XENON®
High Performance Computing

Liquid Cooling

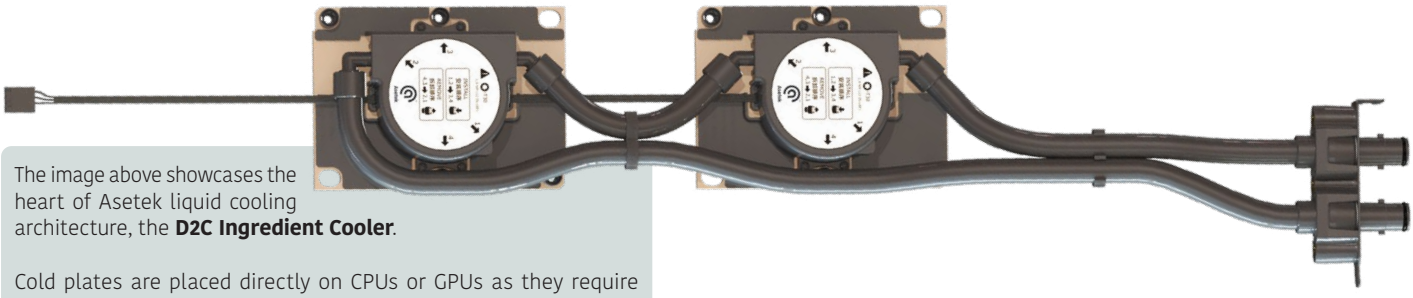
Exceeding the Limits of Air Cooling to Unlock Greater Potential in High Performance Computing

More than good business practice, reducing energy consumption and heat is key to computing on a global scale

Computer chips have been consistently redesigned to break free of performance bottlenecks. Tapping into highly advanced physics and smaller and smaller processor structures, these silicon wonders continually deliver improved features and capabilities that drive ever more impressive applications of technology. Yet CPU or GPU (coprocessor) performance is not the only system element that can create limitations. While every system component has an impact, power is perhaps the most basic yet essential factor in maintaining performance. More than one think tank predicts that computers will need more power than the world can produce by 2040.

The rising level of power being used also generates a significant and increasing amount of heat which, if not effectively addressed, puts future computing advancements at risk. Adopting an efficient process to remove heat is critical to the design of systems that run more complex, higher resolution, and more advanced applications. Solving this challenge is of particular interest to designers of High-Performance Computing (HPC) clusters. Battling limitations of heat and energy as primary issues, this kind of dense and superior computing architecture must overcome hurdles to reach its full potential in scenarios in artificial intelligence, genomics research, weather modelling, financial services, government, and more. This landscape is where liquid cooling is poised to have significant impact on long-term computing performance as well as broad design strategies.

What does that mean exactly? Change is on the horizon, as unsustainable power demands are threatening a global infrastructure increasingly reliant on centralised data centre computing resources. The good news is that the right cooling strategies can not only solve the challenge but also keep high density systems performing beyond current thermal limitations – resulting in improved efficiency, extra computing horsepower, and reduced costs.



The image above showcases the heart of Asetek liquid cooling architecture, the **D2C Ingredient Cooler**.

Cold plates are placed directly on CPUs or GPUs as they require cooling. Cold plates absorb and dissipate the heat from the CPU/GPU to the liquid that is cycled through the system. Liquid enters the cold plate at 45°C enabling the CPU /GPU to operate comfortably within this specification.

CPU IMPACT ON THERMAL INGENUITY

Moore's Law predicts the doubling of compute power, essentially by doubling the amount of energy that is pushed through CPUs. A decade ago, 100 watts per CPU was devastating to thermal design. Today, Intel's highest performing CPUs (e.g. Intel Cascade Lake-AP 9282 processor) have a thermal design envelope of 400 watts. There really is no end in sight, and accommodating more power is critical to advancing performance. The ability to dissipate the resulting heat is the hard ceiling that systems face in terms of performance – giving greater importance to liquid cooling breakthroughs.

With liquid cooling, less energy is expended to cool systems – a significant savings in HPC deployments with arrays of servers drawing energy and generating heat. Electrical current drives the CPU and enables it to function. This electrical power is converted into thermal energy (heat). To maintain a stable temperature, the CPU needs to be cooled by efficiently removing this heat and releasing it. Liquid cooling is the best way to cool a system because liquid transfers heat much more efficiently than air. From an environmental perspective, liquid cooling reduces both those characteristics to create a smarter and more ecological approach on a grand scale. The cascade of value continues, as ambient heat removed from systems can then be used to heat buildings and augment or replace traditional heating systems. It's an intelligent approach to thermal management, distributing the economic value of reduced energy use and transforming heat into an enterprise asset.

LIQUID COOLING IN ACTION

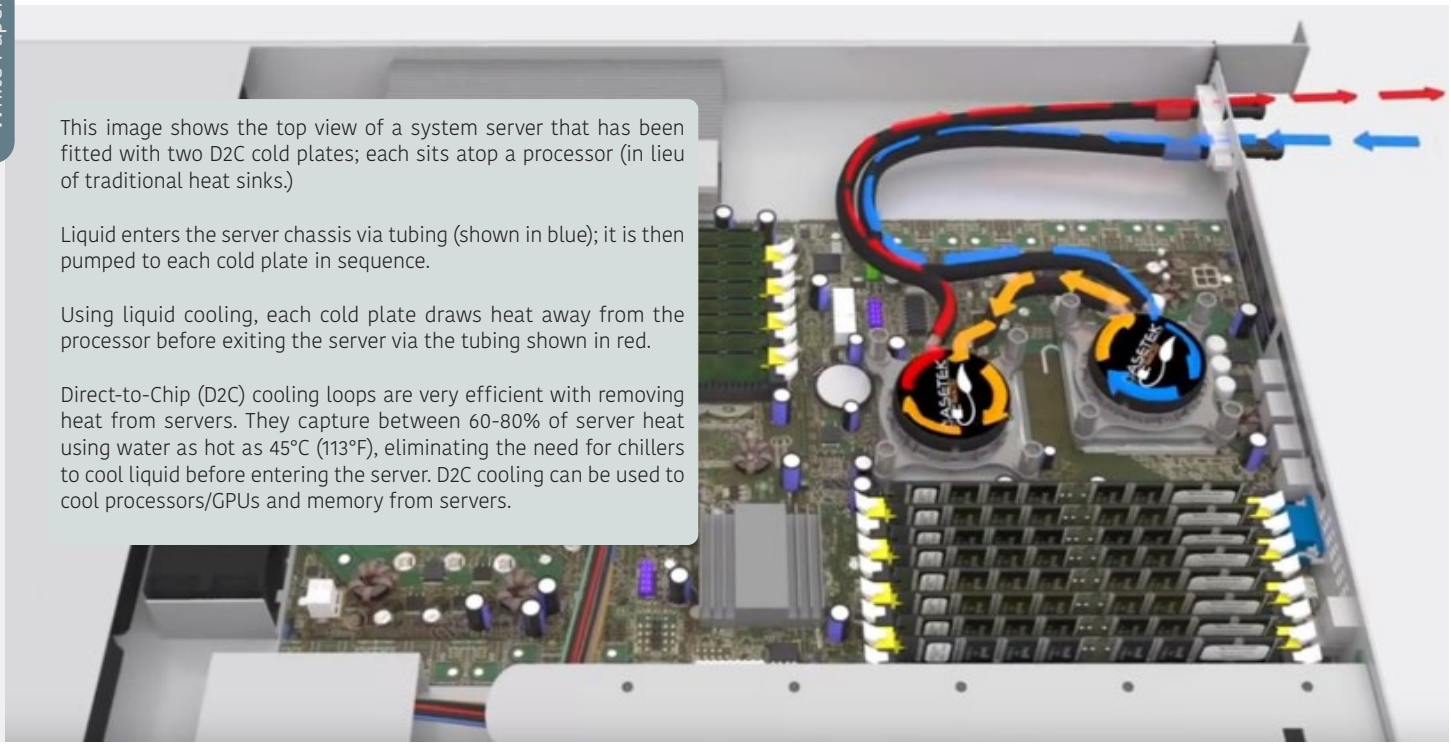
The data centre of a well-known government-run scientific organisation provides an example of the urgency driving thermal innovation. Jumping into GPU computing more than a decade ago, the group deployed the largest GPU cluster in the Southern Hemisphere which was quite groundbreaking at the time. Many of its applications are GPU-accelerated on this system and each rack operated at ~40 kilowatts (kW), a comparatively high heat level resulting from dense, high performing systems. With hardware upgrades providing greater capabilities and better density, compute equipment is reaching 60kW per rack. Heat was impacting the system, keeping GPUs from reaching even their base clock speed.

A liquid cooling proof of concept demonstrated that improved cooling for the CPU and GPU could restore performance and even push it beyond the standard clock speed into turbo speeds. Liquid cooling allows the processor to remain in a turbo state for extended periods of time without overheating.

	AIR COOLED	WATER COOLED
CPU SERVER TEST Running LINPACK		
CPU1 Temp	73˚C	**52˚C**
CPU2 Temp	76˚C	**55˚C**
GPU SERVER TEST Running GPU burn tool		
CPU1 Temp	67˚C	**48˚C**
CPU2 Temp	63˚C	**46˚C**
GPU1 Temp	66˚C	**48˚C**
GPU2 Temp	63˚C	**43˚C**
GPU3 Temp	63˚C	**48˚C**
GPU4 Temp	62˚C	**46˚C**

Table 1. Showcases the efficiency of D2C liquid cooling over traditional air cooling

Image courtesy of Asetek (http://www.asetek.com)



This image shows the top view of a system server that has been fitted with two D2C cold plates; each sits atop a processor (in lieu of traditional heat sinks.)

Liquid enters the server chassis via tubing (shown in blue); it is then pumped to each cold plate in sequence.

Using liquid cooling, each cold plate draws heat away from the processor before exiting the server via the tubing shown in red.

Direct-to-Chip (D2C) cooling loops are very efficient with removing heat from servers. They capture between 60-80% of server heat using water as hot as 45°C (113°F), eliminating the need for chillers to cool liquid before entering the server. D2C cooling can be used to cool processors/GPUs and memory from servers.

LIQUID COOLING ESSENTIALS

The success of this cooling technology is based on direct-to-chip liquid cooling. As CPUs become more powerful, they use more energy and generate more heat. Direct-to-chip liquid cooling makes a difference here, enabling the removal of heat without impact on any of the systems' neighboring components. A dense, traditional server system provides an example: The system may have two CPUs, one behind the other, with spinning fans working very hard to blow hot air across the heatsink. Not only is air already pre-heated as it reaches the second CPU, but it also induces vibrations in the chassis. In addition, boards, memory modules, and other components heat up, which is compounded by the fact that these systems are likely stacked in a dense rack formation. Direct-to-Chip liquid cooling instead moves heat directly away via a cold plate and tubing. No heat is radiated around the components, and much less energy is required for any spinning fans that may still be part of the thermal design.

NO CHILLERS REQUIRED, JUST PLANNING

Next generation Direct-to-Chip liquid cooling can use water as hot as 45˚C to remove heat which alleviates the need for expensive chillers to first cool this liquid. Innovators in Europe and North America are using this to their benefit – driving new kinds of centralised heating systems, or hydronic heating, where this heat is then transferred to a building's heating system. The building is warmed and the water cools as is continues its journey back to the data centre. This cooling and heating method relies solely on pumps to move the water. No chillers are required, and every step of the process brings additional value to the integrated system.

Traditionally, chilled water was required for cooling data centre systems, with refrigeration systems central to bringing liquid temperature quickly down to or even below ambient. Today the cold plate used in Direct-to-Chip methodologies has only a three to five degree Celsius differential between its inlet and outlet, which is all that's required for the efficient cooling of the system components.

To ensure future innovation, data centres must look at this holistically – with improved power usage efficiency (PUE) as the goal. In the old days, if a server generated 1000 watts of heat, it might require 1000 watts of energy to remove the heat – a 100% commitment but a necessary evil to ensure reliable system performance. Today the goal is free cooling by using the available heat energy meaningfully and wasting no energy beyond what is required to run the server. This magic number is reflected in just a 1% ratio of power required for cooling to heat generated by the server system itself.

ADDING COMPETITIVE VALUE TO THE DATA CENTRE MODEL

This is a remarkable evolution from the early days of the internet, when organisations began to realise their cobbled together server rooms didn't provide sufficient space to accommodate growing computing needs. Enter the data centre, a large, offsite facility offering businesses a range of systems deployment and operations solutions and services. The cost model for "co-locating" a company's server farm was initially based on how much data was moving in and out of the organisation's servers via the internet. As companies

Image courtesy of Asetek (http://www.asetek.com)

added more and more servers, requiring more and more physical space, data centre operators devised a new and more lucrative model in which the number of server rack units determined cost.

Today, as HPC clusters and cloud computing have become integral to the world's infrastructure, more and more global computing is centralised in data centres. The profit model has again shifted and is no longer about data used, but rather the power consumed by servers and systems. System operators are working with devices generating 1000-2000 watts or more, which becomes monumental when there are 10,000 or 100,000 systems at play. It's a quick jump to megawatts, and even gigawatts, enough power to run a small city. Yet the ever-increasing use of energy is unsustainable using legacy cooling methods.

TURNING HEAT INTO AN ASSET

Computing needs are only getting more intense as applications such as HPC and artificial intelligence proliferate worldwide. And processing demands power, which means more energy used and more heat generated. Forward-thinking data centres are now leveraging liquid cooled solutions to reduce energy use (and costs) while driving performance, value, and their competitive edge.

The right cooling solution solves performance problems and reduces total cost of ownership. When cooling is efficient, systems can be designed with greater density and can maintain performance beyond current thermal limitations by remaining in turbo state for extended periods. This ultimately translates to operating more servers more efficiently at less cost.

CONNECTING WITH THERMAL MANAGEMENT INNOVATION

Energy consumption and heat generation are by far the most expensive and limiting facets to high performance computing. Liquid cooling, however, mitigates these concerns, providing a smarter, greener option over air cooling methods. It is much more efficient in removing heat. It enables higher performance levels allowing CPUs and GPUs to run at higher clock speeds. Less reliance on computer room air conditioning (CRAC) units result in cost and energy savings. Liquid cooling also requires fewer, slower fans which creates a quieter environment.

Other methods are simply no longer enough and will soon be obsolete for large scale computing environments. Organisations looking for greater value need only understand liquid cooling's effects on long-term computing performance to see its many benefits.

XENON is at the forefront of this shift, as a vendor agnostic solutions provider closely tracking technology evolution and investing in innovation. The company has played a leadership role in managing the sea change in cooling technologies, particularly as HPC cluster systems have moved from research arenas to big data analytics to the spectrum of enterprise computing demands. A holistic approach is central to XENON's philosophy, solving the whole computing problem including software and hardware layers and facilities design. It's this broad expertise and worldview on technology that is critical to firms working to navigate the significant evolution in HPC technology and its immense cooling requirements.

For more insight on how to reduce costs and increase your large-scale system performance with smarter thermal strategies, connect with the engineers at XENON at 1300 888 030 or info@xenon.com.au or visit www.xenon.com.au for more insight.