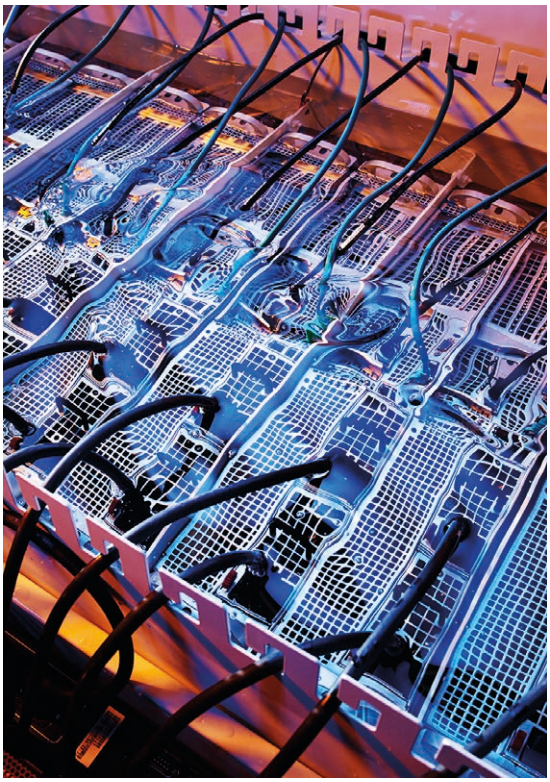

Cool computing – the evolution of datacentre energy efficiency

The ongoing increase in internet use globally is outstripping efforts to control the energy used both to operate and cool datacentres. But some new cooling technologies are available, reports *Elliot Tawney*.

In July 2019, 4.3bn people were active internet users – or more than half of the global population, according to some. Together, we perform upwards of 63,000 Google searches per second, and Ofcom reports that the average person in the UK spends more than a day per week online.

We mostly conduct our online activity without considering how it is made possible – but such a substantial digital appetite requires huge amounts of energy. This job falls to datacentres, warehouses of racks containing servers and computer systems through which most online traffic flows. The International Energy Agency (IEA) reports that global datacentre electricity demand in 2018 was approximately 198 TWh, roughly equal to the electricity consumption of Thailand, or almost 1% of global electricity demand.

Impressively, the electricity intensity of data transmission across networks has decreased by half every two years since 2000, according to a 2017 study in the *Journal of Industrial Ecology*. However, efficiency improvements have struggled to keep pace with the growing demand for computing and the uptake in online traffic, which is set to more than triple by 2022. Avoiding the knock-on rise in electricity consumption – and the associated carbon emissions – will require sustained efforts in innovation and efficiency.

A rack of servers submerged in cooling liquid.

Photo: Green Revolution Cooling

Keeping cool

We are all familiar with the sound a laptop makes as its fan struggles to keep it cool. Similarly, the servers within datacentres generate heat – and must be kept at the right temperature to avoid overheating, which can lead to data loss, equipment damage, system failures and the loss of reputation for the datacentre operator responsible. With data quickly becoming central to most business needs, datacentre downtime now carries a hefty price tag: IT research firm Gartner estimates that system outages in datacentres cost the operator $300,000 per hour on average.

Given the high stakes, cooling is absolutely essential to datacentre operations. Unfortunately, cooling systems are often huge energy consumers, accounting for up to 40% of the total power used in a typical datacentre. As such, they are also responsible for a lot of a datacentre's carbon emissions, and therefore a key area in which to reduce energy consumption and improve energy efficiency.

A major driver of change in the datacentre industry over the last decade has been the 'hyperscale shift', the widespread move of processing power from small, inefficient server rooms to 'hyperscale' datacentres. These are huge facilities with tens of thousands of servers offering cloud computing and storage services to third parties – which can demand over 200 MW of power, but benefit tremendously from purpose-built technology and economies of scale.

Hyperscalers helped establish now-common cooling practices, such as aisle containment. 'People realised there was a real difference between air temperatures coming on to the rack of servers to cool it, and temperatures coming out of the back – the back is probably 10°–20°C warmer than what is coming in the front,' explains Robert Tozer, Director of engineering consultancy Operational Intelligence and visiting professor at London South Bank University. 'Today, most datacentres will have a containment system which segregates cold air supplied to the front of the servers and the hot air that's coming out the back.'

This separation stops hot air flowing into the cold air intake of adjacent racks, meaning the fans can work less, leading to energy savings of up to 30% compared to systems which do not segregate hot and cold air flows.

Another development is free air cooling, where colder external air is drawn in as a cooling source. This reduces or sometimes entirely replaces the need for traditional air conditioning using power-hungry mechanical refrigeration.

The cooler the climate, the greater the impact – thus we have seen a boom of hyperscale datacentres, built by tech giants such as Google and Facebook, across Scandinavia where the air is cold most of the year. But hardier

equipment that can run at higher temperatures has allowed for more widespread uptake, even in climates where air temperatures can be over 25°C. Free cooling is available all year round in more than 75% of North America and in 97% of Europe, and free cooling with air is now used in 40% of the US's datacentres.

Beyond air cooling

However, we may be reaching the limits of what air cooling can achieve. At the heart of the IT equipment that makes up these datacentres are computer chips, and it is here that the rising demand for processing power – crucial to the growth of high-performance computing applications such as AI, big data and cloud computing – is starting to have an effect. More advanced chips, which are crammed into servers at an increasingly high density, have higher cooling demands that exceed the capabilities of air cooling.

Some datacentres are now turning to liquid cooling as a solution, as liquid has a higher heat transfer capacity than air. The simplest and most available liquid is water. Take the Lefdal Mine datacentre in Norway, for example, which uses seawater from a nearby fjord to cool a separate internal freshwater circuit from 30°C to 18°C.

Taking the idea further is US firm Nautilus, which has developed the idea of a floating datacentre. It announced the construction of a €35mn floating datacentre in Limerick's docklands area recently, which would draw water from the River Shannon, pass it through a cooling system and return it to the river less than 2°C warmer, well within recommended environmental limits. The company claims that their cooling system is 80% more energy efficient than the industry average for large and medium data centres, with 30% lower operating costs.

A similar project has been proposed off the shores of Singapore by Keppel Industries, which would not only use water drawn in from the Singapore Strait but, by sitting offshore, would avoid the pitfalls of trying to locate a datacentre on the island, where space is tightly controlled and expensive. Comparable solutions are also now being explored by larger companies, such as Microsoft (see box).

Direct-to-chip cooling

For the most power-intensive processing, direct-to-chip liquid cooling may provide the answer. Liquid (whether water or otherwise) is used in ever-closer proximity to the IT equipment to transfer heat in a more precise manner than air cooling. According to IT infrastructure firm Align, direct-to-chip liquid cooling can provide up to 100 kW of cooling capacity per rack of servers, compared to 20 kW for air cooling systems.

An emerging field is immersion cooling, where IT equipment is submerged in leak-proof cases of non-conductive, non-flammable, non-corrosive fluid. Green Revolution Cooling, a supplier of this technology, advertises a cut to the energy used in cooling of up to 95% using an immersion cooling system, with an upfront cost reduction of 50%.

Once thought of as a niche solution, it seems this type of cooling may soon experience a renaissance. 'It's definitely below 1% of the market now,' says Tozer,

Put the datacentre under the sea

Microsoft is taking the waterborne concept one step further. Rather than submersing server racks or individual components in liquid, they have developed Project Natick, a fully submersible 864-server datacentre designed for the seafloor.

The design utilises water heat exchangers found in submarines to cool the equipment inside, drawing from the reliably cold ocean water and returning it with a negligible effect on the surrounding water temperature. 'While every datacentre on land is different and needs to be tailored to varying environments and terrains, these underwater containers could be mass produced in just 90 days for very similar conditions underwater, which is consistently colder the deeper it is,' said Microsoft.

No staff need to be there permanently, as the datacentre is controlled remotely once submerged. This also removes the need for conveniences like light and space. Combining these advantages resulted in a more energy efficient datacentre than Microsoft's most efficient land-based facilities.

Multiple modules could be submerged in tandem, scaling up to the size of a regular datacentre, or individual units could be deployed in certain circumstances – rapid deployment following an environmental disaster, or near coastal populous cities where most people live.

Deployment anywhere would allow Microsoft to tackle the problem of latency, or the time delay when data is transmitted. Hyperscalers are often located far away from the user, causing delays in data transmission, so the ability to locate closer to users would result in very low latency – which is crucial for some industries, such as finance and healthcare.

In tests, Natick was also powered by renewable energy, fuelling hopes that future iterations could act as anchor tenants for offshore wind farms or tidal turbines, allowing the datacentre and renewable industries to evolve in tandem.

Project Natick is submerged off the coast of the Orkney Islands. *Source: Microsoft*

'but I think this is definitely going to increase big time in the next couple of years.'

In 2018, Google unveiled its TPU 3.0 computer chip, designed to cope with the processing needs of AI applications. The chip is 'so powerful, that for the first time we've had to introduce liquid cooling in our datacentres,' said Google's CEO at a 2018 conference.

Since then, the company has been busy retrofitting its datacentres with liquid cooling capabilities, and other hyperscale companies – such as Chinese conglomerate Alibaba and Microsoft – are on similar paths.

How widely this form of cooling will be adopted remains to be seen, but an uptake in the growing field of high-performance computing is guaranteed.

Waste heat reuse

A final strategy for tackling datacentre efficiency is to reuse the wasted heat – something that is increasingly possible thanks to liquid cooling, says Tozer: 'You can get rid of the heat at a higher temperature, so that lends itself to heat recovery in a much better way than using air – you have liquid available at much higher temperatures.'

Existing schemes are currently hard to find outside of Scandinavia, where the demand for heat is high and district heating networks are more common. But liquid cooling could change that. LiquidCool Solutions tested eight servers by submerging them in dialectric fluid at the National Renewable Energy Laboratory in Golden, Colorado, and found that they were able to recover 90%–95% of the heat, allowing them to heat water to 48°C – hot enough to be useful to the building.

Unfortunately, cooling systems are often huge energy consumers, accounting for up to 40% of the total power used in a typical datacentre

A 2018 study in *Applied Thermal Engineering* showed that a 3.5 MW datacentre using liquid cooling in London could not only reduce carbon emissions by providing waste heat to local heat networks, would also make considerable energy savings in doing so. In the UK, the waste heat would also be eligible for the Renewable Heat Incentive.

However, the problem of location would still remain: 'There are some datacentres in city centres where you can get your datacentre near a heat sink or somewhere where the heat is needed,' says Tozer, 'but that's not the majority of cases. Most are quite remote. There need to be clear policies on where to build datacentres in terms of things like housing schemes to be able to fully utilise this potential.'

Continued efforts

Cooling the world's datacentres, has been a key area of focus in the battle to reduce energy consumption. We have moved from ordinary air conditioning to specialised setups and free cooling and are witnessing the start of a foray into the worlds of liquid and direct-to-chip cooling and waste heat reuse.

But cooling is just one piece of the datacentre energy puzzle. Tackling the energy use problem behind datacentres will also require developments in policies around integration with renewable energy to promote increasing sustainability. Without these, the electricity that datacentres consume could surge to as much as 13% of global electricity in 2030, says a 2015 study in *Challenges*.

The IEA asserts that global datacentre energy demand has so far been kept in check due to the huge progress made in energy efficiency and cooling especially. However, with global traffic set to increase dramatically, and with an increasing need for high-performance computing to run our digital world, the foot cannot be taken off the accelerator. ●