QCT QOOLRACK:

An Optimized Liquid Cooling Solution for HPC/AI Workloads

Produced by TCI Media Custom Publishing in conjunction with:

Executive Summary

Modern computer infrastructures provide the ability to run artificial intelligence (AI) and high performance computing (HPC) workloads. AI machines need much more than powerful microprocessor chips and use central processing units (CPUs), graphics processing units (GPUs), and acceleration chips to carry out compute-intensive tasks. But higher performance comes with higher power consumption and generates more heat. There are limitations to using traditional data center air cooling with fans and heatsinks when moving towards HPC or AI workloads. Therefore, liquid cooling is required to meet the heat and performance needs of HPC/AI computing workloads and do it in an energy efficient manner. This paper describes how the QCT QoolRack direct-to-chip liquid cooling solution meets IT data center cooling needs for AI and HPC workloads.

Challenges IT Faces with Air Cooling Solutions

Energy price hikes and compute-intensive workloads pose a cooling cost challenge for data centers. The thermal design power (TDP) of next-gen CPUs and GPUs continues to increase, demanding more power and posing a greater thermal challenge for IT managers. Quanta Cloud Technology (QCT) developed an advanced liquid cooling solution powered by Intel® to meet the increased need for cooling in data center infrastructure while lowering the overall power at the rack and data center level.

Why Liquid Cooling is Needed

The key cooling factor of a system's thermal design focuses on its heat dissipation. Data centers traditionally use air cooling for heat reduction. A big heatsink can be installed behind a server rack for heat exchange. While a heatsink provides cooling, there is still a long distance between the chip and the hot aisle of a data center where the heat is blown off. System fans are also used to provide air flow to cool internal system components. However, increasing the speed of multiple system fans increases power consumption and raises the electricity cost of the data center. Increasing airflow through increased fan speed results in diminishing returns in terms of thermal benefit. However, fan power continues to increase as a cubic function of airflow. Increasing fan speed beyond certain point is no longer energy efficient for cooling and one has to consider moving to liquid cooling.

Liquid cooling removes heat efficiently and uses less power than air cooling. Due to increased power usage with air cooled systems, liquid cooled systems will be increasingly replacing air cooled solutions for data centers running HPC and AI workloads. Among currently available liquid cooling methods, direct-to-chip liquid cooling uses a cold plate module that has much smaller dimensions compared to newer (and much larger) heatsinks used for air cooling. Extra space behind the server also allows more real estate for other components and peripherals. Liquid cooling solutions efficiently remove the majority of the heat from chips with less power compared to air cooled solutions.

How QCT QoolRack Solutions Meet Data Center Cooling Needs

Quanta Cloud Technology (QCT) is a global data center solution provider that combines the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation data center design and operational challenges.

QCT developed QCT QoolRack, a rack-level direct-to-chip liquid cooling solution that adopts a cold plate design to meet customers' thermal demands. During the development of 4th Gen Intel® Xeon® Scalable processors, Intel was aware that the growth of TDP could be a great design bottleneck for servers. QCT collaborated with Intel to innovate new, smart, and sustainable cooling solutions such as the QCT QoolRack solution that made its debut at Intel Innovation 2022.

Depending on the data center environment, QCT QoolRack can be operated without adjusting the heating, ventilation or air conditioning (HVAC) to reduce overall power consumption. The QoolRack liquid cooling solution uses the QuantaGrid D54Q-2U server that features 4th Gen Intel Xeon Scalable processors as a gold standard to validate the cooling solution.

The D54Q-2U server is optimized for AI acceleration and can support up to 2x dual-width accelerators in a 2U system. It includes an advanced system cooling architecture supporting top bin 350W TDP CPUs, that is both air and liquid cooling ready. The server provides up to 10x PCIe Gen5 expansion slots and DC-SCM architecture to meet different configuration requirements.

"QCT QoolRack design is an excellent technology that combines cold plate technology with RDHx technology and is capable of cooling high performance HPC and AI servers in an energy efficient manner," said Sandeep Ahuja, Senior Principal Engineer of Intel Corporation.

QCT QoolRack Direct-to-Chip Liquid Cooling

The QCT QoolRack solution uses a rear door heat exchanger (RDHx), a coolant distribution unit (CDU), and cold plate modules to dissipate heat. The RDHx is an integrated radiator that combines a tank and sensors with fans to remove air from the radiator and achieve heat dissipation. The radiator is a big heat exchanger, which is a core location of heat dissipation. Fans at its rear draw the air flow to the hot aisle as the radiator cools down the liquid from the hot manifold. Liquid temperature and liquid level sensors at the CDU allow the cooling rack to function appropriately.

QoolRack Coolant Distribution Unit (CDU)

The CDU is the heart of the liquid cooling system, and it has two pumps with one filter. The CDU pushes liquid to the cold plate module for extraction of heat from key components in the server. Liquid continues to move to the radiator in the RDHx for further heat dissipation from the coolant to the air. The CDU has a liquid temperature sensor, liquid pressure sensor, and air ambient temperature sensor.

QCT QoolRack Cold Plate Module

The QCT QoolRack cold plate plays a crucial role in removing heat from key components inside the server. The chip is the highest point of exhausted heat, and the cold plate module passes liquid through a metal tube to carry out heat from the chip through the liquid outlet that goes to the hot manifold. This results in temperature decreases at the processor(s) and inside the server. The cold plate is metal and meets IEC 62368-1 [2] requirements to help prevent liquid leakage and shorting which could damage electrical components in a server. In addition, the cold plate has tubing and inlet/outlets for easy maintenance inside a server.

Liquid Loop in the QCT QoolRack Liquid Cooling Solution

The Figure 1 shows the liquid loop of the QCT QoolRack solution. Liquid is pushed out by the pump in the CDU located at the bottom of the server rack. Cooling liquid then goes through the cold manifold along the rear side of the rack. The cold plate module(s) extract heat from the chip(s) of each node. Liquid carries heat off the chip(s) which goes through the hot manifold to the RDHx where radiators dissipate heat from hot coolant and use fans to remove heat from the liquid and out of the server rack. The cooled liquid goes back to the CDU for another cooling cycle.

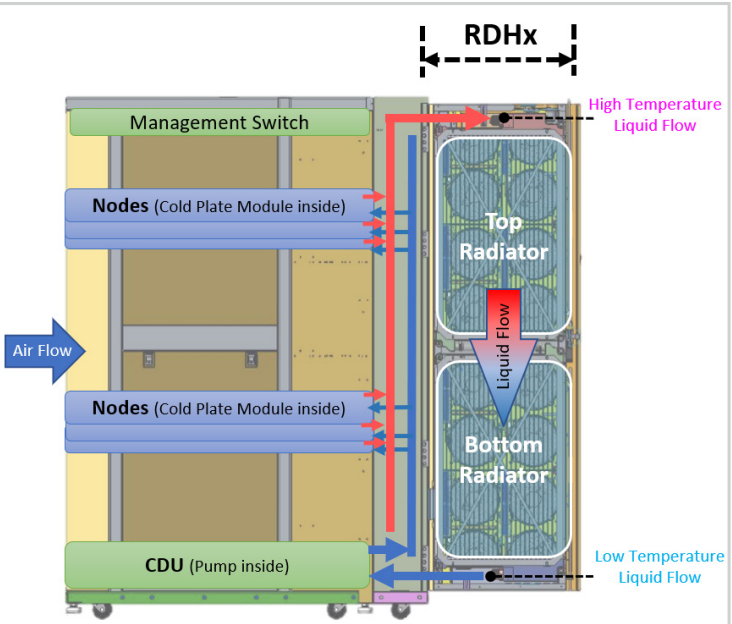


Figure 1. Liquid Loop in the QCT QoolRack Liquid Cooled Rack

Coolant Selection

Coolant is the "blood" of a liquid cooling system which goes through the server to extract heat from the CPU chip(s). The coolant needs to have characteristics to prevent corrosion, biological inhibition, and freeze-proofing in low ambient temperatures. Coolant is usually a hybrid liquid which can be either deionized water with Ethylene Glycol (EG) or deionized water with Propylene Glycol (PG). Each of these coolants are compatible with QCT QoolRack. However, QCT validation focuses on PG considering the lower toxicity rating for humans of PG compared to EG.

QCT QoolRack Smart Management Controller (DC-SCM)

QCT QoolRack executes smart power consumption management according to real workloads. The DC-SCM adjusts the rear door fan speed and the CDU pump speed to balance the coolant temperature. In addition, it adjusts workloads when detecting thermal conditions to enhance power savings.

CDU Smart Management Benefits

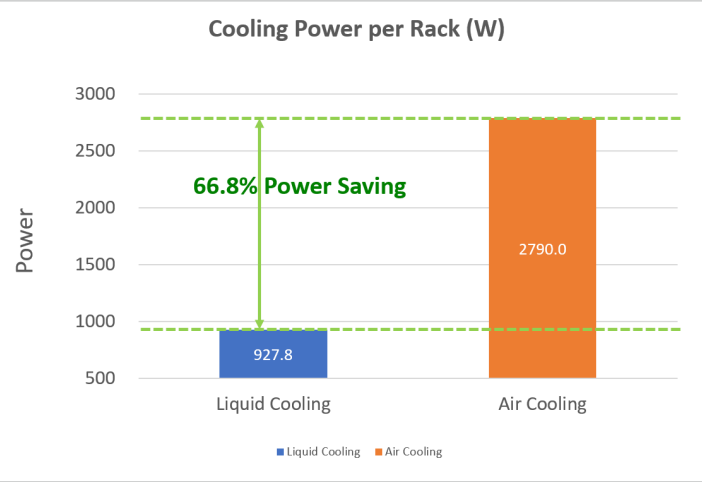
Server-grade DC-SCM in the CDU manage rear door cooling fans and coolant pumps. RDHx fan control via two flow zones dynamically balances coolant temperature, decreasing power usage for idle nodes.

Dynamic coolant adjustment from the CDU pump is based on the aggregated processors' workload. The smart commute includes auto detect of server BMC data and monitoring nodes for overheating and leakage values.

QCT QoolRack Liquid Cooling Thermal Test Result

Power savings are a major QCT QoolRack benefit. QCT performed a stress test under 35°C ambient temperature on two Intel 350W CPUs. The test found that the CPU temperature differences between a traditional air cooling solution and QCT QoolRack liquid cooling solution are remarkable. The test used servers with air cooling and QCT liquid cooling. The CPU temperature of the air cooled solution was 68.0°C, while that of QCT liquid cooling 56.2°C (11.8°C gap). Lowering CPU temperature can save more cooling power in individual server rack. During the test, liquid cooling consumed 927.8W, whereas air cooling consumed 2,790W cooling power per rack, resulting in a 66.8% cooling power saving on the QCT QoolRack liquid cooled server.

Figure 2. QCT QoolRack Liquid Cooling Power per Rack (W)



Benefits of QoolRack Liquid Cooling in Lowering Carbon Emissions

Today's prosperous human lifestyle comes from a large amount of energy consumption. Burning fossil fuels such as coal, oil, and natural gas produces carbon dioxide—one of the greenhouse gases that cause global warming. These carbon emissions left by human activities are called a carbon footprint. Saving power through cooling methods such as QCT QoolRack liquid cooling also reduces the earth's carbon footprint.

The QCT QoolRack smart management reflects a 66.30% carbon footprint saving based on Taiwan Power values.

Conclusion

A wide variety of organizations have large workloads that require infrastructure to run AI and HPC workloads. Data center infrastructure running AI and HPC workloads requires powerful microprocessor chips and the use of CPUs, GPUs, and acceleration chips to carry out compute intensive tasks. Data centers traditionally use air cooling solutions including heatsinks and fans. However, AI and HPC processing generates excessive heat, which results in higher data center power consumption and additional data center costs.

Traditional data center air cooling solutions may not be able to reduce energy consumption while maintaining infrastructure performance for AI and HPC workloads.

There are limitations to using traditional data center air cooling with fans and heatsinks when moving towards HPC or AI workloads. Liquid cooled systems will be increasingly replacing air cooled solutions for data centers running HPC and AI workloads to meet heat and performance needs.

QCT worked with Intel to develop the QCT QoolRack, a rack-level direct-to-chip cooling solution that meets data center needs with impressive cooling power savings per rack over air cooled solutions. It also reduces data centers' carbon footprint with QCT QoolRack smart management.

"QCT has a dedicated team devoted to solving thermal challenges related to running heavy computational workloads on our latest servers," said Mike Yang, President of QCT. "Because it's difficult to operate at maximum performance while keeping the temperature of silicon chips down, QCT developed its rack level liquid to air cooling approach to not only increase the efficiency of our own systems, but also to improve sustainability at the data center level with reduced OPEX"

For more information on QCT and how QCT QoolRack can help your organization, see: https://blog.qct.io/qct-qoolrack-solution-serves-as-the-latest-innovation-for-modern-data-centers