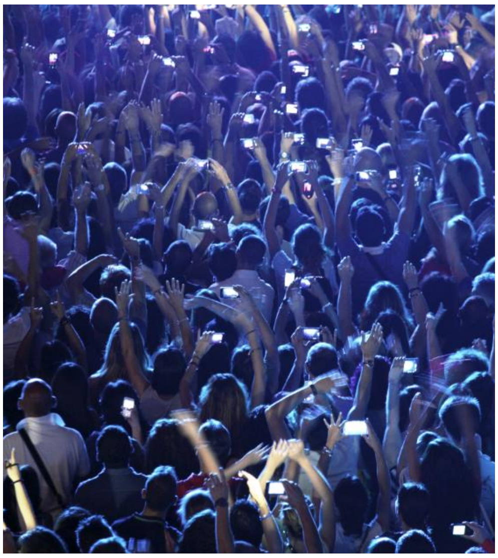
Rittal High Density Cooled-by-ZutaCore

White Paper No. 36
October 2021

Content

Content

Executive Summary



Technological innovation is one of the foundations of the improved quality of life of people around the globe. Smart speakers, smart refrigerators and watches that understand multiple spoken languages directly enhance the lives of everyday consumers. Augmented reality, predictive algorithms and AI-enabled controls enable unheard-of efficiencies in logistics, manufacturing and complex system maintenance. All of these recent innovations are built on the foundation of high-density computing.

High-density computing has two meanings: the first refers to the computing power required to provide the machine learning, augmented reality, etc., demanded by many applications; the second is that these computers literally require high levels of electrical power. Unfortunately, the conventional means for handling the heat generated by computers is just not able to keep up with the rapid densification at all levels.

A new approach, Direct-on-chip, two-phase, waterless liquid cooling (2PLC), can now be integrated into racks and data centers seamlessly and painlessly. 2PLC can support today's requirements while saving space and energy. In addition, it is practically unlimited in its ability to handle any high-powered processor that technological innovation may require in the future. Please read on to learn how 2PLC can remove the constraints imposed on innovation by other cooling technologies.

Christian Ludwig
Vice President Channel and OEM IT (Rittal)

Tim Shedd
Director Product Management (ZutaCore)

Challenges in Datacenter Cooling

Innovative Technology: Opportunities and Challenges

Figure 1: Autonomous Vehicles

Computing becomes more pervasive with each passing year. Innovative products continuously appear that take advantage of the ever-improving capabilities of computer processors. From smart refrigerators that can alert you - or the supermarket - when you are low on a staple food, to smart speaker systems that know your favorite music to listen to when making dinner, and that can talk to the refrigerator when you ask it, in plain language, whether you have any more eggs left. Autonomous vehicles are a reality and semi-autonomous vehicles are quite common. Businesses now apply advanced data analytics and machine learning to improve customer experience and maximize profits, while “fintech” is changing the world of finance and banking, opening up new avenues for individuals and corporations to invest their money precisely how they desire. High density computing is powering increasingly detailed and accurate weather predictions that save thousands of lives and billions of dollars in property damage, while artificial intelligence algorithms are able to identify abnormalities in medical images that even trained doctors can miss.

IT companies support this innovation with faster, bigger, more powerful chips. The intent is not just bigger and faster, though. These new processors provide non-linear improvements in computing capability, meaning that they are more efficient, offering more computation per watt of electricity they use. Still, the demand for computation is growing voraciously.

The computers used for everything from processing our spoken requests and providing detailed weather prediction, to the more mundane email and never-ending scrolls of social media content, are housed in data centers. Within these data centers can be found racks, typically about 0.6 m wide and over 2 m tall, that provide the physical support and electrical infrastructure to store and operate the computers.

Where in the past each rack would typically have computers requiring 6 kW to 10 kW of electrical power, the demand for more and more compute is driving this to 15 kW, 30 kW and higher. On top of this, functions that we now take for granted, like having a phone recognize our voice, or having up-to-the-second product inventory updates and predictions, require that many computers work together, communicating over high-speed systems that require the computers to be located near one another.

These trends – greater innovation leads to more powerful processors, leading to more electrical power in each rack of computers, ultimately leading to more power required per data center - all drive density: more computing power available in a given space, and more electrical power expended in that space. All of that electrical power going into the computers for the processing is turned into heat. And getting rid of all of this heat generated in a small space is now one of the key limitations on innovation and technological progress.

Getting the heat out

Traditionally, data center engineers used the technologies they were most familiar with to remove the heat from computers. If a room is hot, an air-conditioner can be used to keep it cool. The air-conditioner uses fans to push air across a coil, or heat exchanger, filled with an evaporating refrigerant, and the air gives up its heat to the refrigerant in the heat exchanger, coming out of the cooler. This cold air is directed into the data center, typically via a raised-floor ducting system, and air is supplied to vents sitting in front of each rack of computers. Fans within each computer draw the cold air in and blow it across the hot components, keeping them cool while warming up the air in the process. The warm air leaves the back side of the rack and is pulled back to the air-conditioner by fans in the air-conditioning unit.

Figure 2: CRAC based datacenter cooling

A data center air-conditioner is often called a CRAC unit (Figure 2), short for Computer Room Air-Conditioner. A similar strategy that works much like that described above uses a refrigerated water chiller that pumps cold water into Computer Room Air Handlers, or CRAHs.

These CRAC and CRAH units are not like the air-conditioners in a house or a commercial building; the computer chips are sensitive to temperature, and they can go from generating just a few watts of heat to 100’s of watts in milliseconds. Thus, the CRAC and CRAH units have sophisticated controls and are typically required to blow enough air continuously to cool the worst-case computer load, even if many of the computers are in idle.

Using air to cool computers appears to be a convenient choice. It is “free,” and can be distributed easily to 1000’s of computers with just fans. However, air is not a good fluid for moving heat around. This is because it absorbs and releases heat energy via sensible heating or cooling, which means it must change its temperature, increasing temperature when it absorbs heat, and decreasing temperature when it releases heat. The amount of heat that can be absorbed by the air depends on the amount of air that flows past a hot component. The physical law describing this (the First Law of Thermodynamics), can be shown in the first equation on the right.

($\dot{q}_{air}$, *Heat energy per second absorbed or released*) = ($\dot{m}_{air}$, *mass flow of air*) x ($c_{p,air}$, *specific heat capacity of air*) x (ΔT_{air}, *the temperature change*)

$$\dot{q}_{air} = \dot{m}_{air} c_{p,air} \Delta T_{air}$$

As a detailed example, consider air flowing through a server at 0.05 m3/s (180 m3/hr, 106 cfm). It enters a computer server at 25°C, which means that the density of air is 1.18 kg/m3 and the specific heat capacity is 1021 J/kg-°C. If the server generates 500 W of heat, the temperature difference of the air is given in the second equation on the right and thus the air leaves the server at 33 °C.

$$\Delta T_{air} = \frac{\dot{q}_{air}}{\dot{m}_{air} \cdot C_{p\,air}} = \frac{500W}{\left(1.18\frac{kg}{m^3}\right)\cdot\left(0{,}05\frac{m^3}{s}\right)\cdot\left(1021\frac{J}{Kg\,K}\right)} = 8{,}26K$$

The challenge comes with new processors that are exceeding 400W per package, like high-end, general-purpose CPUs, as well as the GPUs and FPGAs that accelerate the machine learning and artificial intelligence applications that are the foundation of so much technological innovation. Not only do we have to get enough air to the computer servers, we have to focus that air on the processors themselves and the air has to be cold enough to cool multiple processors in a line. Further analysis with the First Law of Thermodynamics shows that for servers with multiple 400+ W processors, the air may need to reach literally hurricane-speeds and be quite cool (around 10 °C or 50 °F). In most cases, this is just too noisy, too expensive and too impractical to do.

So why be limited to air? For years, computer gaming enthusiasts have used water pumped through tubes to cold plates mounted directly on the processors to beat the thermal limitations of air. The First Law of Thermodynamics says that, for a 400 W processor, the ΔT_{water} is given in the equation on the right for a flow of water directly to the chip at 1 lpm. So, a 3000x lower flow of water can cool more effectively than air.

$$\Delta T_{water} = \frac{\dot{q}_{water}}{\dot{m}_{water} c_{p,water}}$$

$$= \frac{400\ W}{(997\frac{kg}{m^3})(1.67\times10^{-5}\frac{m^3}{s})(4180\frac{J}{kg\,°C})} = 5.8\ °C$$

On the other hand, as most computer gaming enthusiasts can sorrowfully attest, a single drop of water in the wrong place on a computer circuit board can mean permanent damage to very expensive processors or other components. And, over time, water is corrosive to many materials.

An effective, low maintenance solution

At this point, it is clear that cooling with air is low risk and convenient, but it is expensive to use when high heat loads are present and limited by the First Law of Thermodynamics as we move to the future. Pumping water through cold plates attached to processors has several advantages, but also carries significant risk and significant maintenance costs. What are the alternatives?

A number of different fluids exist that might be the answer. These fluids, designed for heat transfer applications, are dielectric, meaning they don't conduct electricity and won't harm any electronics that come into contact with them. They are also inert, non-corrosive, non-toxic and non-flammable. Almost perfect. But they usually have a specific heat capacity close to that of air due to their chemical structure. The First Law of Thermodynamics says that if the heat capacity is 4x less than water, you will need about 4x more flow rate to achieve the same heat transfer performance, and this presents a potential limitation for these otherwise great heat transfer fluids.

Now, imagine that one could eliminate this heat transfer limitation. That we could find a fluid with the great qualities of dielectric fluids, but with a heat capacity that was very high. Such a fluid could absorb a lot of heat without increasing in temperature very much. Even better, could we find a fluid with infinite heat capacity? A fluid that could absorb as much heat as we put into it and its temperature would not change. This is actually possible if we design a system to take advantage of evaporation (the so-called latent heat of the fluid) rather than warming up the fluid (sensible heating). An evaporating fluid has a specific heat capacity that is essentially infinite ($c_{p,evap} \rightarrow \infty$). Then, as shown in the example calculation on the left, the temperature of the fluid does not change at all.

$$\Delta T_{evap} = \frac{\dot{q}_{evap}}{\dot{m}_{evap} c_{p,evap}}$$

$$= \frac{1000\ W}{(1400 \frac{kg}{m^3})(1.67 \times 10^{-5} \frac{m^3}{s})(\infty \frac{J}{kg\ °C})} = 0\ °C$$

This technology is called Direct-on-chip, two-phase, waterless liquid cooling or 2PLC. It is highly effective at managing heat loads from small to large, allows for light, compact design and high densities, has no chance of damaging IT hardware, and the dielectric fluid requires no monitoring or maintenance. And it is available today.

Direct Chip Cooling

In-Rack System Overview

The Direct-on-chip, two-phase, waterless liquid cooling (2PLC) technology is available as a self-contained, in-rack system, alleviating cooling boundaries at the chip, server, rack, POD and data center levels as shown in Figure 3.

The in-rack solution uniquely supports up to 15 kW computing power with an in-rack air-cooled condenser and 70 kW computing power with an in-rack water-cooled condenser. In-rack units can scale in parallel into multiple rack units, which are supported by in-row condensers.

The waterless 2PLC system brings a distinct combination of self-regulated, on-demand, closed-loop, low-pressure cooling in a well-integrated solution. The system is densifying data centers from small to large scale, for both data center retrofits and new builds.

This plug-and-play solution triples computing densities in a fraction of the footprint, consistently performing at the lowest PUE and highest efficiency, in any climate. The result – the data center shrinks and scarce energy, water, land and construction resources are saved.

2PLC hardware comes with a Software-Defined-Cooling (SDC) platform that enables remote monitoring of the cooling system and its status. In addition, the SDC platform provides a family of upgrades that can provide much more detailed monitoring of the thermal state of some or all of the processors in the data center, as well as providing optimized processor utilization to minimize power consumption.

For example, the SDC's dynamic frequency scaling reduces server power consumption.
Based on the knowledge that power required by a processor is exponentially related to the working clock speed, management of the CPUs in the servers is based on the "race to halt" method

Figure 3: In-rack system

at highest clock speed or idle. In data center design, servers run at full capacity only during peak demand, but for the majority of the time, servers are at 40-60% load.
The dynamic frequency scaling, real-time algorithm, evaluates the workload on the server and calculates the minimum CPU clock speed that will provide the computing power needed to support the work load. This in turn significantly lowers power consumption by up to 20%.

By reducing the data center design to commissioning cycle and halving costs in many cases, the 2PLC system empowers data center owners and operators to accelerate return on investments and maximize real estate assets by supporting the proliferation of autonomous and central data centers. With increasingly powerful CPUs and GPUs pushing the envelope into multi-hundred watts, the 2PLC system addresses chip and server-level hot spots and is well suited for edge computing requirements.

System Components

The Enhanced Nucleation Evaporator (ENE)

At the core of the DCEC system is the ENE, enabling a single, closed-loop, two-phase cooling solution that yields unparalleled heat dissipation on-chip. Unlike water-based solutions that expose the risk of permanent IT damage, the system utilizes a safe, non-conductive, commercially available liquid refrigerant. It is part of a complete hardware system, enhanced by a Software-Defined-Cooling platform.

Figure 4: Enhanced nucleation evaporator

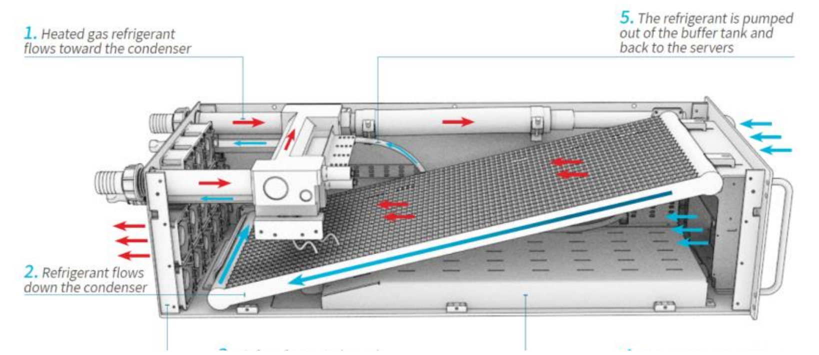
The ENE, a pool-boiling evaporator device, provides on-demand heat transfer response and enables the scalability of the system to cool all chips in the data center in parallel, Figure 4.

The ENE's self-regulation supplies exactly the amount of refrigerant needed to cool the heat generated by each component. The generation of vapor removes the thermal energy efficiently from the devices, which then transports the energy to the condenser where it is removed from the rack via a flow of air or water.

The Heat Rejection Unit (HRU)

The HRU (Figure 5) is an integrated, automated package that includes intelligently controlled pumps, a buffer for storing refrigerant, and a heat exchanger for rejecting heat to either air or facility water. The controller automatically maintains the optimum temperature for the refrigerant to ensure the most efficient removal of heat from the servers.

Figure 5: Heat Rejection Unit



It contains and monitors several sensors used for this control function and also to provide early warnings if preventive maintenance might be necessary. In tandem with the Software-Defined Cooling package, the HRU can provide early warning of leaks throughout the system.

Leaks, of course, are unlikely due to the rigorous quality and reliability programs integrated into the design and manufacturing processes. In addition, other key reliability features are incorporated, such as N+1 redundancy on the only moving part, the pump, automatic A/B power supply switchover, complete power failure ride-through, etc.

Figure 6: Refrigerant Distribution Unit

Figure 7: Plug & play solution

The Refrigerant Distribution Unit (RDU)

The RDU (Figure 6) is a passive device, about the size of a vertical rack PDU, that installs in an IT rack to ensure efficient distribution of refrigerant to the ENEs as well as to collect the vapor they generate. State-of-the-art, dripless quick connects ensure simple maintenance of servers should they need to be removed from the rack. The RDU is also available with low-force engagement blind-mate quick connects for further efficiency and higher level rack integration.

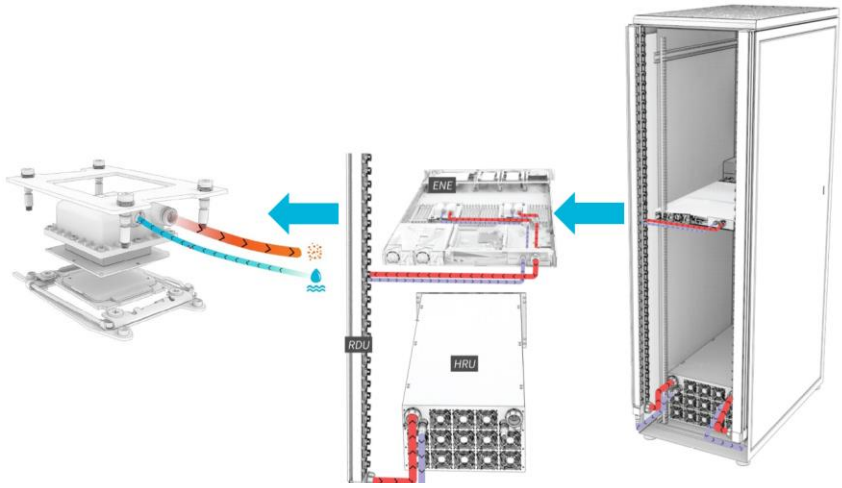
The 2PLC system

Figure shows the entire, integrated plug & play solution mounted within a rack. The ENEs of the various processors are connected via the RDU mounted at the rear sidewall of the IT rack. The condenser is a 6U unit installed in the 19" IT rack or the 21" OCP rack. Figure 8 explains the basic operation principle.

The operation sequence of the system consists of the following steps:

- The liquid refrigerant flows into the ENEs of the various CPUs / GPUs Inside the different servers.
- The refrigerant evaporates by absorbing heat from the CPU / GPU.
- The RDU collects the refrigerant vapor and directs it to the condenser.
- The refrigerant vapor condenses back to a liquid in the condenser.
- The HRU pumps ensure that the liquid refrigerant flows again via the RDU into the cooling system.

Figure 8: Basic operation principle



The Refrigerant

Although the 2PLC system is not necessarily restricted to one type of refrigerant, the dielectric fluid from 3M Corporation, 3M™ Novec™ 7000 Engineered Fluid, is a great match to IT cooling applications. The refrigerant boils at 34°C in atmospheric pressure and allows the ENEs to leverage latent heat (phase change) to provide 90% reduction in liquid pumping compared with a water-cooled system, providing superior heat transfer performance.

Novec™ 7000 Engineered Fluid is non-toxic, non-corrosive, non-flammable and easy to work with. It does not support biological growth like water does, it does not break down with use and it never needs to be treated beyond passing through the built-in filter within the HRU. Like other common refrigerant systems (i.e., air-conditioners and refrigerators), the system is closed and, under normal operation, refrigerant does not need to be added. However, should maintenance require it, adding refrigerant is a simple matter.

Integration into a rear door cooling system

Figure 9: RHx integrated system

The rear door of a standard rack or OCP rack is replaced by a heat exchanger (RHx - Rear Door Heat Exchanger), as shown in Figure 9. The RHx consists of fans and the air-cooled condenser itself. The fans ensure that the required air volume is moved through the rack - from front to back - transporting the heat out of the cabinet. The First Law of Thermodynamics can be used to calculate the air volume if the heat load of the servers is known.

In parallel, the air stream through the cabinets cools down the rest of the active IT components not served by the 2PLC cooling system. It cools these components so that the warmer air is used to condense the refrigerant vapor in the condenser.

The benefits of the Rear-Door-Air approach:

- Minimal additional space required (the RHx door is slightly thicker compared to a standard rear door)
- No additional cooling system required to remove the heat load of servers
- The additional heat load is transported in two steps out of the data center:
    - 2PLC system from CPU to RHx
    - RHx in conjunction with direct or indirect free cooling from rack to the outside ambient

Perhaps the greatest benefit is this: with the air-cooled RHx, high-density compute for today's most advanced technological innovations can be deployed in existing, air-cooled data centers with no modification whatsoever to the data center infrastructure. Cutting cooling power infrastructure needed from the server to the data center is as easy as rolling in a new rack or replacing the rear door of an existing rack.

Integration into an inline cooling system

If a significant amount of new, high density compute needs to be deployed, the data center cooling infrastructure will be impacted. Thus, when considering whether to deploy the elegant RHx system, another elegant solution may be preferable. In this application, the additional heat load is conducted out of the data center by the warm water (or by the hot vapor of the refrigerant, respectively). Outside the data center, the water must be cooled down again (or the refrigerant liquefied again), but dry coolers can be used in most environments, eliminating water-consuming cooling towers.

Like traditional inline coolers, the in-rack solution provides cooled air to the rack to remove the heat from the computers that does not come from the processors, which are all cooled with ENEs.

The benefit of the Inline approach:

- Easily couples to a data center cooling water distribution system.
- No impact on the air-cooling within the data center (it is designed to be room-neutral).
- The inline system is able to deal with very high heat loads where it would otherwise be difficult to drive an adequate volume of air through a server rack.

Figure 10: Rack integrated system

Integration into the rack itself

When it is not possible to use any additional data center space, the high-density 2PLC solution can be integrated right into the rack (Figure 10) to support technological advancements that high-density compute can bring.

The benefits of the in-rack HRU include:

- Multiple connection options to RDU or blind mate connectors, pre-assembled for easy installation
- Easy access and no refrigerant piping above or below the racks
- Easily mountable in 19” or wider racks, it is 6U tall by 40” deep

The air-cooled HRU handles a load of 20 kW, with a partial PUE of 1.03. By using water to transport the heat out of the rack, a cooling capacity of 70 kW is provided at a pPUE of 1.1.

The partial PUE is defined as the ratio of "total power in one section of a data center: $P_S(DC)$" and " total power of the IT devices in this section $P_S(IT)$". Hence, the pPUE covers only the rack - but with all components (server, 2PLC system and HRU).

Partial Power Usage Effectiveness:
$pPUE = P_S(DC)/P_S(IT)$

Keeping innovation moving forward with 2PLC

If we are to continue to see exponential innovation in technology that enhances the human experience globally, new, higher density compute will need to be deployed in data centers large and small around the world. Unfortunately, the conventional approach of air-conditioning these computers will just not be able to handle the load and are quite inefficient even with current levels of heat load.

Direct-on-chip, two-phase, waterless liquid cooling (2PLC) provides a low maintenance, extremely high density, sustainable cooling solution that enables innovation today and is ready for any high-powered processor that IT innovators may think up in the future.

Are you ready to innovate? Do not be constrained. Contact Rittal today for integrated solutions that will enable you to deploy seamlessly and painlessly the compute you need, where you need it.
To learn more about Rittal High Density Cooled-by-ZutaCore solutions please call or email us: info@rittal.de or info@zutacore.com.

List of images

List of images

Glossary

2PLC	Direct-on-chip, **two**-**p**hase, waterless **l**iquid **c**ooling: the heat source of a server (CPU) is directly contacted to eliminate the waste heat by using a non-conducting fluid.
CPU	**C**entral **P**rocessing **U**nit: central computing element Inside a server or personal computer
GPU	**G**raphics **P**rocessing **U**nit: a CPU designed for the calculation of graphical vectors and elements
CRAC	**C**omputer **R**oom **A**ir **C**onditioner: a unit providing chilled air through a raised floor or aisle containment. The cold air Is generated by an appropriate heat exchanger with CW or DX technology.
CW	**C**hilled **W**ater: Indirect cooling of a data center, where cooling is provided using chilled water.
DX	**D**irect E**x**pansion: The cold is transported through a phase transition of a refrigerant into the data center.
ENE	**E**nhanced **N**ucleation **E**vaporator: this component replaces the traditional heat exchanger of a CPU or GPU by a direct contacting liquid.
HDC	**H**igh **D**ensity **C**omputing: servers systems dedicated for high computing power (super computer, AI nodes, blade units, …)
HRU	**H**eat **R**ejection **U**nit: compact DCEC unit for smale scale applications (6U unit)
IIoT	**I**ndustrial **I**nternet **o**f **T**hings

IoT — **I**nternet **o**f **T**hings: All active components from machines to products are assigned an IP address that they can use to communicate. In this way, data can also be exchanged between machines and between computers without people being involved. The term Internet of Things is derived from this.

LCP — **L**iquid **C**ooling **P**ackage, available as CW (30 kW, 55 kW) and DX (12 kW, 20 kW, 35 kW) for rack and row based climatization

POD — **P**oint **o**f **d**elivery: the combination of several racks including all infrastructure components to serve as a host for active IT devices (server, network, storage)

RHx — **R**ear Door **H**eat E**x**changer: a passive heat exchanger is Installed Instead of a rear door. The exhaust air from the servers is moving through the door and cooled down. The entire data center hall is at low temperature. The hot aisle is reduced to a small space at the backside of the rack in front of the RHx.

SDC — **S**oftware-**d**efined-**c**ooling: Management suite controlling the entire cooling system in order to balance between compute load and cooling capabilities

S-RDU — **S**mart **R**efrigerant **D**istribution **U**nit: Cooling fluid distribution and collection system between the condenser and the various ENEs Inside the different server.

You can find the contact details of all
Rittal companies throughout the world here.

www.rittal.com/contact

RITTAL GmbH & Co. KG
Auf dem Stützelberg · D-35745 Herborn
Phone +49 (0)2772 505-0 · Fax +49 (0)2772 505-2319
E-mail: info@rittal.de · www.rittal.com