intensive construction phase. This method contrasts with traditional nuclear, which often bypasses smaller-scale demonstrations, leading to cost overruns and delays. In turn, this has prevented many companies from backing SMRs.

Rolls Royce SMR has also adopted this measured approach. According to Harry Keeling, the company's head of development of new markets, Rolls Royce's "approach gives customers certainty that when we commit to timelines, we will be able to deliver, and this creates trust among investors."

Subsequently, Keeling has contended that "in the next ten years, we are likely to see a consolidation around a few leading SMR technologies akin to Boeing and Airbus in aviation."

This, in turn, could support the industry in achieving the fleet-level economies of volume necessary to serve the data center market. In doing so, SMR developers hope to meet one of their biggest challenges: creating cost certainty.

Cost certainty through modularity

The path to cost certainty involves careful, measured steps. Data center commitments have provided developers the flexibility to avoid hastened deployment. However, concerns remain over whether SMRs will be financially viable for widespread use.

Harnessing SMR modularity will play a central role in creating the cost certainty required. The modularity of SMRs compares favorably to renewable energy, especially solar, as they can be built in-house off-site and shipped as modules. Moving away from the single monolith model of traditional nuclear to the many small modules model will mean that economies of scale can be achieved during the component manufacturing process, reducing costs and easy scalability.

Keeling argues that modularity makes SMRs much more attractive to the financial community, as it "de-risk projects." Therefore, unlike traditional nuclear, where 50 percent of the energy cost comes from debt, "modularity allows a reduction of on-site construction risks, streamlines operations, significantly shortens project timelines, making nuclear power accessible to a wider range of customers."

Rolls-Royce SMR has embraced this approach, constructing its entire power plant using standardized modular pieces. "Every plant uses the same 1,000 modular pieces, ensuring standardization and volume economies," he says.

Modularity not only reduces costs but also aligns well with data centers' scalability needs. Clayton Scott, chief

commercial officer at NuScale, another SMR developer, highlighted the appeal: "This modular approach provides data center operators with greater options when selecting the right size power plant to meet capacity and economic considerations."

In addition, as they are not location-dependent like many other clean energy systems, modular reactor deployment flexibility is much higher, making them more suited for scaling in Edge and remote installations that may lack good grid connections or clean energy access.

This has led some companies to focus explicitly on the data center market. Start-up Deep Atomic is doing that, seeking to offer a 60MW "behind-the-grid island power solution for data centers," according to a company representative. The small size will support greater deployment flexibility, with potential hybridization with energy storage systems and renewable to support the data center operator's drive to net zero.

In addition, by streamlining factory-built components and minimizing on-site risks, SMRs can control costs and timelines more effectively than traditional nuclear projects. Keeling emphasizes that from financing to licensing to operation, "standardization reduces uncertainty at every stage, making SMRs far more commercially viable than traditional nuclear projects."

For data centers, this provides a compelling case for adoption, as SMRs can deliver consistent, low-carbon energy tailored to their needs.

Challenges in financing and regulation

Despite growing private sector interest, SMRs face significant hurdles in achieving cost certainty and delivery reliability.

A report by Germany's Federal Office for the Safety of Nuclear Waste Management highlights the steep construction costs associated with SMRs. BASE found that achieving the same global output as today's large-scale nuclear plants would necessitate scaling SMR deployment by three to 1,000. This translates to constructing approximately 3,000 SMRs globally to make their production economically viable.

> *"In the next ten years, we are likely to see a consolidation around a few leading SMR technologies akin to Boeing and Airbus in aviation"*
>
> *>>Harry Keeling, Rolls Royce SMR*

Critics have also argued that several SMRs have sold the market on inflated unit economics while grossly underestimating the time and capital it will take to commercialize their products.

In a recent report, Kerrisdale Capital claimed that SMR developer Oklo, which has signed several supply agreements within the data center sector including Equinix and Prometheus Hyperscale, is beyond optimistic in its timelines.

The firm is working towards submitting a license application in 2025, hoping for a first reactor deployment by late 2027. However, according to a former NRC Commissioner, its near-term projections are steeped in "hubris," as the company lacks the long-term supply of enriched uranium required for its reactor technology.

Given that Oklo has already signed several agreements with data center firms, the accusation that it is unlikely to meet its timelines weighs heavily, as it may impact the overall confidence that SMRs can be successfully commercialized.

Both Nano Nuclear and NuScale have also stoked controversy, with both firms facing charges from short seller Hunterbrook Media that their timelines are unrealistic and their products may not be able to live up to lofty claims. NuScale canceled a project in 2023, citing a lack of demand, which provoked further concern throughout the sector.

In addition, with most SMRs still in the concept exploration phase and more than 80 designs under development globally, there is still fundamental and inherent uncertainty about what designs can acquire regulatory approval and subsequently scale their solution to a commercial market.

However, SMR developers such as Rolls-Royce argue that the trick is to center product development around the regulation process.

"We've spent significant time with regulators in all our target countries. The feedback we receive is consistent: our reactor is 'boring,' and in nuclear, boring is the highest compliment," notes Keeling.

Developers also point to crossover between regulatory frameworks, which could expedite market licensing in new jurisdictions. This is more apparent with Gen III reactors, as they are already well understood. Gen IV reactors, on the other hand, are still highly experimental, and that may subsequently impact their timelines.

"While US regulatory approvals provide a strong foundation, each country has unique regulatory environments that require careful navigation for international market penetration," said Laufer.

Consequently, while challenges remain - particularly around cost certainty and regulatory hurdles - the SMR sector is positioned to benefit from its alignment with data centers' energy needs and modular design.

Partnerships between the data center sector and SMR developers could catalyze a new era of reliable, scalable, and low-carbon nuclear energy, addressing the data center sector's power demands and supporting the sustainable development of SMR technologies.

GLOBAL COMMISSIONING

DATA CENTRE COMMISSIONING SPECIALISTS

The Industries Leading Data Centre Commissioning Provider

UNLOCKING POTENTIAL

EMPOWERING SUCCESS

SERVICES

ELECTRICAL SERVICES

- Electrical Services: Ranging from HV SAP to 400kV, LV APs, and Electrical Safe Systems.
- Cloud Solutions: Including permit-to-work systems and commissioning software.
- System Training, Familiarisation, and Technical Authoring

PROJECT MANAGEMENT

- On-site Project Management, Support & QAQC Compliance
- Commissioning Programme Creation & Information Management
- Construction Management

COMMISSIONING MANAGMENT

- Comprehensive Commissioning: From concept to handover.
- MEP Validation Services
- Design Reviews & Commission ability Studies Test Script Production

CONTACT US

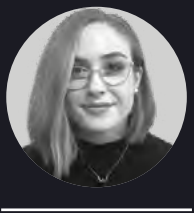
+44 (0) 1227 649087

admin@global-cxm.com

Radio House, John Wilson Business Park, Whitstable, Kent, CT5 3QP, UK

Global-cxm.com

Frosty with a chance of GPUs



Georgia Butler
Senior Reporter
Cloud & Hybrid

The compute behind our weather forecast system

"Earlier on today, apparently, a woman rang the BBC and said she heard there was a hurricane on the way. Well, if you're watching, don't worry, there isn't!"

These were the iconic words of meteorologist and weather forecaster Michael Fish on October 15, 1987. That night, the worst storm to hit South East England for three centuries crashed into the county, and 19 people died.

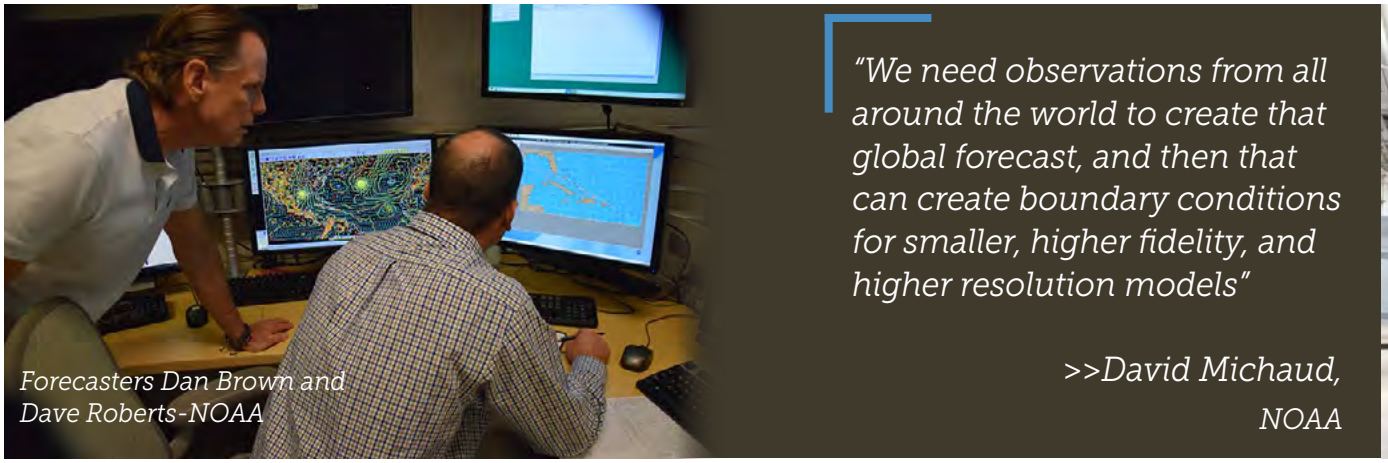
Whether Fish misspoke (he later stated he was referring to a hurricane in Florida at the time), or was making a poorly timed joke, the incident led to the coining of "the Michael Fish moment" - a forecast on any topic that turns out to be completely, and horrifically, wrong.

While Fish would presumably wish he could suck those words back up and into oblivion, never to have left his lips, the reality is that unless had superpowers, the responsibility of weather forecasting did not solely lie on his shoulders. In fact, the weather gods of today are supercomputers, whirring machines handling quadrillions of calculations a second. These numbers are set to grow further as more compute power is added, giving meteorologists a more detailed picture than ever before.

Human meets machine

Prior to our reliance on supercomputers, responsibility for weather observation and forecasting lay with human 'computers.' As depicted by Andrew Blum in his 2018 book The Weather Machine, weather reporting in the US during the mid-1800s relied on the newly installed 2,100 plus miles of telegraph lines.

Telegraph operators in various cities would communicate with one another what the weather was like in their respective locations. Because the lines worked poorly in the rain, if communication was unavailable, there was a reasonable expectation that



Forecasters Dan Brown and Dave Roberts-NOAA

"We need observations from all around the world to create that global forecast, and then that can create boundary conditions for smaller, higher fidelity, and higher resolution models"

>>David Michaud, NOAA

somewhere along that line it was raining.

Though a rudimentary approach to weather forecasting, this limited data could provide a basis for assumptions about storms that may or may not be approaching.

By the late 1800s, Norwegian physicist Vilhelm Bjerknes had begun to realize that there was also a mathematical basis for the way weather behaved.

His first hypothesis argued that, in circumstances where pressure and densities are unequal, they will battle against one another until they become more balanced. Dubbed 'Circulation Theorem,' the hypothesis could, in theory, predict the direction and intensity of that circulation.

Once tested on meteorological data, Bjerknes' idea proved to have legs, and he refined it further. Bjerknes has been credited with the founding and formulating of many of the equations that are still used today in numerical weather prediction and climate modeling.

These calculations were still made by human computers for a long time, and the sheer complexity of the parameters meant that truly meaningful forecasting was limited - predictions often took longer to formulate than the weather did to emerge.

In 1922, English mathematician and physicist Lewis Fry Richardson estimated that a global workforce of 64,000 human computers would be necessary for a global forecast. As he wrote at the time: "The scheme is complicated because the atmosphere is complicated."

It wasn't until 1950 that computers as we think of them today found their

role. The *Electronic Numerical Integrator and Computer* (ENIAC), the first programmable computer, was used for the first forecasts. ENIAC was made up of 18,000 vacuum tubes, and 1,500 relays, as well as hundreds of thousands of resistors, capacity, and inductors. In total, ENIAC was capable of around 5,000 calculations per second. The first calculations for a 24-hour forecast on ENIAC took the machine almost 24 hours to produce - not exactly practical.

As computers have become more powerful, forecasting has become more accurate.

The global scope of the problem

It is easy to think of weather as just what is outside of your window at any given time, but the reality, says David Michaud, is much more complicated.

"You can almost think of the atmosphere as a fluid - it needs to be continual," explains Michaud, director of the Office of Central Processing at the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service.

"You have to view the forecast in a global context. We need observations from all around the world to create that global forecast, and that can then create boundary conditions for smaller, higher fidelity, and higher resolution models."

As Michaud puts it, if the weather comes from the East Coast and moves to the West, "there isn't just a void behind it." Even if you are just looking for a regional weather forecast, you still need to be able to fill it in from a global scale model.

To work effectively within these conditions, meteorological services across the world collaborate and share observations.

"We have agreements and standards around the way that we share data and format data. So each government agency or entity around the world has its own platforms, and then all that data is shared," says Michaud. "All those observations are constantly flowing and we have sharing agreements so that we have this continuous set of information coming in."

This was reiterated by Alan Hally, scientific lead of the AI transformation team at Met

NOAA Dogwood. Credit: GDIT

Éireann, Ireland's meteorological institution. He notes that national meteorological services share data via a global transmission system where observations are collected in different global centers, and then distributed to all the regional offices.

"Recently, we had the tail end of Hurricane Kirk that transitioned across the Atlantic and impacted parts of Europe," Hally says. "When those things occur, we have conversations directly with the National Weather Service, specifically the National Hurricane Center in the US, because they have responsibility for those types of weather events.

"Then, as it was going to impact France and possibly the UK, it would have been a collective call between all of those countries to discuss where the hurricane is most likely to move."

In the case of an approaching hurricane, the conversations would surround the simulations created using Numerical Weather Models - the models that underpin our understanding of the atmosphere. However, for those simulations to be created and interpreted, data must first be collected.

Observing the atmosphere

Weather observations are made from a variety of platforms and instruments, Michaud tells *DCD*, and those methods cover every layer of the atmosphere - from the ground, to in the air, to out of the air in space.

Satellites - both GEO and LEO - contribute significantly to the gathering of weather data. "There are various instruments that exist on each satellite platform that can look at different things - different wavelengths and with different parameters," he says. "It's not just a camera that looks at the clouds. You can actually profile through a vertical in the atmosphere and get information at different layers with a satellite."

Beyond satellites, observations are also gathered from weather balloons, which have been used for the past 150 years and are still released twice a day, every day, from around 900 locations globally. The balloons reach heights of around 20 miles - twice that of airplanes - and sensors on those balloons measure elements such as temperature, humidity, wind, and atmospheric pressure.

"We even have sensors from airplanes, so as a plane takes off, they're taking weather observations through the atmosphere," adds Michaud. Observations are also taken from ships.

Then there are radars "looking up from the surface," Michaud says, adding: "All of that is brought in, and then you have to quality control it because now you have to make that initial condition a continuous state.

"If you have one bad sensor out there, that's reporting a temperature of 500 degrees or something like that, you need a way to figure out how to throw out that observation so that it doesn't create these hotspots and this discontinuous environment."

The scale of the challenge facing Michaud and his colleagues around the world is hard to take in. At its most fundamental level, calculating the weather is a computational fluid dynamics (CFD) problem, but one that covers the entire globe.

The atmosphere has to be broken down into three-dimensional segments, with each segment impacted by those surrounding it. Nothing can be viewed in isolation.

Numerical weather models typically run on either spectral spacing or grid spacing, explains Michaud.

In the case of grid spacing - which is perhaps more simple to understand - values have to be associated with each grid point, but that is not always possible. Planes and ships move, and not every surface has an observation at that point.

"You have to find ways to extrapolate all that. There can also be sparse areas where you don't have a lot of observations, and you need to be able to fill that in. So to create that initial condition, you take all those observations in, and then you can look at a previous forecast and then fit the observations to that forecast," says Michaud.

"You quality control against other observations and previous forecasts to get a continuous atmosphere as your initial starting condition."

From the initial starting condition, meteorologists move incrementally forward in time through the physics-based equations that predict the weather.

"One thing we do with our models is take an initial condition and perturb it in many different ways, and then run the same forecast out from different initial conditions," Michaud says. "We call that a model ensemble. If you look at all the models and they are tightly coupled, you have pretty high confidence, but then the more they diverge, the lower the confidence," he explains.

But, Michaud acknowledges that "ensemble modeling is another way to chew up computing cycles."

The direct correlation between more compute and better forecasts

National weather services are continuously expanding their compute capacity to better serve weather forecasting.

NOAA completed an upgrade of its Weather and Climate Operational Supercomputing System (WCOSS) in the summer of 2023, giving it twin supercomputers each with 14.5 petaflops of computing capacity, an increase of 20 percent compared to the previous solution.

Met Éireann also recently launched a new supercomputer in collaboration with Denmark, Iceland, and the Netherlands dubbed "*UWC-West.*" The supercomputer is housed in the Icelandic Met Office data center and, according to Hally, the collaborative approach delivers "more bang for our buck."

"By pooling the resources of each country, we were able to purchase a supercomputer that was more powerful

than any one country could on its own," he says. "It's also a pooling of expertise. So each of the four institutions has expertise in different areas of numerical weather modeling, so we can bring all of that expertise together."

Hally adds that housing the computer in Iceland, with its cool temperatures and abundant clean energy, means it is carbon neutral, and runs on 100 percent renewable power.

"Better" forecasts are provided in a variety of ways.

The new UWC-West supercomputer pales in comparison to NOAA's machines - UWC-West can manage four quadrillion calculations per second compared to its US counterpart's 27 quadrillion - but the upgrade has still made a significant difference to Met Éireann's forecasting.

"The numerical weather model we would have previously run on our supercomputers would have been updated eight times a day - so every three hours. The new supercomputer allows us to update the model every hour," Hally says.

Hally also draws attention to the issue of ensemble models - the importance of them, and how they necessitate more computation.

"When you do numerical weather modeling, you don't just do one simulation of weather, you do many of them," he explains. "This is to account for natural uncertainty - things like chaos and the butterfly effect all come into it. There is a natural inherent uncertainty in weather forecasting, you have to do many simulations in order to capture all of that.

"Previously, we were doing 15 different simulations every three hours. But now we're doing 30 different simulations."

More computational power can also improve models on a resolution level. Reducing the grid spacing provides a greater level of precision in forecast outcomes, but it also complicates the process.

As Michaud puts it: "Let's say you double the resolution - from a 30km to 15km model by grid space. That's not just double the computing - you are doubling in several different directions. So instead of double, you could need more like 16× the computing capacity."

Additional detail can be added to a model by introducing elements such as ocean components, or new cloud physics - both of which add physical complexity to an already physically complex model.

All of this - the initial observations, and then the simulations derived from the models - creates astronomical amounts of data, all of which has to be stored.

Housing the atmosphere

The European Centre for Medium-Range Weather Forecasts (ECMWF) has one of the largest stores of atmospheric data in the world.

Approaching its 50th birthday, ECMWF has been running operational models since the end of the 1970s and, according to Umberto Modigliani, its director of forecast and service development, has been archiving all the forecasts it produces both operationally and for research purposes.

"The size of that archive is around one exabyte," he says. "It's one of the largest archives of meteorological data in the world."

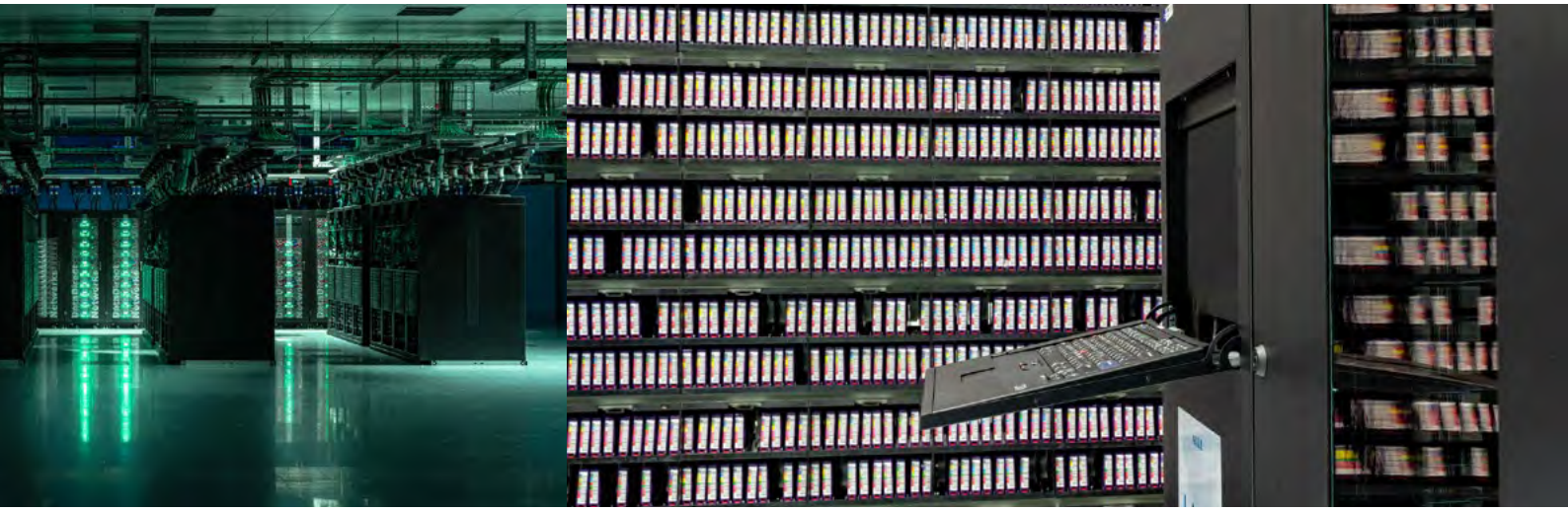
Not only are the forecasts looking towards the future, but extensive amounts of effort are put into re-analysis - or reproducing the situation of weather in the past. According to Modigliani, they have been running these sorts of forecasts from around 1940.

"The data produced a long time ago is still relatively small in size because, at that time, the computing resources were not as powerful," he says. "They were using one of the first Cray systems which, in comparison to what we use today, had a lot less power. The first one, ECMWF, had around 16 CPUs."

Modigliani says the bulk of the archive has "been produced more recently." He explains: "Typically, what we run right now operationally is a global model with a resolution of around 10km. Every day, we produce more than 100 terabytes of data, and we archive something like half a petabyte per day. The size of the archive is growing very quickly, because of what we are doing now, not because of what we had in the past."

The vast majority of that data is stored on-premises using a distinctly old-school storage medium.

"Our archive is still mostly based on tapes," Modigliani says. "We have a parallel file system. If you are accessing recent data, it will typically be from the disks for the supercomputer. If you were looking at trends in temperature somewhere over the last 20 years and need to see our



operational forecasts for that period, then that is sitting on tapes which are then transferred to disks."

Modigliani adds: "There is also a catalog accessible online. We handle something in the order of between half a million and a million requests per day from people."

The use of tape storage is also the preferred method across the Atlantic ocean, at NOAA.

NOAA has a component dedicated to the archiving of data - the National Centers for Environmental Information (NCEI). "It's important because as you get better computational technology and better model physics, you can go back and rerun models with past data and then use those to make sure that you are capturing extremes in different situations," Michaud tells *DCD*.

While also using tape for much of its storage, the solution comes with its own challenges. "If you're storing on tape, which is an affordable solution, the tape densities and the tape technologies and the tape drives change over time, and then you've got to migrate all that onto the new technology. So, as the data is growing, the ability for technology to store that in a higher data density is growing as well," he explains, adding that tape migration is an ever-ongoing process at NOAA, one which has been underway for decades.

NOAA has not, however, made a move to cloud storage, though Michaud notes that others in the industry are looking at that.

Cloud computing in the cloud

The UK's Met Office notably signed a partnership with Microsoft in April 2021 that would see the two collaborating on a new supercomputer.

It would be incorrect to describe the supercomputer as "based in the public cloud," however. As previously reported by *DCD*, the supercomputer is actually housed in a dedicated supercomputing facility within a Microsoft cloud data center in the UK.

"Microsoft is hosting the multiple supercomputers underlying this service in dedicated halls within Microsoft data centers that have been designed and optimized for these supercomputers, rather than generic cloud hosting," Microsoft told *DCD* in a 2022 statement.

"This includes power, cooling, and networking configurations tuned to the needs of the program, including energy efficiency and operational resilience. Thus, the supercomputers are hosted within a 'dedicated' Microsoft supercomputing facility for this project."

The solution holds its own storage capabilities, but can also leverage offerings available from Microsoft cloud.

Amazon Web Services' (AWS) general manager of advanced computing Ian Colle, argues firmly in favor of doing weather forecast modeling in the cloud.

During the AWS Re:Invent 2024 conference, Colle told *DCD* about a particular client - MAXAR - which uses AWS for just that purpose and, Colle claimed, to great effect.

"They (MAXAR) take the data set from the National Weather Service, and they can actually get their model out before the National Weather Service does because they can fan out on the flexible resources on AWS as opposed to what NOAA has to do with their fixed on-premises cluster," says Colle.

Colle notes that it is reliable enough that this is a big part of their business value, with MAXAR customers then able to use the weather report to look at commodities and investments.

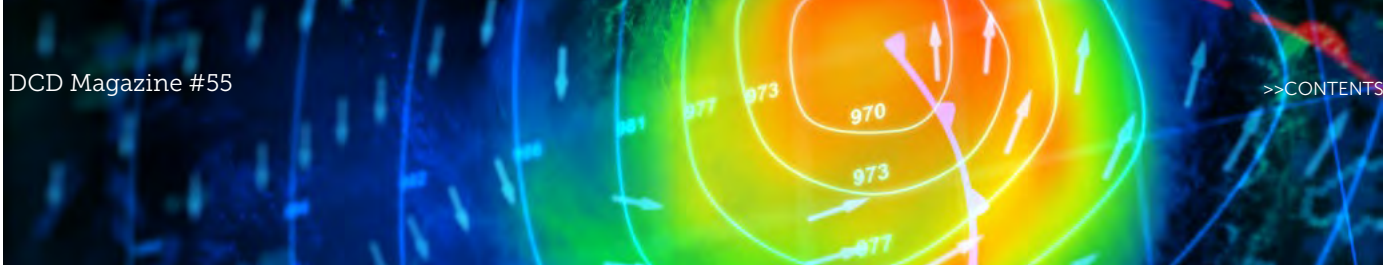
According to Colle, MAXAR is doing this work at the same level or greater fidelity as NOAA, in less time.

The convergence of cloud computing and meteorological data can also, to an extent, be seen in the work of ECMWF. Under the ECMWF umbrella lies the European Weather Cloud (EWC).

The EWC became fully operational in September of 2023 and describes itself as a "hub for the meteorological community in ECMWF and EUMETSAT member and co-operating states, so that users from different countries and organizations can be brought together to collaborate and share resources."

The EWC, Modigliani explains, is hosted in ECMWF's data center - the same as its supercomputer - but is a "physically separate system." It is connected to the archive, and the supercomputer via an internal network.

"The data we produce on a daily basis was also one of the main reasons for having this cloud service so we could offer more flexibility on the sort



of application you could run," he says. "We wanted to have that system on-premise, because of the amount of data, in principle, that users could access. We discussed using a public cloud provider, but if you want to use the 100 terabytes produced daily, that would need to be transferred, which isn't really feasible."

The cloud offering also gives users more flexibility.

It is not only the ECMWF using its supercomputer for forecasting - member state users can also access it. However, the HPC system is managed by ECMWF, and "users don't have a lot of flexibility on what kind of applications they can deploy on the system."

"The idea is that they get their own infrastructure [on EWC]. We create tenancies where they have their own virtual machines (VMs). At the moment, the basic service is VMs, but there are some also running Kubernetes clusters and then, if the user wants, we can manage things at an infrastructure-as-a-service" level.

"Some of the workloads just can't be done on an HPC. For example, creating a web-based service that they can provide access to someone else, that couldn't run on an HPC. So, for us, the EWC is more like a compliment to the main HPC to facilitate applications that were not there before," says Modigliani.

Another stark difference is the makeup of hardware.

The ECMWF's HPC system is mostly CPU-based, with only a relatively small segment utilizing GPUs. The EWC operates on both CPUs and GPUs, which Modigliani says is down to demand for GPUs "particularly in the last couple of years," with the growth of AI systems making the accelerators harder to come by.

GPUs and AI models

Met Éireann's UWC-West supercomputer is solely CPU-based. Met Éireann's Hally says the procurement for that system started in 2019 or 2020 and, at that tim,e AI and weather forecasting "weren't really the big topic they are now."

This focus on CPUs for weather is due to the fact that numerical models used for weather forecasting are not designed with GPUs in mind.

"The traditional weather modeling code - the actual computer code - was written originally to work on CPUs in terms of the way it was parallelized and sent to individual processors, but there are ongoing efforts to change the code so it can be used on GPUs because that is the way the industry is going and where the technology is going," says Hally.

ECMWF is similarly working on making its code work for GPUs.

"The problem is that the model that we are using - the integrated forecasting system (IFS) - is a physics-based model," says Modigliani. "Firstly, it is mostly written in [programming language] Fortran at the moment. It has been developed over the last 20 years or so, and it certainly has more than one million lines of code.

"It's a large bit of code with a lot of different components, and in practice, you need to rewrite the whole of it. It's not an effort that you achieve in a small amount of time."

The ECMWF has been doing some testing on GPUs - with Modigliani citing the Destination Earth project as an example, which has access to EuroHPC resources including the Lumi (see page 86), Leonardo, and MareNostrum 5 supercomputers in Europe - but the performance thus far hasn't radically improved: "there isn't really an advantage for us."

"There are some areas that could benefit from the use of GPUs, but we still need to improve the GPU version of the model by at least a factor of five."

The winds are certainly blowing in the direction of GPUs. In September, NOAA announced a $100 million grant from the Bipartisan Infrastructure Law and Inflation Reduction Act, which will be used for the procurement of a new supercomputer dubbed Rhea. Rhea will be equipped with an unspecified amount of GPUs and, according to NOAA, will be used to "strengthen NOAA's exploration and application of artificial intelligence and machine learning capabilities."

As put by Met Éireann's Hally, the outlook is "rapidly changing." He says: "Every couple of months, I read a new scientific article about an advancement in AI and weather forecasting."

Notably, this December, Google announced that its DeepMind research lab had developed a 'GenCast' AI weather prediction model that it claims outperforms traditional methods on forecasts up to 15 days. This "marks something of an inflection point in the advance of AI for weather prediction," Ilan Price, a Google DeepMind research scientist, said in a statement.

Hally adds that, once a model is trained, the number of GPUs needed to run it is much smaller. "It can also be deployed on very simple hardware and architecture, meaning you can make models available to those in developing countries that don't have as much access to high-performance computing and they can then do their own simulations for their own targeted area."

What remains a certainty, is that the pursuit of more compute will not reach an endpoint any time soon.

"I don't see an end to it," says Michaud. "In all seriousness, we've been doing this for a long time, and I think what will end up happening is that scientists and researchers will have to prioritize the balance of resolution increase, complexity, and certainty.

"You may choose to deal with better data simulation and not do as much with the model, or you may do less simulation and increase complexity in the model."

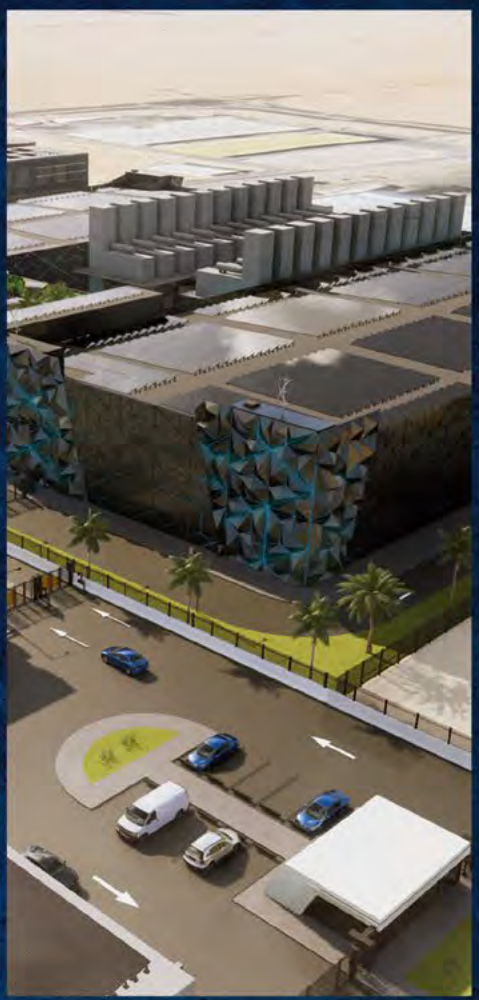
This sentiment was reiterated by Modigliani. "It's very difficult for us and other organizations to say 'let's take a break and wait a few years.' The goal is always to provide the most useful forecast."

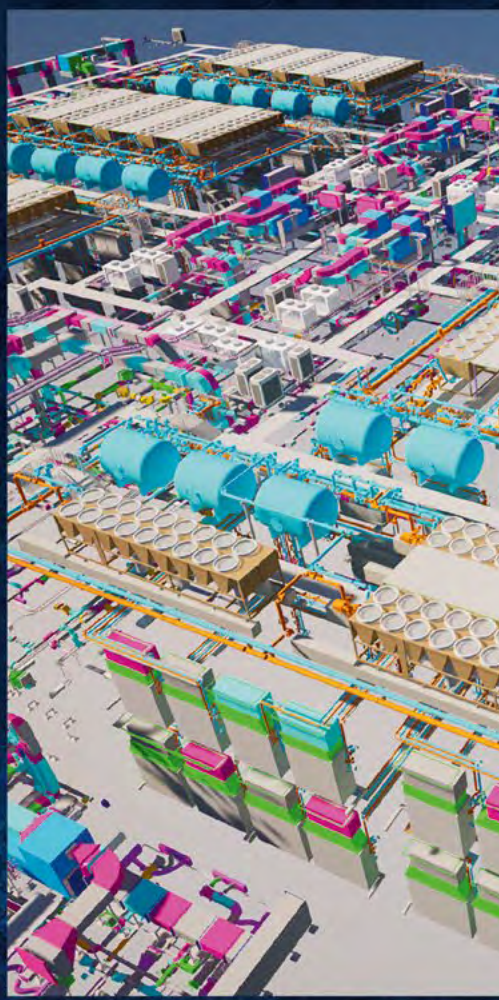
TV presenters looking to avoid their own "Michael Fish moment" will certainly be grateful for the extra help.

WE DESIGN THE BUILDINGS THAT CONNECT THE WORLD

RED

A company of TRACTEBEL ENGIE





20 YEARS

Celebrating over **20 years** of engineering excellence in Data Centre design.

RED Engineering Design, est. 2004

Handling a nation's confidential data

Georgia Butler
Senior Reporter
Cloud & Hybrid

GCHQ loves its secrets. It also loves reading yours

The Government Communications Headquarters - better known as GCHQ - is one of the most mysterious government departments in the UK.

The agency, which handles signals intelligence and monitors communications similarly to the US' NSA, has long been dramatized along with its human intelligence counterparts MI5 and MI6 in film and television.

There is a great deal of intrigue about the agency and the work it does. After all, who doesn't love a secret?

Among those 'secrets' is the digital infrastructure GCHQ relies on to conduct its operations. While unable to get into the nitty-gritty of the exact technology storing the country's confidential information, *DCD* was able to corner GCHQ's ex-CTO Gaven Smith at the October DTX event in London and subject him to a light interrogation.

Smith's sheer joy of his years at GCHQ - a total of 30, with the last seven as chief technology officer - was plain to see. Having left, he now holds positions as a non-executive director at Beyond Blue Limited and Interrupt Labs, as well as a part-time professorship at the University of Manchester.

"I think it's the best job in the country for an engineer - which is controversial with a lot of other engineers," he says. "Running the technology half of GCHQ is a massive privilege, and you are conscious that today's Alan Turing could work in the organization that you are responsible for."

Turing looms large over GCHQ and the technology it uses. The agency was originally established after the First World War, then known as the Government Code and Cypher School. But it is perhaps best known for the work it did at Bletchley Park during the Second World War.

Dramatized in the 2014 film The Imitation Game, staff including Turing used the Colossus computer - a machine weighing several tons and made up of thousands of vacuum tubes to function - to analyze vast sets of communications data from the Nazi military and help crack their codes. That work is estimated to have shortened the war by several months.

Since the war was won, the technology propping up GCHQ has evolved significantly, and "computer" barely has the same definition.

Supercomputers and data centers

Today, GCHQ utilizes a mixture of on-premise data centers, and in the cloud. "GCHQ still has lots of different data centers - the biggest one is at headquarters - and these are big operations," says Smith.

The headquarters in question is affectionately known as The Doughnut. GCHQ had been housed in two sites on opposite sides of Cheltenham since 1952, but in 1997 decided to redevelop its accommodation, signing a contract with IAS Limited in June 2000 for the new headquarters in an agreement set to last thirty years.

The move to the new location faced significant backlash and the National Audit Office conducted a review of the move in 2003 which found that, while GCHQ had budgeted £40 million ($51.93m) for the technical transition, its estimated cost rose to £450 million ($584.22m) - today that would be worth around £891.65 million ($1.15bn).

The data center that lies under GCHQ has long been the subject of speculation and rumors. While Smith confirmed its existence, and said that its data halls are

home to supercomputers, he declined to share detailed information on the compute power and hardware that lies in that facility.

"It's a mixture of things. It has lots of very boring computers - similar to what you'd see in any data center - but there are also supercomputers," Smith says. "GCHQ does use supercomputers - some of the problems it has to solve require that kind of compute."

He offers Colossus at Bletchley Park as a comparison: "Essentially, the computing estate does the modern equivalent, plus a bunch of things that help analysts make sense of data," he explains. "Sometimes those are really simple analytics - searching for something, or sometimes you are looking for patterns in data."

In a follow-up email to GCHQ from *DCD*, the agency responded with a blanket "neither confirm nor deny" response to all queries, including regarding the compute capacity that lies in the bowels of the building.

While hard facts are difficult to come by, speculation is abound. In a 2014 book - *Shaping British Foreign Defense Policy in the Twentieth Century* - R.J. Aldrich, a professor of International Security at the University of Warwick, noted: "The exact size and type of these computers are secret, but GCHQ is rumored to have several machines each with a storage capacity of 25 petabytes (25,000 terabytes) equipped with over 20,000 cores to provide rapid parallel processing. Such computers are required for only a few specialist scientific tasks: simulating complex weather systems, mapping the human genome, designing nuclear weapons, and of course cryptography - the science of making and breaking ciphers."

That prediction was made over a decade ago. In 2014, the most powerful recorded supercomputer was the Tianhe-2 which had a performance of 33.86 petaflops on the Linpack benchmark. Today, the number one spot is held tenuously by Frontier, with 1.2 exaflops of performance, or 1,200 petaflops. With or without confirmation, it's likely the compute power GCHQ has at its disposal has also grown exponentially in the last 10 years.

The quantity of data that GCHQ handles has also increased. In 2016, the Investigatory Powers Bill - later rebranded as the Investigatory Powers Act 2016 (IPA) - came into force, and made significant changes to the pool of data GCHQ could access, and how.

"GCHQ has access to a lot of data - it's authorized to do that," Smith says. "There are rules - it's got to be necessary, it got to be proportionate, and it's got to be authorized."

The Investigatory Powers Act brought together the powers of law enforcement and the security and intelligence agencies to obtain communications and data about communications. Within strict regulatory boundaries, the agencies can effectively pool their resources, though access to this data and communications is, of course, still restricted.

Smith explains that accessing that information is held within a "double lock," so every warranted access is signed off by the Secretary of State and has to be approved by an independent Judicial Commissioner from the Investigatory Powers Commissioner's Office.

For an interception warrant to be issued, the data can only be accessed if it's in the interest of national security, the economic well-being of the UK, and to support the prevention or detection of serious crime, and the data accessed by be proportionate to the need.

According to Smith, even that process in itself can lead to significant amounts of data but "the policy, the ethics, and the legalities are super important to everybody that works there. We're all acutely aware of our responsibilities. It's an intrusive power."

Despite this, the IPA - and GCHQ's past behavior - has been controversial.

In 2013, reports emerged that GCHQ, along with the US National Security Agency (NSA,) had cracked a lot of the online encryption used to protect people's personal data, online transactions, and emails according to documents revealed by whistleblower Edward Snowden.

The agencies were accused of using "covert measures" to set international encryption standards, supercomputers to break encryption with "brute force" and collaborating with technology companies and Internet service providers to put in backdoors. There was allegedly a GCHQ team dedicated to finding access to encrypted traffic on Hotmail, Google, Yahoo, and Facebook.

While these accusations precede the IPA, the IPA in itself has been criticized as enabling more violations of privacy.

Liberty Human Rights group took a case to the UK High Court against the IPA in 2017, arguing that it intruded upon the private life of individuals and interfered with the rights of journalists and lawyers to communicate confidentially with sources and clients.

The court ruled in favor of the IPA. It upheld its legality in 2019, however following Liberty's appeal in 2022, it was revealed that MI5 had unlawfully mishandled personal data leading to a separate case being started. In 2023, the Court of Appeal ruled in favor of Liberty that sharing data from bulk personal datasets with overseas states was unlawful.

In 2024, some safeguards were added to protect journalists from having confidential journalistic material accessed by state bodies "easily."

Regardless of the ethics and legality of the act and mass surveillance, when it comes to data storage, the answer is predictably complicated.

"It's a mixture of things," says Smith of the agency's data storage capabilities. "It's a phenomenally complex infrastructure under GCHQ - made up of hundreds of thousands of systems. Some of this is cloud-based, and some are on-prem. It is large scale."

GCHQ, along with MI5 and MI6, signed up to be an Amazon Web Services (AWS) customer in October 2021 with plans to

> *"It's a mixture of things. It has lots of very boring computers - similar to what you'd see in any data center - but there are also supercomputers"*
>
> *>> ex CTO Gaven Smith*

Gaven Smith

host "top secret material" on the cloud platform.

At the time of the announcement, the data was said to be held in AWS data centers in the UK, and was hoped to help share data internally more easily including enabling the agencies to search each other's databases faster.

Response to the contract was mixed, with many expressing concern over the advisability of putting such sensitive data into the hands of a US-based private company. At the time, Gus Hoesin, executive director of Privacy International, told the Financial Times: "If this contract goes through, Amazon will be positioned as the go-to cloud provider for the world's intelligence agencies. Amazon has to answer for itself which countries' security services it would be prepared to work for."

Having contracts with the CIA in place since 2013, AWS notably scored a huge contract earlier this year with the Australian Government to develop a data center that would host "top secret" data in the country.

Smith notes that moving to the cloud is not a perfect solution in all circumstances. "Lots of organizations are trying to get out [of on-prem], but when you run highly classified infrastructure, you've got to put it in secure spaces," he says.

As for the department's use of cloud computing, Smith emphasizes the importance of trust in your provider.

"When it comes to the use of the public Internet, clearly we are going to follow the National Cyber Security Centre (NCSC) guidances about doing that in the same way any government department does," he says. "You have to know where your data is, you have to be happy with the end users' licensing and that tends to mean using known services that you have a degree of trust with - for example, Amazon, Microsoft, and Google."

Follow-up questions for GCHQ regarding its use of cloud computing were similarly met with "neither confirm nor deny" responses.

Future technologies

GCHQ is already an established user of AI - something the agency has been relatively open about.

"There used to be a thing about 'how to catch a terrorist,' and it talks a lot about the analytic processes that go into finding patterns in data and looking for the known unknowns and the unknown unknowns in data. Of course, increasingly, that's about AI," says Smith.

In 2021, the agency released a paper titled *Pioneering a New National Security: The Ethics of Artificial Intelligence*, in which it laid out its AI strategy, noting that "an increasing use of AI will be fundamental to GCHQ's mission of keeping the nation safe."

According to that strategy, AI is mostly used to deal with large amounts of data and solve "well-defined, narrow problems" that are too time-consuming to be handled by a human alone.

Some of this will be in tackling cyber security breaches through the NCSC. AI will be able to identify malicious software by analyzing activity on networks and devices at scale, identifying patterns, and then updating its "known patterns" of malicious activity.

What differs in the agency's use of new and emerging technologies, however, is its level of transparency. The strategy recalls Winston Churchill's description of the veterans of Bletchley Park as "geese that laid the golden eggs and never cackled." The work of Bletchley Park remained a secret for decades, but Smith says there is an increasing attitude of openness.

"GCHQ has been open about its use of AI, which I think is a good thing," he argues. "We should be talking about the things we do so that people understand.

"One of the things I was responsible for [as CTO], was getting us onto the Internet. So you can see there are around 60 GitHub software development projects, and I think that's important for a bunch of reasons."

Smith argues that, by increasing transparency, the codes used are actually improved: "It's a sort of retention tool for software developers. You write better code if your peers are going to review it and publish it to the Internet."

Additionally, if people then download and use that software for other purposes, Smith says it is a way of introducing "cybersecurity and resiliency" where it may have previously been lacking.

This attitude extends beyond AI, and into the world of quantum computing. "One of my first public speeches was about quantum computing," recalls Smith, adding that it was probably what brought him to public consciousness.

"I did a speech at a quantum showcase, and I spoke about what GCHQ is doing. It basically said: we didn't talk about Colossus for 50 years, we didn't talk about [encryption algorithm] RSA and [its inventor] Clifford Cocks for 25 years, so let's talk about quantum now. Let's admit that quantum is a really important issue for us now - and most important is being quantum safe."

In October of this year, the UK government opened the National Quantum Computing Centre (NQCC), located in Harwell. The site is set to house 12 quantum computers, of which the likes of Infleqtion and Rigetti are already known to be involved.

Quantum is mostly in the remit of the NCSC, which is looking at quantum-safe algorithms that can resist quantum machines powerful enough to crack RSA encryption, which has become the standard method of protecting data.

"Nobody really knows when that will be," admits Smith. "It might be in five, 10, or 15 years, but it will certainly be in our professional lifetimes. So the most important work GCHQ is doing is to support the cybersecurity effort around quantum."

Beyond that, the agency is investing in quantum computing for data analysis. Quantum computers, still under development, come in many forms and rely on a variety of technologies. Which type, or types, will become dominant in the market, so GCHQ is hedging its bets.

"There is some early-stage research going on - this is where the National Security Strategy Investment Fund comes in," Smith says. "There's quite a lot [of quantum technology] on our doorstep in the UK, which has a really exciting quantum computing ecosystem. But I don't think anyone has decided which quantum computer is going to be the solution yet, so you've got to look at them all."

Smith recalls that Patrick Vallance, when he was chief scientific officer, used to call this attitude "optionality."

"I would call it spread betting," he says. "Nobody knows which technology will win out, but it matters so much for us to use these capabilities to keep us safe, and organizations like GCHQ have to be at the cutting edge of that to use it and advise the government.

"I just wouldn't want to pick the wrong horse."

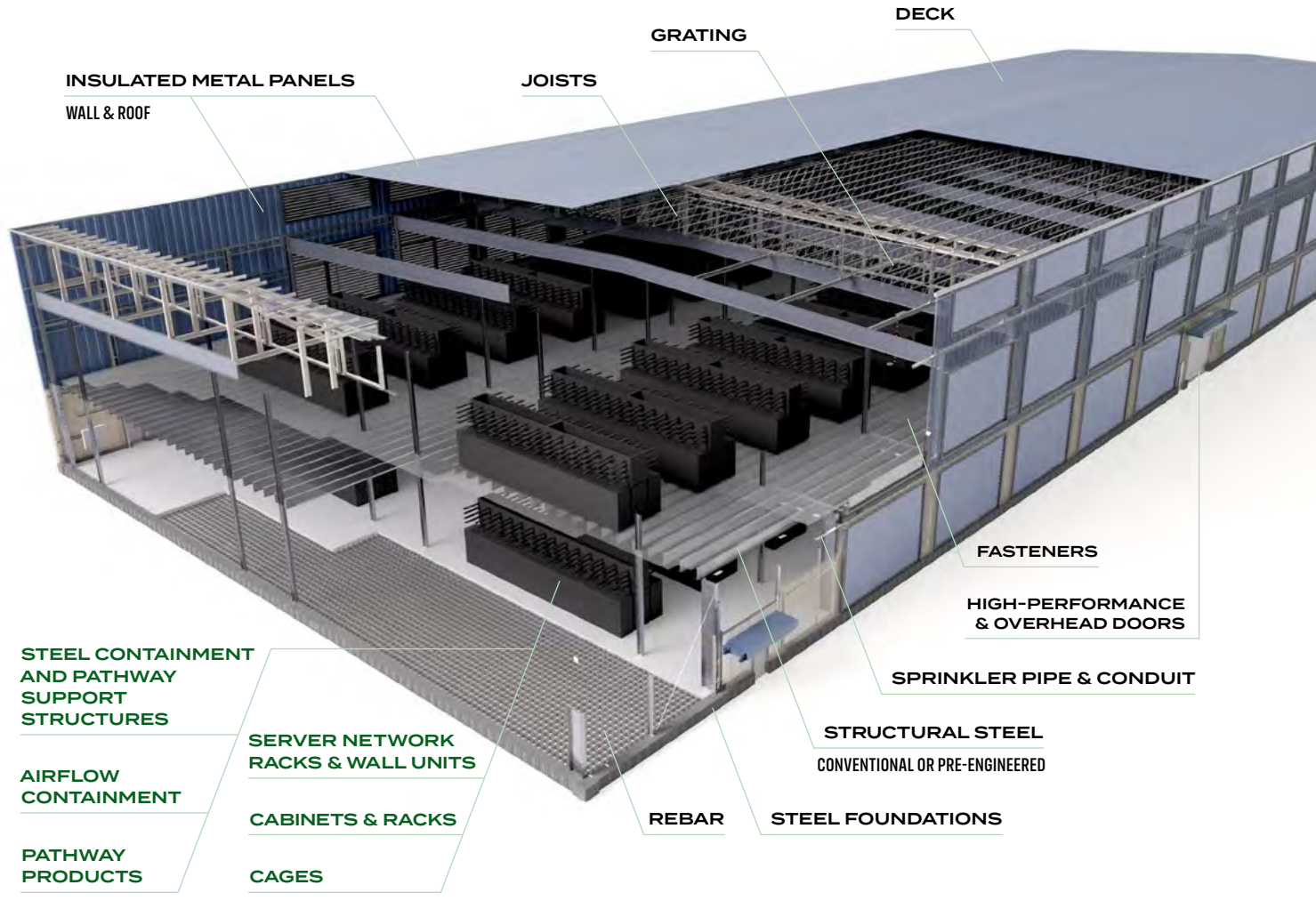
YOUR COLLABORATIVE PARTNER FOR SUSTAINABLE DATA CENTER CONSTRUCTION

As North America's largest and most diverse steel producer, Nucor can provide sustainable steel products and services for your next data center project. All Nucor's steel products are made via electric arc furnace (EAF) technology resulting in high-quality, low-embodied carbon steel to help meet your project's sustainability goals.

- Core and shell products
- White space products & installation

BENEFITS OF PARTNERING WITH NUCOR

- SPEED TO MARKET
- BREADTH OF PRODUCTS & CAPABILITIES
- DESIGN PARTNER



NUCOR®

nucor.com/data-centers

The quest for space sovereignty

Laurence Russell
Contributor

Many governments are looking for their own, secure, orbital infrastructure. This may not be a realistic goal

Rapid commercial build-out in connectivity and communications infrastructure worldwide has been the bread and butter of the digital age, bringing record numbers online, connecting countless operations, and establishing an expansive foundation of data-based awareness across all industries.

One consequence of this has been a heavy reliance on private, monopolizing, and increasingly political Big Tech groups, leading to an uncomfortable relationship between the powers of oligarchs and foreign states, and data security.

At September's DCD>Connect event in London, UK government representatives were in attendance to reaffirm Whitehall's understanding of the importance of data infrastructure in Britain, and what industry needed for this to be delivered at scale, securely.

"Government engagement [with the private sector] has started to ramp up," explained Ethan Thornton, the deputy director for data infrastructure security and resilience in the Department for Science, Innovation and Technology (DSIT). "There are things we know that they don't, and certainly things they know that we don't. Industry is going to have to trust us, even with proprietary content."

The UK is not unique in its aspirations to expand and secure its data infrastructure from the ground up to orbit, especially when it comes to the most vital systems.

"Commercial space systems - such as imagery providers, RF systems, and communications systems - are providing more optionality for government use," explains Nick Saunders, chief cybersecurity and data officer for government systems at satellite operator Viasat, which completed a $7.3 billion merger with Inmarsat in May 2023.

Defining sovereignty

Though the widespread privatization of connectivity technologies and companies in the last 40 years has fueled competition and growth across many sectors, it has also placed many essential services under the control of forces driven by the market, leading to uncertainty among certain national security bodies.

> *"Securing the strategic value of space is increasingly becoming a must-have for power blocs"*
>
> *>>Glenn Katz, Telesat*

In July 2020, the UK government announced a total ban on the purchase of Huawei 5G technologies, adding that all Huawei equipment would be phased out of British 5G networks by the end of 2027 following a review from the National Cyber Security Centre. This was in response to similar sanctions imposed by the US government on Huawei, and more general growing distrust in Chinese technology across Nato member states.

Detailed reasons for the ban were outlined in an October 2022 Designation Notice, which stated "The Chinese state and its associated actors continue to seek to exploit weaknesses in telecommunications service equipment, and/or in how providers of public electronic communications networks build and operate their networks, in order to compromise their security."

The report concluded that "dependency on Huawei significantly increases the potential impact of any systemic failures or hostile exploitation and therefore gives rise to unacceptable risks to national security."

Are commercial networks truly secure?

As Western countries take an increasingly hard line with China, communications technology providers, perhaps unsurprisingly, do not believe digital sovereignty is compromised by the commercial nature of their solutions.

"There's nuance to sovereignty," Glenn Katz, chief commercial officer at Telesat told *DCD* in June.

His company's Lightspeed Low Earth Orbit (LEO) network seeks to connect commercial, government, and defense markets worldwide, including for mission-critical and secure applications customers demand security guarantees about.

"Lightspeed has the capability to hand off full control of our satellites and network to a country," he explains. "We've been having preliminary discussions with customers around the world interested in securing that application."

Lightspeed was funded in part by the Canadian government, suggesting a state interest in the development of the solution for sovereign satellite assurance, though Katz insists the constellation could serve any allied government through the use of their own private gate station and full-mesh private network.

"We believe the target addressable market for government use of satellite is bigger than analysts believe because many of these global bodies haven't yet recognized what they want and how it works," Katz says. "But these discussions are happening today, and sovereignty is going to be a big priority."

Viasat's Saunders agrees. He says: "Many governments are planning to integrate sovereign satellite systems and commercial services to increase resilience with a hybrid communications architecture through which information can be sent across government and/or commercial satellites based on the mission need."

"Securing the strategic value of space is increasingly becoming a must-have for power blocs," Telesat's Katz continues. "That's what's driving coverage of polar satellites. From our side of the fence, we see Nato taking this very seriously. We believe this market will one day become one of our bigger markets."

But there's another side to this story. Real sovereignty often means proprietary technology, which can come at the cost of interconnectivity with other systems.

"The idea that every space-faring country could have its own sovereign, global constellation to operate independently without any integration or interoperability to allow data/information exchange, is a hindrance to the operational capabilities that can and will need to be achieved by joint and allied forces," Viasat's Saunders argues.

Both Viasat and Telesat have been reliable developers of communications hardware and networks for the highest

levels of US government branches as well as mission-critical procurers around the world.

The question of Starlink

Commercial operators launching more satellites more quickly than state-owned agencies and militaries have tested how comfortable governments are using privately-owned satellite capacity in times of conflict. SpaceX's Starlink has provided services to the Ukrainian army since 2022 to aid it in the fight against Russian invaders. This has enabled the use of connected drones for surveillance and strikes.

"Starlink is indeed the blood of our entire communication infrastructure now," Mykhailo Fedorov, Ukraine's digital minister, said in a 2023 interview with The New York Times.

Konstantin Sotnikov, manager of the regional operation division, Lifecell Ukraine, also highlighted Starlink's impact in an interview with *DCD* last year.

"Starlink solutions allow us to restore the network quickly, especially in those areas where the transport network is seriously damaged and needs time to repair," he said.

This is not without its problems. In September 2023, SpaceX's founder, Elon Musk denied the service to Ukrainian forces for the purposes of an attack on Crimea, a territory that has been occupied by Russia since 2014 and that is likely to be hotly contested in any peace deal between the two countries. He said that he did so to avoid complicity in a "major act of war."

Ukraine's use of the technology has inspired attempts by Taiwan to develop its own service, for which it has been seeking financing for since early 2023

"Our primary concern is facilitating societal resilience, to make sure, for example, that journalists can send videos to international viewers even during a large-scale disaster," Audrey Tang, Taiwan's digital minister told the Financial Times in January 2023.

A recently released Wall Street Journal report claimed Musk and Russian President Vladimir Putin had been speaking regularly since 2022, which reportedly included Putin asking Musk for "a favor" on behalf of Chinese leader Xi Jinping not to activate Starlink services over Taiwan, restricting the nation's options for satellite intelligence.

> *"Discussions are happening today, and sovereignty is going to be a big priority"*
>
> *>>Glenn Katz, Telesat*

Kremlin spokesperson Dmitry Peskov denied that such regular conversations had taken place, but confirmed one telephone call on "space as well as current and future technologies," had occurred.

Taipei has expressed an ambition to not "tie ourselves to any particular satellite provider," but rather work with as many as possible to achieve better resilience.

Taiwan's Taiwan Accelerator Plus (TAcc+) space accelerator program attended Paris Space Week 2024 earlier this year, seeking to invest in European aerospace technologies.

"By working more closely with expertise we can build something better together, and support Taiwan's aspirations to become a more capable space nation," Jessi Shen Jye Fu, project manager for TAcc+, told attendees.

Is a wave of space data nationalization ahead?

While not all tech companies are headed by erratic billionaires with authoritarian tendencies, the behavior of some has not reassured decision-makers in intelligence circles about private sector dependency. Demand for sovereign mission-critical systems is on the rise globally.

In August this year, UK Space Command launched a satellite, Tyche, from Vandenberg Space Force Base in California on a SpaceX Falcon 9. Built by Surrey Satellites Technology Limited (SSTL) and fully owned by the UK Ministry of Defence, the washing machine-sized satellite was financed as part of the 10-year ISTARI (Intelligence, Surveillance, Target Acquisition, and Reconnaissance) program, which has a £970 million ($1.2bn) budget.

At DCD>Connect, we asked DSIT's Thornton about what the UK intends to do with its golden share in OneWeb, which survives the company's $3.4 billion merger with Eutelsat.

"The government wouldn't close routes of telecommunication before exploring them, so it's a conversation that's ongoing," he explained.

Late last year, the Indonesian Ministry of Defence established a contract between a state-owned electromechanical giant PT Len Industri and Thales Alenia Space to develop a cutting-edge Earth observation constellation for an undisclosed sum.

The United States and China have been leading this trend for decades, but since the war in Ukraine, many more countries are realizing their interest in securing secure national connectivity.

The European Commission has experimented with pan-state solutions for problems like satellite assurance for years, with their Copernicus, Galileo, and Iris fleets, offering climate data, GPS, and aviation management, respectively.

These systems are financed by the combined purchasing power of the EU state, making use of developers that have become experienced with government contracting at this scale.

Iris is being delivered in partnership with Viasat, using their existing 14 geostationary satellites to manage the wealth of air traffic data that has already caused congestion at European airports across the continent.

But these trends may not be driven wholly by geopolitics.

"[Another important factor is] about countries wanting to exert more technical control and oversight with sovereign space capabilities," Viasat's Saunders argues.

The return of powerful nation-states and isolationist trade policies, epitomized by Donald Trump's re-election in the US, means technological sovereignty is likely to remain a popular goal for many governments.

But achieving it will be an elaborate journey for nations that have pursued healthy and interconnected trading relationships for decades.

Data centre battery solutions

Safety first

Remote health monitoring

Wide temperature tolerance

High energy in compact design

fiamm.com

FIAMM

Grappling with the inference time concept

Charlotte Truemanl
Compute, Storage & Networking Editor

Walter Goodwin CEO
Credit: Fractile

As inference becomes the new buzzword of the AI industry, Fractile founder Dr. Walter Goodwin talks about the need to create hardware to support it at scale

In 2022, Dr. Walter Goodwin was trying to build "general purpose robot brains."

While making robots that are good at many things - most are currently good at just doing one specific thing over and over again - was the focus of Goodwin's efforts, he says at that point in his studies at the UK's Oxford Robotics Institute, his interests became less about how well a robot could pick up a mug, and more about scaling laws.

Having spent four years working on big vision and language AI models that had been trained on images and texts scraped from the Internet, Goodwin says he saw how scaling laws had been "rocking that AI world," particularly when it comes to model training and the idea that increasing the training flops for a foundation model's training run would bring a deterministic improvement in how that model works.

At the time, Goodwin says he was part of a group that started pushing the idea that, as foundation models continue to permeate our lives, this would be accompanied by an inevitable shift in how we think more broadly about AI.

He explains that, from 2011 to "maybe 2020," every problem had its own specific neural network, meaning companies would collect a data set, find the right neural network architecture, and then train it for a specific application until it became good enough.

"Towards the end of my PhD, what I was increasingly convinced by was that [the application-specific neural networks] were going to be eclipsed by this idea of the 'do it all' AI model that is trained on just huge swathes of data and would generalize very well," Goodwin says. "I was seeing that in robotics, and I could see that the same would start to happen with language and vision.

"And what happens is, when you get that shift, the big story in AI stops being: 'How do we train a slightly better model;' it actually stops being so much about that training at all, and it shifts much more to: 'If we're going to be running this small set of models at this vast scale, how are we actually going to do that in a sustainable way?'"

With that in mind, Goodwin returned to his electrical engineering background, forming Fractile to answer the question: "If we're heading towards a world where most compute power is focused on inference, is our current hardware fit for purpose?"

For Fractile, the chip company he founded, the answer was a resounding no.

In Inference is your new best friend

While training has dominated AI conversations in the last few years, recently companies of all sizes have publicly announced a move away from training AI models to focus on the less computationally intensive inference, where, simply put, an AI model uses the patterns it's been trained on to make predictions.

On Microsoft's Q1 2025 earnings call in October, CEO Satya Nadella said the company was on track to generate $10bn

in annual revenue from AI inference and, as a result, the company was turning away requests to use its GPUs for training "because we have so much demand on inference."

Goodwin says that platforms built based on von Neumann architecture - the basis for most general-purpose computers - are full of memory-bound stages which cause a latency cost trade-off. This is less of an issue when training models because throughput, not latency, is the focus. However, when it comes to inference, users expect it to be done at speed, what he calls the "inference time concept" – "[OpenAI's] o1 is the best model in the world, but you have to wait 15 seconds for an answer."

Goodwin says that, for the best part of the last two years, Fractile has been talking to people about this concept of inference scaling.

"We've had scaling laws in training where you increase the training flops, but what inference scaling says is that AI performance is actually about two things," he explains. "How good you can make the base model, and how to use more compute to get better AI outcomes."

To achieve these better outcomes, the UK-based company has been developing chips that use in-memory compute, an approach that allows processors to run calculations directly in computer memory. Goodwin says that, by taking this approach, the company hopes to create hardware that reduces power consumption and improves performance, all while allowing for faster and less expensive inference at scale.

In July 2024, Fractile emerged from stealth, having raised $15 million in seed funding from a round co-led by Kindred Capital, Nato Innovation Fund, Oxford Science Enterprises, and a number of angel investors, including Stan Boland.

While Fractile has yet to bring its product to market, the company believes its hardware will ultimately be able to run large language models 100× faster and 10× cheaper than Nvidia's GPUs, and have a 20× better performance per watt of energy than any other AI hardware currently on the market - although by the time it does launch, competitor hardware will have advanced substantially.

Goodwin notes that, while there are a few companies that are also exploring this concept of more on-chip memory, what Fractile is looking to do differently is remove the need for a separate memory bank and processor, allowing the company to better address what Goodwin believes is the most critical limitation in compute scaling right now, power.

"[With Fractile's approach] what you can achieve is far, far, far higher than you would get if you just had that near memory compute piece. While [near memory] is good for driving up the bandwidth, it doesn't drive up your TOPS per watt, so you still have a chip that ultimately is going to be thermally limited. For a long time now, we've been thermally limited in how we scale up these systems.

"[For Fractile] it's about building a system that will allow us to run inference at scale for these very large models, far faster. That means more tokens per second, more words output per second per user, but also doing all this in a much less expensive way as well."

And unlike most companies focusing on in-memory compute that have thus far mostly deployed hardware in low-power Edge devices, Goodwin says what's exciting for Fractile is that it's one of the only companies trying to bring this technology to data center scale workloads.

"That's one of the things that's more unique about what we're doing," he says.

However, despite the company's ambition, Goodwin explains that one thing Fractile has been careful not to do is simultaneously reinvent too many things, as it's important that the company can achieve not only a good time to market but also scale up production and play within the rules of existing semiconductor manufacturing.

"Fractile is doing things that might be kind of radical in terms of the circuits that we're designing for in-memory compute and how we're thinking about architecture and software but at the lowest level, from a silicon design point of view, we're doing our test chips at TSMC process nodes and our production chips will be at cutting-edge FinFET nodes on standard foundry processes. In that sense, we're seeking to be as normal as possible from a manufacturability point of view."

The zero-billion-dollar market

Given the costs associated with getting silicon to market – Goodwin says a mask set alone costs upwards of $10 million – the first large-scale silicon the company produces will be its first product.

Fractile has been working on prototype test chips, but to date, the designs have only been tested in computer simulations. While he declines to divulge the company's anticipated timeline for bringing a product to market, Goodwin says tape-outs are expected in the next few months.

When asked if the semiconductor industry is at an inflection point, and whether we might begin to see a split form between the ever-dominant incumbents who are having great success with their tried and tested chip architecture and the enthusiastic startups who think there's a new, better approach to be found, Goodwin is rather optimistic about the whole affair.

"When you have a rapid emergence of a very large scale and arguably new workload, which I guess we have today with the inference of these very large models, I think the exciting thing for startups is that there are these whole new markets that emerge.

"It's the Jensen Huang quote: 'The zero-billion-dollar market.' In Fractile's case, I think blazingly fast data center scale inference is today, in some ways, a zero-billion-dollar market. There's nobody that can meet the needs for that, there is no hardware that exists. Fractile is on a path to produce that hardware and so I think what we're excited about is entering that entirely new space and creating a whole set of applications that we can enable."

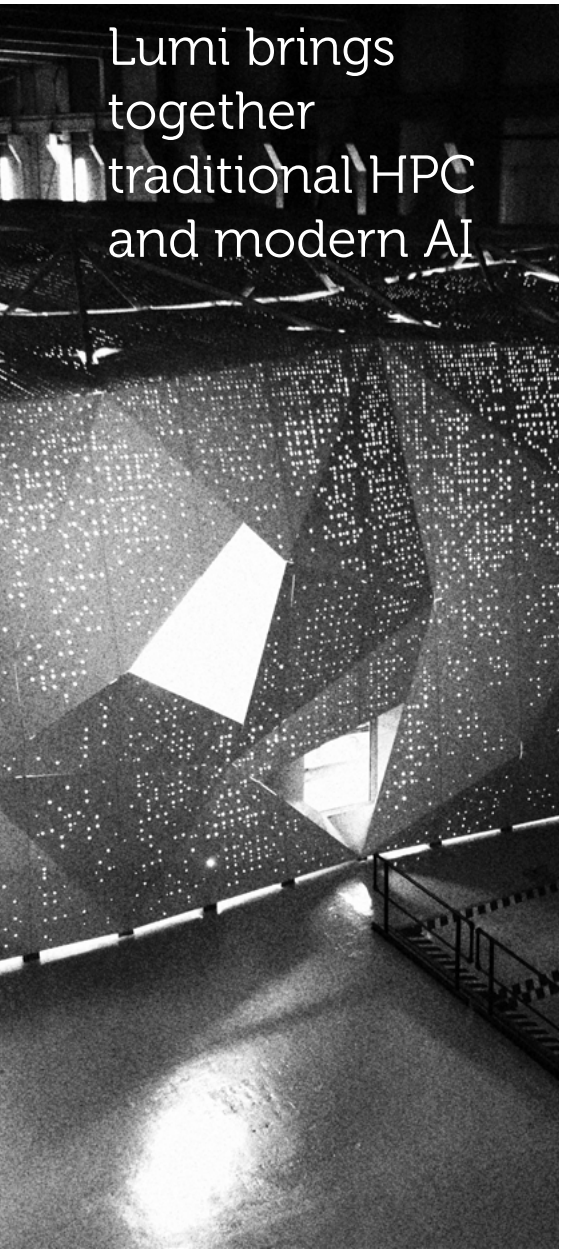
Goodwin says the last six months have been exciting for Fractile, revealing to *DCD* that, as of October 2024, the company has just opened a new office in Bristol and is looking to add a further 10 or 15 people to its current 23-person workforce.

"The critical things that we're working on at the moment actually, beyond the silicon which we've already talked about, is a very large portion of what Fractile works on is entirely in the software layer.

"So, in terms of the markets that we're looking to serve, it's very clear that what needs to be done in order to provide a turnkey solution is to have a hardware platform with a software stack."

Touring Europe's fastest AI supercomputer

Sebastian Moss
Editor-in-Chief



Tucked away in a sleepy Finnish town, inside the giant hall of a disused paper mill, *DCD* comes across a vast glowing box.

We are in Kajaani, located in central Finland, to see Lumi, a 7.1MW European Union project set to help the Bloc as we enter the first innings of an era of artificial intelligence-assisted scientific research.

Now the world's eighth most powerful supercomputer (it recently slipped out of the top five), Lumi was meant to look good. 'The Queen of the North,' as her operators call her, is a building within a building.

The cavernous hall of the former United Paper Mills site is so large that it acts merely as a first layer protective shell around the data center itself, built to comparatively modest proportions, at 300 sqm (3,230 sq ft).

An HPE Cray EX235a system supercomputer, it features 2,978 AMD Epyc Trento CPUs, 11,912 AMD MI250X GPUs, and another 2,048 dual-socket AMD CPUs in a separate partition. Altogether, it has a sustained performance of 379.7 petaflops (HPL).

Its facade is intentionally irregular, with straight walls replaced by odd angles. The face is pocked by slits, from which a glittering light emerges. Unlike the usual grey data center, this one appears closer to a futuristic Christmas bauble.

"The aluminum coating and look was a relatively trivial question in the budget, and we pushed the product manager and the architects to come out with something that looked special," Dr. Pekka Manninen, director of science and technology at Lumi, says.

"Just for the stakeholder work and political goodwill, I think it's paid off."

In an age of ever-present economic and political crises, where science is often lambasted as a luxury boondoggle, Dr. Manninen and the other workers at Lumi are keen to repeatedly tout the broader virtue of supercomputers. Lumi has been good for locals, for the EU, and for humanity more broadly, they say.

The collapse of the UPM mill, once the town's largest employer, was not the economic death knell it could have been for Kajaani. UPM spent millions on supporting laid-off staffers and converting the site into a business park following the closure in 2008, while the Finnish government pumped in tens of millions more.

Around the same time, the increasing price of Finnish wood fiber and decreasing demand for paper had claimed another victim - Stora Enso's Summa Mill in Hamina, to the south.

That site was picked up by Google in 2009, which converted it into a unique data center cooled by seawater. This May, Google announced a €1bn ($1.055bn) expansion to the site, bringing its total investment in Finland to more than €4.5bn ($4.75bn).

That deal gave Dr. Manninen and Finland's IT Center for Science (CSC) an idea: To search for their own disused paper mill. CSC was operating out of Keilaniemi, near Helsinki, at the time. "It's the most expensive square meter price for space in the country, and then it turned out that we wouldn't be able to bring more

than one megawatt online," Dr. Manninen says. "We saw what Google did and looked at three different mills before we ended up in here."

Kajaani benefits from low but not extremely low ambient temperatures, and has a series of hydroelectric power plants that were built to support to mill, with some 235MW available.

CSC soon shifted Finland's national supercomputers to a site on the Renforsin Ranta business park in the town, with the government again providing additional funding to support Kajaani. SGI, Cray, and even Russia's T-Platforms all provided supercomputers for the agency over the years.

Images by Sebastian Moss

> *"During the summertime, basically all the heating needs of Kajaani are from us, but in the winter they still need to burn."*

It is currently home to Finland's national supercomputers Mahti and Puhti, which are set to be replaced by the 49 petaflops Roihu.

The wider park is also home to a Borealis data center, more government systems, and soon a system from British high-frequency trader XTX Markets. In a nice quirk of history, Google itself this November bought land in Kajaani (albeit not at the mill) for a potential data center - citing the growing data center community as its rationale.

Lumi came in 2021 as a joint project with the EU. "We were running procurement negotiations during Covid-19, and then started the installation in June 2021, beginning with the heavy processors and the storage systems," Dr. Manninen recalls.

"We then started on the second phase, the GPU deployment. But we were hit by another black swan event - a global shortage of electronics. It was suddenly impossible to get some small components - power controllers, FPGAs, etc. We were held up by $1 chips."

Still, the team brought the system fully online, with its GPU partition, in the summer of 2022, just before ChatGPT launched and changed everything.

"When Lumi was designed, the generative AI [era] was not foreseeable, but we knew that deep learning would be huge," Dr. Aleksi Kallio, the manager for CSC's AI development program, tells *DCD*.

A colleague of his worked at the education ministry that CSC reports into when the chatbot first debuted. "They have these regular meetings, and at the previous meeting, they decided that we are not so interested in this AI topic anymore, because it's maybe going away," he said. "And then ChatGPT came out: The next meeting, we decided that we were once again very interested in the topic."

Lumi has been used by the University of Turku to develop Finnish and Nordic-language-specific large language models (LLMs). These languages have so far been ignored by those AI system developers with commercial concerns. Beyond its impact in helping maintain a language for the potential next step of computing, it also benefits AI more broadly, Katja Mankinen says.

"There is not that much Finnish in the Internet to use," says Mankinen, CSC's senior data scientist. "The data is quite limited, but they combined Finnish resources with larger resources such as English, and they developed some quite nice tricks on how to make high-performing Finnish models without compromising their quality. So this is one of the things that we believe will lay the base for the future around writing AI applications."

Dr. Kallio adds: "Commercial players don't do things openly. They don't tell what kind of data is given by the models. They don't talk about the model architecture. There are so many hidden things, but these models, they are public for everyone to use and everyone to learn from, for other researchers to build on top of the models."

It is but one project on the supercomputer, Mankinen says: "In Lumi, you can build everything from the really smallest scales - from simulating the matter of particles, how they interact, what the properties are, to what happens in the universe at the galactic scale. Then we also have projects that help more like societal issues, how to cure cancer, how to create personalized medicine, how to help people who have health problems, etc."

Lumi operates like other supercomputers under the European High Performance Computing Joint Undertaking (EuroHPC JU) umbrella – member states and project funders get a portion of the system, while the hosting country which paid more gets to use more.

Finland offers a section of its component to businesses - but access is only free if they open source the results, otherwise they have to pay. Their work is still expected to be broadly in the public interest. Military workloads are banned.

Another workload is Destination Earth (DE), an ambitious EU project to simulate the entire planet, first exclusively covered by *DCD* in 2020 (see Issue #37).

"This is a European Commission project to develop an information system to support decision-making so they can make policy based on scientific facts," Dr. Mankinen says. "Can we do something about the global climate, how can we adapt, what will happen with food, wildfires, precipitation, etc? These are societal questions."

Unlike any other project, it has a dedicated section in the data center for storing data, consisting of around 100PB of capacity. It does not currently have dedicated compute, competing for time along with everyone else, but that could change in the future as the project grows in scope.

There is no upper limit to the demands of simulating the Earth, Dr. Mankinen notes, when asked how large a Destination Earth supercomputer could be.

DE also represents a potential inflection point in the development of supercomputers.

Historically, high-performance computing has focused on high-precision floating point 64 simulations, but AI has pushed for ever lower floating points, potentially going as low as FP4. "GPU vendors seem to focus a lot on AI flops," Dr. Mankinen says.

"Traditional FP64 workloads are not necessarily any faster on the upcoming generation of GPU than they are currently, maybe they are even slower.

"You are probably going to see bifurcation in supercomputers, where some will be more like AI systems and some more like HPC systems."

Dr. Kallio is hopeful this does not come to pass: "We really don't know the future of AI and the safest way to build for it is to have some general purpose capacity well connected together. I think and hope that it will not go down different paths, but we can stay on one single path with one technology.

"Of course, it might mean that we are going to have some accelerators and special purpose hardware connected through the general purpose GPU cluster."

Looking to projects like DE, he says that "what we see as extremely important in the future is the fusion of AI and HPC methods in science. Usually, you start with simulating through the HPC side, then you have an AI model built on that data."

All of this brings the Lumi workers - who are planning to deploy a larger EU AI system in the years to come - back to their core argument: The supercomputer has been a boon for society.

On the local scale, the system is also providing benefits, albeit not in jobs - "one fallacy is that you need to put your data center where your staff is," Dr. Kallio says. "Our staff are not here; they are in Helsinki. We are a skeleton crew in Kajaani."

Lumi produces 40°C (104°F) waste heat water, which CSC then bumps up to 80°C (176°F) with heat pumps to send to the existing district heating network. The heating system "burns pretty much everything it can burn, including food and oil," Dr. Mankinen explains. "So that's why we call Lumi carbon-negative, because they can reduce the amount they are burning.

"During the summertime, basically all the heating needs of Kajaani are from us, but in the winter they still need to burn."

CSC is also using the supercomputer's uninterruptible power supply (UPS) to run a pilot scheme for grid frequency regulation, but Dr. Mankinen admits that they only deployed UPS systems to meet the procurement requirements.

When it was a paper mill, the site experienced only a single two-minute power outage during its 38 years of operations, thanks to the stable, nearby hydropower.

A future exascale supercomputer may cut the UPS entirely, he posits, as he lays out a vision for a number of ever larger systems in Kajaani.

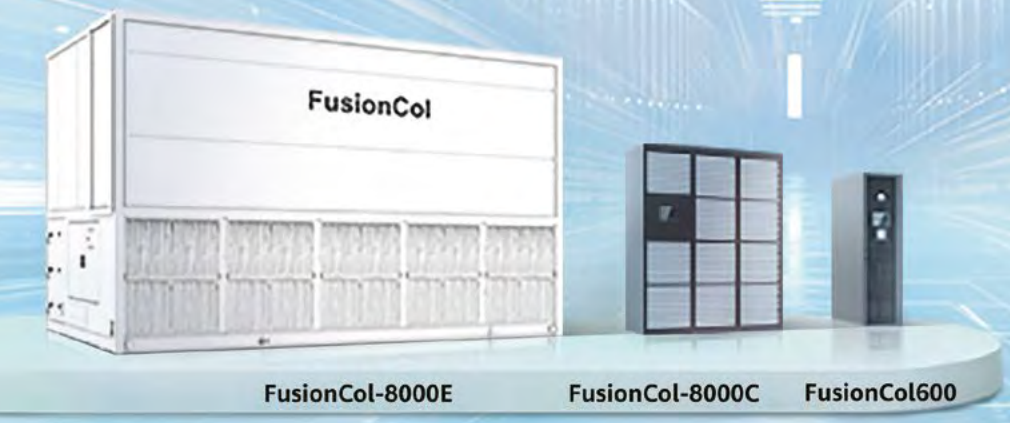
"Supercomputing has a big footprint," Dr. Mankinen says, both in power and cost. "We can and must use supercomputers to improve the quality of our future."

HUAWEI

Building a Fully Connected, Intelligent World

Huawei Data Center Facility

Builds a Large-Scale Computing Center Facility



FusionCol-8000E FusionCol-8000C FusionCol600

FusionPower6000 FusionPower9000

Power the Digital Era Forward

In harmony with nature

Matthew Gooding
Features Editor

Prometheus Hyperscale has an ambitious plan for a network of carbon-neutral data centers

In Greek mythology, the god Prometheus is credited with stealing fire from Zeus, hiding it in a giant fennel stalk, and delivering it to humanity, in the process gifting the world science and technology.

Prometheus Hyperscale, the data center company formerly known as Wyoming Hyperscale Whitebox, has taken the Titan's name and intends to deliver society a different kind of gift: a lot of sustainably-powered AI data halls.

With ambitious plans to develop its flagship campus in Evanston, Wyoming, and roll out its model to other parts of the US, Prometheus has been thinking big since recruiting cleantech entrepreneur Trevor Neilson as president.

In November, the company raised eyebrows when former BP CEO Bernard Looney joined as chairman. Looney left the oil and gas giant under a cloud in 2023 (he initially resigned, and was later sacked for serious misconduct after failing to disclose relationships with colleagues), but his arrival at Prometheus Hyperscale is likely to give the company

Wind turbines in Wyoming

additional credibility in some circles. After several years developing its offering, Prometheus says it is finally ready to deliver.

Down on the ranch

The Prometheus Hyperscale story dates back to a time when AI had barely progressed beyond the pages of a Mary Shelley novel (the full title of which, incidentally, is *Frankenstein; or, The Modern Prometheus*).

In 1869, Trenton Thornock's family headed West and set up a homestead on land at Aspen Mountain, southeast of Evanston in Wyoming's Uinta County. Since then, the site has grown in size and passed through six generations of the family, becoming a well-established cattle ranch.

"My family have been custodians of this land for 155 years," Thornock, Prometheus's founder and CEO, tells *DCD*. "The actual parcel of land we're building on was bought by my dad at auction in 1990, so in that sense it's our 'new' acquisition because we've only had it for 34 years."

Prior to founding Prometheus, Thornock spent his career in the energy industry, working in the Far East to deliver a zinc recycling facility outside Shanghai, China, and running a similar project in the Philippines.

"When I started out, I didn't know much about data centers, but I have developed stuff before and worked on big industrial projects," Thornock says. "For this site, we knew we couldn't just build a data center in isolation, some big noisy box sucking in water and power.

"This is going to be in my backyard. My family has a responsibility to think about what the best use for this land is - a lot of it is virgin farmland that has never been developed before - and we want the data center to be a community asset, an anchor at the center of an ecosystem that adds value for Evanston."

Thornock started Wyoming Hyperscale Whitebox in 2020, with a vision for a 120MW fully liquid-cooled campus using zero water, with power supplied by local wind farms and waste heat being used to warm an indoor farm and grow fresh produce which, it is hoped, will help reduce the reliance on food brought in from elsewhere.

> *"We want the data center to be a community asset, an anchor at the center of an ecosystem that adds value for Evanston"*
>
> *>>Trenton Thornock*

"One of the first things we said from the start was that we would be 100 percent liquid-cooled," he says. "I remember about three years ago someone from one of the hyperscalers telling me that liquid cooling was a niche that would never go mainstream, and that I was wasting my family's money."

Then AI happened.

"Fortunately for us, because we made that early decision around liquid cooling and having a way to reuse the waste heat created by compute, we were already focused on AI, machine learning, and high-performance computing as applications," Thornock says.

Back in 2021, Thornock agreed a deal with Submer to use its immersion cooling pods on the site, but he says direct-to-chip liquid cooling will also be supported in Wyoming.

He says: "We're not going to dictate to our tenants which kind of deployments they make. Our CDU is designed to be multi-liquid and designed to handle the kind of power surges you get within AI clusters."

Going fishing

As the AI revolution was taking hold, Neilson was enjoying a serendipitous fly-fishing trip to Wyoming.

The self-described climate activist has enjoyed a varied career with a heavy emphasis on philanthropy, working in the family office of Bill and Melinda Gates, and later serving as the founding director of public affairs for the couple's foundation. He then launched a consultancy, Global Philanthropy Group, to advise high-net-worth individuals on their philanthropic strategies, teaming with A-List stars including Brad Pitt, Demi Moore, and Shakira on charitable campaigns, a role that led the *New York Times* to dub him "Charity fixer to the stars."

More recently, Neilson worked with Howard Warren Buffett, grandson of Warren Buffett, to run i(x) Net Zero, an investment company focused on sustainability, and founded WasteFuel, a company transforming agricultural waste into low-carbon fuel. He has funded over 100 climate action groups through another project he set up, including Extinction Rebellion and Just Stop Oil, but decided to withdraw his backing for both in 2023, having become disillusioned that their tactics are mere "performative" stunts that "don't accomplish anything."

So how did a man with a contacts book that would be the envy of many a Hollywood agent end up in the data center industry? It's back to that fishing trip, which Neilson embarked upon after deciding to leave his CEO role at WasteFuel.

"Any self-aware entrepreneur eventually realizes they eventually need to replace themselves," he says. "I started that process with WasteFuel and then began to think about what I would do next. This led me to do a lot of fly fishing, and I spent a lot of time in Wyoming."

Neilson wanted a new challenge that involved tackling "the biggest consumers of electricity and other natural resources." Naturally, this included data centers, and he says he became "fascinated by the excess energy produced in the State of Wyoming, and the policy environment there which is very enabling for renewables."

Wyoming, which has the second smallest population of all the US states, produces 12 times more energy than it uses and is the third largest net supplier to other states, according US Energy Information Administration figures. Much of this energy comes from fossil fuel-burning power stations, but efforts are being made to change this. Work has started on a 345MW sodium-cooled nuclear plant, which is being developed by TerraPower on the site of a retiring coal power plant. Wyoming also has a growing wind power sector, which accounts for 90 percent of the renewable energy used in the Cowboy State.

After investigating several projects, Neilson put in a call to Thornock. "I had done enough research on Trenton to know that he had a whole bunch of land

Render of Prometheus Hyperscale Wyoming campus

in a part of the state that I happen to just really love," he says. "I knew he was doing something very smart in southwestern Wyoming and I ended up wanting to be a part of it.

This led to what Neilson describes as "a strategy process" that involved "three or four months of figuring out how to build upon the exciting work that Trenton and his team had already done," and evolve it into "a multi-project pipeline under the auspices of this new entity."

Scaling up

And so Prometheus Hyperscale was born. The company intends to stay true to Thornock's original vision, but is thinking about things on a larger scale.

That starts in Wyoming. Whereas back in 2020, a 120MW campus would have seemed enormous, now it is dwarfed by some of the data centers being pitched by the hyperscalers and their partners. The Evanston site is now targeting an eventual capacity of 1GW, and Thornock says the company is hopeful of having the first 120MW online in the next 18 months ready for hyperscale tenants.

Most of the power for the campus will be supplied directly by four windfarms located in the vicinity of Evanston, and the company claims less than 10 percent of its energy needs will be served by the state grid. Thornock says this percentage will eventually be reduced to zero, with Prometheus having signed a deal with small nuclear reactor company Oklo, which has agreed to deliver the data center 100MW of clean power.

"We're not going to mess around with PPAs and paper 'swaps,'" Thornock says, referring to the renewable energy offset deals often signed by data center operators. "We're in an area where there are underutilized wind assets and they will effectively be our power grid. The utility will be happy because we're increasing utilization."

Prometheus also intends to install solar panels on the site, combining them with grid-scale batteries to ensure they have adequate power available when the wind is not blowing.

Further afield, the company has announced a pipeline of five other projects at locations in Pueblo and Fort Morgan, Colorado, and Phoenix and Tucson, Arizona. These data centers will be attached to existing renewable energy assets such as microgrids, Thornock says, and run on the same carbon-neutral basis as the Evanston site.

"The reason we have projects in Arizona is because we had a solar microgrid company come to us and say 'we're looking for a developer that knows about liquid cooling and heat reuse,'" he says. "These deals are structured as joint ventures with the microgrid operator, so we don't need to worry about land acquisition and things like. We just need guaranteed power and long-haul fiber, then we can bring out the data center ecosystem."

Thornock says this approach has led to a number of interesting opportunities, many of which have yet to be made public.

"We've had a whole series of these distributed generation developers approaching us to say 'we like what you're doing in Wyoming and we would like you to offtake our renewable power,'" he says. "It's mostly solar, with some wind, some connected to the grid and some not. We've announced the five sites but we have more than a dozen in the pipeline, some up to 1GW."

Looney's arrival, announced after *DCD* interviewed Neilson and Thornock, is likely to open more doors for Prometheus Hyperscale. For the moment, the company is focused on getting its first site up and running.

Neilson says the project will maintain the spirit of the original homesteaders who settled on the land in the 19th century. "The families that came West had to work in harmony with the land, or they would die," he says. "It was a very dangerous undertaking, and that pioneer spirit and cooperation with nature is very apt for those of us thinking about the future of AI and how we build infrastructure to support it. If we do not work in harmony with nature, we will not survive, and that's at the core of who we are."

The tale of the original Prometheus concludes with the god tied to a rock and having his liver repeatedly pecked out by an eagle as part of punishment from Zeus. All involved will be hoping the story of Prometheus Hyperscale has a much happier ending.

Forcing a decision: AI investments in the cloud and on-prem

Dan Swinhoe
Senior Editor

Google Cloud's infrastructure GM Sachin Gupta on the infrastructure of AI

Though the company has often lagged behind the likes of Amazon and Microsoft in terms of its cloud business revenues, Google is pitching itself hard as the place to be when it comes to AI infrastructure investments.

Though debatable, during the Google Cloud Summit in London in October 2024, Tara Brady, Google's EMEA president, claimed it was a "fact" Google created generative AI – likely referring to the 2017 Google Research paper *Attention Is All You Need*, which introduced the transformer concept underpinning many of today's genAI models.

Whether startup or enterprise, Google wants your AI dollars. Brady said that today 90 percent of AI unicorns use GCP, and this year has seen the company announce AI-centered deals with companies including Vodafone, Warner Bros. Discovery, Mercedes-Benz, Bayer, Best Buy, Orange, PwC, and others. Pfizer, Hiscox, Toyota, Lloyds Bank, Bupa, and Monzo were also named on stage in London as AI customers.

"We're super excited," Google Cloud GM for infrastructure, Sachin Gupta, tells *DCD*. "When you look at the number of industries, where they're moving from experimentation to scaling and production, I think that's very, very exciting."

"AI is forcing a decision," he says, speaking to *DCD* at the summit in London. "Unlike legacy and traditional applications, AI, in most cases, requires a new infrastructure investment."

Hypercomputer – efficiency gains

Google is certainly investing. Like its fellow hyperscalers, the company is rapidly building out new locations globally and expanding existing campuses.

When asked if the rapid build-out of AI capacity is breaking the traditional public cloud data center architecture of large regions serviced by multiple nearby availability zones, Gupta notes approaches may have to be different for training versus inference.

"When you're doing large-scale training, you want contiguous, large clusters. There are two ways to achieve that; you put them in the same location, or they're close enough and put so much network bandwidth that it doesn't become a bottleneck," he says. "Both options exist, and depending on the location and what's available, we'll pursue a certain design option."

Trillium TPU

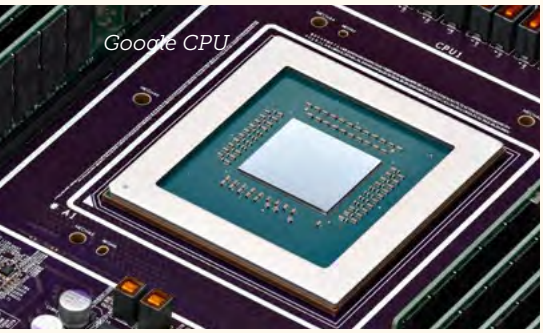
"For inferencing, customers see that as more you're now serving an end customer, and so the reliability and availability of that, the latency of that option, can change sort of how you think about designs."

A key part of Google's AI ambition is its hypercomputer concept. Announced in late 2023, the hypercomputer is described as a unified architecture combining and optimizing all elements of the software and hardware stack.

"We think of it as an architecture of how we build the entire AI-optimized infrastructure with all of the components," he says. That holistic concept runs from the physical data centers' electrical and cooling systems, to the storage, networking, and IT hardware, right through to the software stack including different services and load balancing.

"How do you get the most benefit for customers out of that infrastructure in the most cost-effective way? Every single one of those components matters."

Those gains combined, he says, means Google thinks it can glean around four times the performance from a GPU or TPU - the AI chips the company



Google CPU

developed in-house - compared to a siloed approach.

In an age where companies are buying tens of thousands of GPUs for millions of dollars and building out gigawatt-scale clusters to house and power them, this improvement is significant.

On the company's ongoing data center build-out, Gupta says Google's thinking on AI infrastructure is generally spread across three broad buckets that consider different latency and sovereignty requirements.

"If it's latency insensitive and you don't have sovereign requirements, it's probably best for the customer to just put it in a few locations in the world, and you can use any of our larger clusters to go do that,"

Sachin Gupta

> *"AI is forcing a decision. Unlike some legacy applications, AI requires new infrastructure investment"*

he says.

"There are also many countries where you can train the model on general data anywhere. But to serve the model, or fine-tune the model with your own data, it must be in your country. For that, we look at which countries have those needs and how we put both GPUs and TPUs in-country to support customers."

"The third way to think about it is sheer latency requirement, where now I need to be, wherever my end customers are. There are going to be unique use cases where I need single-digit millisecond latency and have to be much closer to the end user. But I think the latency-sensitive use cases are [in the] earlier stage."

For the first two buckets – large clusters and in-country requirements – Gupta says the company's current and planned region footprint is sufficient. For the low-latency and stricter sovereignty requirements, the company is seeking to offer on-premise cloud solutions.

Bringing AI to the on-premise cloud

Most of the major cloud providers are long past trying to convince large enterprises they should move every single workload to the cloud.

It was interesting, however, at Google Cloud's London summit, to see a slide claiming the industry is now 'past the last cycle of enterprise data center refreshes' and that even large enterprises were no longer just 'cloud first' but aiming for 'cloud only,' though various headlines about *cloud repatriation* may suggest otherwise.

Gupta is more measured on the idea of the death of enterprise data centers.

"For AI, I have to do something new. But which new path do I take?" he says. "If you want the maximum scale up and down, the maximum flexibility, the latest innovation delivered, and those economies of scale, public cloud is the best place to do it.

"But I work with a lot of customers where that's just not going to work. If something keeps you on-premise, we want to make sure that you get that same cloud experience and you can get AI anywhere you need it," he adds.

"Defense, federal government-type, highly regulated industries, central banks, the energy sector; there are just many use cases where they just are not ready to or cannot leverage the public cloud for some of their workloads and data."

Google Distributed Cloud is the company's answer to bringing cloud-like capabilities and consumption models to customers' on-premise or Edge locations.

Like Amazon Web Services (AWS), Microsoft Azure, and even Oracle Cloud, Google offers pre-configured and managed on-premise servers and racks that customers can place in their own data centers or other Edge locations, that offer access to Google's Cloud services. It uses third-party hardware from the likes of HPE or Dell rather than its own proprietary servers.

The connected version of the service starts at a 1U server, scalable to hundreds of racks. The air-gapped offering starts as a small appliance and also scales to hundreds of racks.

A notable customer is McDonald's, which is putting Google Distributed Cloud hardware into thousands of its stores worldwide to analyze equipment. Several companies are using the service to offer their own sovereign cloud to customers.

For customers that want AI capabilities on-premise but want a cloud-like experience – and cost model – without the need to invest in GPUs, on-premise cloud offerings from the hyperscalers could be a viable option, potentially

offering lower latency and costs.

AI-based use cases Google is seeing for on-premise customers include running translation, speech-to-text models, enterprise search capabilities, and local inferencing.

"We hear sometimes that customers are OK to train their model in the public cloud, but then they want to bring it on-premise, fine-tune it, and build their own agent or whatever application they're trying to build," says Gupta.

Application modernization - refactoring and updating existing enterprise applications while trying to build out new AI capabilities - is another driver for these kinds of on-premise deployments.

"Next to that AI application or agent is still a ton of enterprise applications and data in many different locations, a lot of it in VMs, sitting on-premise," Gupta says. "As enterprises look at the investment decision on AI, how much of the rest of that estate also goes to a new cloud model? This can help you migrate those on-prem environments into another on-prem environment that is a cloud environment with Cloud APIs. "

For now, Gupta says the distributed cloud offering only provides Nvidia GPUs; customers are currently unable to get access to Google's own high-end TPUs outside Google data centers – though it offers a mini version known as the Edge TPU on a dev board through its Coral subsidiary.

Google isn't alone in this; neither Microsoft nor Amazon offer their custom silicon as part of their on-premise services. When asked if Google's TPUs could ever make it to customer locations via its Distributed Cloud, Gupta says the company is open to the idea.

"Could we evolve that to AMD or Intel, or to supporting our own servers with TPUs? Yes. But right now, for the use cases we're seeing so far, we feel we've got them covered with Nvidia's A100 and H100, and we will look to support H200 as well."

He adds: "We will continue to look at that market, and if we need to support different hardware there, of course, we absolutely will evolve."

Sovereignty – the real AI investment driver?

Google, like its public cloud peers, is heavily investing in existing markets such as the UK, where the company announced a new $1 billion data center earlier this year.

Amid a major capacity crunch in many established markets, especially in Europe, *DCD* asks if sovereignty was partly driving the company to invest in new AI-focused facilities in these challenging markets, Gupta says that is "100 percent" the case.

"It varies depending on the use case what you need to put in," he says. "How much infrastructure and what infrastructure you put in, really depends on the use cases you're trying to service in the country.

"There are customers who have sovereignty requirements that require data to remain in country. They must do inferencing in country. And so we need to look at how we augment and grow our infrastructure to support that."

Describing this as "a continuum of sovereignty," he continues: "We build data centers of all kinds, all different sizes, and we will put the right infrastructure based on the market needs; it could be TPUs, GPUs. It could be CPUs."

After being the place where genAI began, and watching its competitors take much of the market, Google now hopes that this continuum will finally let it become the home of AI.

CHIPS CHIPS CHIPS

When it comes to AI chips, Google currently offers access to either Nvidia GPUs or its own Tensor Processing Units (TPUs).

"We've all heard Moore's Law is slowing down on the CPU side," Gupta said on stage at Google Cloud's summit. "Thankfully, we can do a lot more processing in the same footprint on the GPU-TPU side as we go from one generation to the other."

TPUs, an AI accelerator application-specific integrated circuit, were first developed in 2015 for the company's internal use – including training its Gemini AI models. They were made available to Google Cloud customers in 2018, and the company is now onto its sixth generation TPU, known as Trillium, unveiled in May 2024.

Oracle, Microsoft, Amazon, and IBM all also offer access to Nvidia and AMD GPUs, with Microsoft and Amazon also offering their own custom silicon. IBM is currently the only one to come out saying it will offer access to Intel's Gaudi 3 GPU.

Despite its rivals all announcing plans to support AMD's latest GPU, the MI300X, Google has been surprisingly quiet. An April 2024 report from The Information suggested the company does not plan to offer AMD's AI chips and "feels good" about its current hardware selection. However, when asked about expanding what AI hardware Google might offer in future, Gupta says the company aims to be flexible to suit customer demand.

"We've been completely open with the CPU side," he says. "In terms of GPUs or accelerators, we'll be open on that side too."

On the CPU side, Google currently offers access to chips from Ampere, AMD, Intel, and its own custom Axion chip.

"There's no restriction for us. We are very open to using the latest and greatest innovations that our customers want there. And so if that happens to be AMD, if that happens to be Intel, absolutely we will go there."

The compromise

Growth requires sacrifice. Try as we might, our species has yet to work out how to develop an advancement without paying a price - be it on land, resources, or human capital.

The data center sector has never been immune to this Faustian bargain, putting the needs of a connected world above local concerns or grid constraints. But, as an industry led by the world's best-funded, and often most scientifically-minded corporations, it has always done better than most.

While renewable energy has now reached a level where it can make better business sense to use (via PPAs) than other forms of energy, it wasn't always so. And yet, hyperscalers pumped money into renewables, losing money in search of reducing emissions.

Similarly, the broader data center industry has pursued sustainability initiatives and pushed to lower the sector's impact on the world. Some of this has been self-serving - either to lower costs, to win over customers, or to head off regulations - but a lot has come from a genuine desire to do the right thing.

But gold rushes have a habit of undermining one's best intentions. As the data center industry dramatically increases the scale of individual facilities and reduces the time to market in pursuit of a promised El Dorado, companies will have to ask themselves how far they are willing to compromise on their sustainability agenda.

The data center industry has turned to natural gas to fuel its next stage of growth. A 2GW Meta mega campus will require the construction of three new gas plants; the CEO of a gas company spoke proudly of how data center demand will fund the expansion of a giant pipeline that spans across natural parks; natural gas operators have begun seeking data center operators to set up shop near their reserves.

This, whether one agrees with the compromise or not, is far from the lofty ideals espoused by the industry. Whatever they may now say about it acting as a temporary bridge, or being less harmful than other fossil fuels, it is undeniable that AI demands are increasing emissions at a time when the planet is already hurtling towards disaster.

This is not inevitable. More sustainable approaches, using renewable energy, large batteries, and careful planning, are possible. They just cost more, take longer, and require a lot more work. But we've done this before.

As a new US president enters office who is certain to further undermine sustainability efforts in the name of unbridled growth, the sector has an opportunity to show that there's a different way.

- Sebastian Moss, Editor-in-Chief


Kohler Energy
is now Rehlko
rehlko.com
rehlko

ZincFive

NEW

BC 2 - 500 UPS Battery Cabinet Powered by Nickel-Zinc

Smallest Industry Footprint & 50% Increased Current Capacity

WATCH VIDEO