Emergence and Expansion of Liquid Cooling in Mainstream Data Centers

White Paper Developed by

ASHRAE Technical Committee 9.9, Mission Critical Facilities,
Data Centers, Technology Spaces, and Electronic Equipment

ASHRAE®

Peachtree Corners

This white paper was developed by
ASHRAE Technical Committee (TC) 9.9, Mission Critical Facilities, Data Centers,
Technology Spaces, and Electronic Equipment.

www.ashrae.org/datacenterguidance
http://tc0909.ashraetcs.org

ASHRAE has compiled this publication with care, but ASHRAE has not investigated, and ASHRAE expressly disclaims any duty to investigate, any product, service, process, procedure, design, or the like that may be described herein. The appearance of any technical data or editorial material in this publication does not constitute endorsement, warranty, or guaranty by ASHRAE of any product, service, process, procedure, design, or the like. ASHRAE does not warrant that the information in the publication is free of errors, and ASHRAE does not necessarily agree with any statement or opinion in this publication. The entire risk of the use of any information in this publication is assumed by the user.

Contents

Acknowledgments

The ASHRAE TC9.9 committee would like to thank the following people for their work and willingness to share their subject matter knowledge to further the understanding of liquid cooling as applied to servers deployed into data centers.

- Paul Artman—Technical Director, Advanced Micro Devices, Raleigh, NC; Corresponding Member ASHRAE
- Tozer Bandorawalla, PhD—Intel, Hillsboro, OR; Corresponding Member ASHRAE
- David Grant, PE, CEM, DCEP—Oak Ridge National Laboratory, Oak Ridge, TN; Corresponding Member ASHRAE
- John Gross, PE, ATD—J.M. Gross Engineering, Richmond, TX; Corresponding Member ASHRAE
- Jason Matteson—Director of Product Strategy, Iceotope, Raleigh, NC; Corresponding Member ASHRAE
- David Moss (co-lead)—Dell, Austin, TX; Corresponding Member ASHRAE
- Chris E. Peterson—Technical Staff, Dell ISG, Round Rock, TX
- Suresh Pichai, PE—Director, Equinix Global Engineering, Simpsonville, SC
- Mani Prakash, PhD—Senior Principal Engineer, Intel, Austin, TX; Corresponding Member ASHRAE
- Matt Shumway—Technology Development, Office of the CTO, Seagate, Shakopee, MN
- Mark Steinke, PhD (co-lead)— Senior Manager, Data Center GPU and Accelerated Computing Thermal Mechanical Engineering, Advanced Micro Devices, Austin, TX; Corresponding Member ASHRAE

A special thank you to the reviewers of this paper: Roger Schmidt, Dustin Demetriou, Dave Kelley, John Groenewold, Terry Rodgers, John Bean, and Gerardo Alfonso.

Objective of the White Paper

The IT industry is at a performance inflection point. When purchasing a new computing device, it is expected to be more powerful than the previous generation. As for servers, over the last decade the industry had a period where significant performance increases were delivered, generation over generation, accompanied by modest and predictable power increases. That period ended around 2018. Large power increases in the compute, memory, and storage subsystems of current and future IT equipment are already challenging data centers, especially those with short refresh cycles. The challenges will only increase. Liquid cooling is becoming a requirement in some cases, and should be strongly and quickly considered. This paper explains why liquid cooling should be considered, rather than the details around what liquid cooling is or how to deploy it.

Introduction

Large increases in IT equipment power will require additional equipment energy use and cooling resources will result in fewer servers per rack. During the 1990s and early 2000s, IT equipment power draw increased regularly. At the time, nameplate power was the typical planning metric, so a refresh may not have been that problematic. This paper will address three time frames: the early time frame where power increases were acceptable, the period following where power remained relatively constant, and the current time frame where power draw is again on the rise. Now that more accurate power levels are the data center planning metric, there is no longer a comfortable margin of power and cooling overprovisioning resulting from the use of the nameplate metric.

Liquid-only processor chips are currently available, and more are coming in the near future. There are many who wish to put off the introduction of liquid cooling into the data center due to its cost and complexity. There are data center impacts associated with the continued push for higher and higher levels of air cooling. These impacts are likely mitigated by liquid cooling and should be a point of consideration whether or not the data center is considering using liquid-only chips. One of the unintended consequences of increasing chip power is the need to reduce the case temperature. The case temperature, sometimes called *lid temperature*, is the temperature on the top surface of the chip, typically at its center; this is often referred to as *Tcase*. Chip vendors characterize the relationship of the case temperature, an externally measurable location, to the critical internal chip temperatures. Tcase is used during the IT equipment thermal design process to ensure the chip is adequately cooled. With case temperatures decreasing in the future, it will become increasingly harder to use higher ASRHAE classes of both air and water. It is recommended that any future data center includes the capability to add liquid cooling in the design of the data center.

Change to ASHRAE Water Classifications

There has been some confusion over the ASHRAE water classifications. As originally written, servers did not necessarily have a clearly defined compliance obligation over the range of temperatures within the classifications. For simplicity, the W classes are being renamed with the upper temperature limits incorporated in the name. All classes will carry a lower temperature limit of 2°C (35.6°F). The classes are newly named as follows: W17 (previously W1), W27 (W2), W32 (W3), W40 (new), W45 (W4), and W+ (W5). Server capability has been strengthened by adding the following: "operating within a particular environmental class requires *full performance* of the equipment over the entire environmental range of the specified class, based on nonfailure conditions." This change is included in the fifth edition of *Thermal Guidelines for Data Processing Environments* (ASHRAE 2021), released in 2021. In addition, the upcoming third edition of the *Liquid Cooling Guidelines for Datacom Equipment Centers* (ASHRAE 2013) will also include the update.

Motivations

Realizing the drastic increases in device power facing the industry, the authors set out to write this paper as a warning of the impending challenge. Responding to the need for real-time data analytics, chip manufacturers are responding with higher-frequency parts, raising thermal design power while lowering package case temperature requirements. There will be an increasing number of device inventory or stock keeping units (SKUs) that require liquid cooling. With multicore parts reaching performance limits, the only option for increased performance is increased power. Expect an increased shift from off-roadmap SKUs to on-roadmap SKUs in the future. Also, expect the power levels of the entire on-road SKU stack power to creep upward.

Socket Power Increases and Case Temperature Reductions

Higher performance demands placed upon servers can be seen in the form of increasing power for the central processing units (CPUs), graphic processing units (GPUs), and field programmable gate arrays (FPGAs). The increasing performance demands and power trends can now be seen across all processing technologies. To generalize the power demand placed upon these devices, the term *socket power* will be used to describe the power requirement for that device, whether it be CPU, GPU, or FPGA. In this way, the processing technology can become independent and the cooling requirements need only be related to that processor, or *socket*.

To describe the cooling solution requirements, a commonly used metric called thermal resistance (Ψ_{ca}) is used to represent the effectiveness of a thermal solution. As a heat sink or a cold plate is used to capture heat from a device, the thermal resistance represents the temperature difference between the processor case and the cooling medium, divided by the device power, typically in °C per watt. The lower the thermal resistance value, the more effective the cooling solution. Figure 1 shows the thermal resistance required to cool past, current, and future socket solutions from the processor manufacturers (IBM, Sun, Intel, AMD, NVIDIA, ARM). Only data that has been released is included in the dataset. The highest available processor power is used as the reported socket power and the case or junction temperature for that processor is also used to determine the thermal resistance. The thermal resistance, represented roughly by the dashed polynomial trend line in the figure, represents the cooling requirement for the socket itself and does not account for upstream components preheating the air prior to

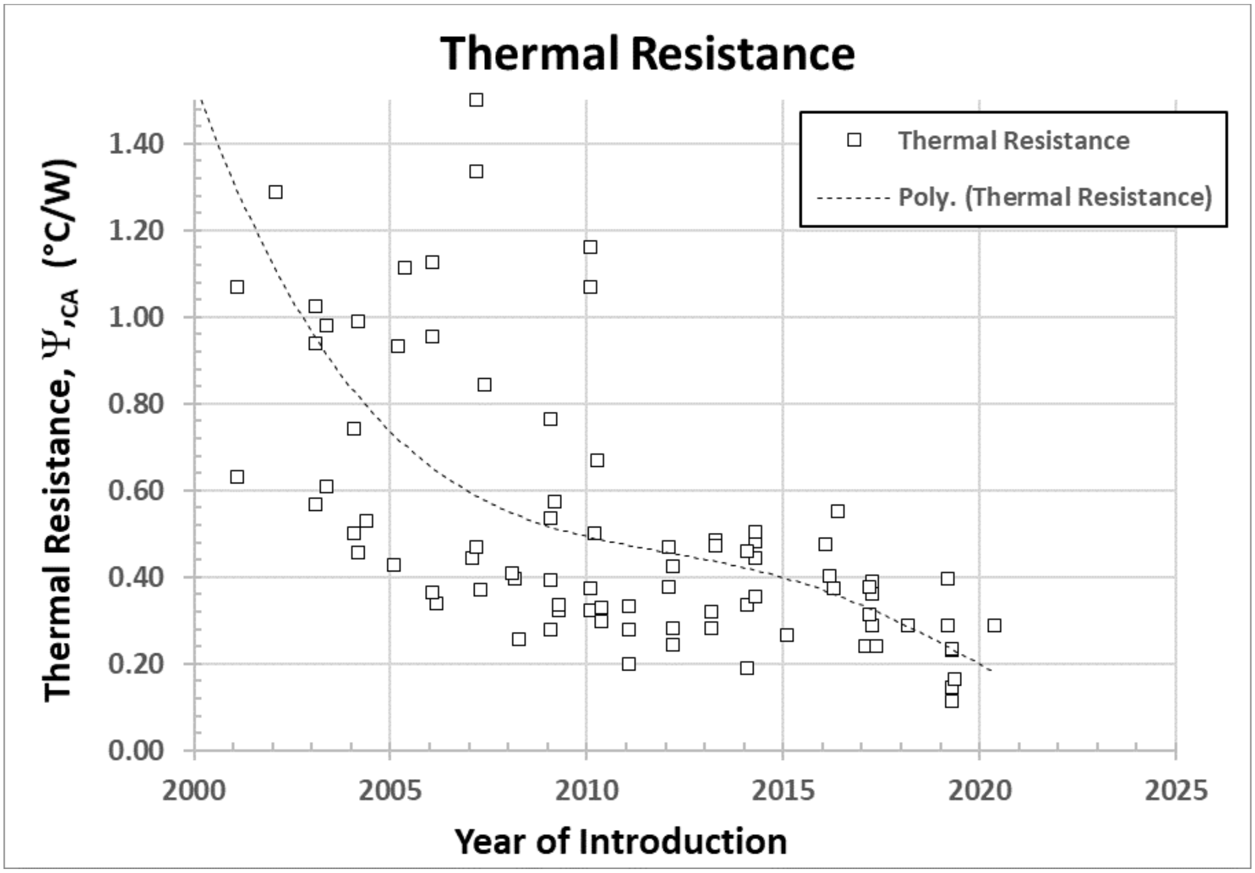

Figure 1 Thermal resistance required to cool socket power.

reaching the socket. The declining trend in thermal resistance over time is clearly displayed. As socket power goes up, so does the temperature difference between the device and the cooling medium, whether liquid or air. As a result, improvement in cooling performance is required, which forces a reduction in cooling inlet temperature or a shift to different cooling technologies. This trend is driving a shift from air to liquid cooling technologies, as liquid cooling is many times more effective at removing heat than air. Hence, the thermal resistance needed to cool the current higher power devices is much lower than it was 10 or 20 years ago. The thermal resistance relates the socket power and maximum case temperature limit to the fluid temperature whether the fluid is air or single- or two-phase immersion.

Another way to look at the data is to express the inverse of thermal resistance, or the *degree of cooling difficulty*. Expressed as the inverse of thermal resistance, Figure 2 shows clear trends over three time periods. In the first period, 2000 through 2010, performance increases were enabled by increasing device power. The industry was very close to needing liquid cooling at the end of this period. Given the looming tipping point or need to shift to liquid cooling in or around 2010, chip manufacturers managed to avoid this transition through a fundamental and architectural design shift. This is shown in the second time frame, the multicore period, where performance increases were enabled by increasing the number

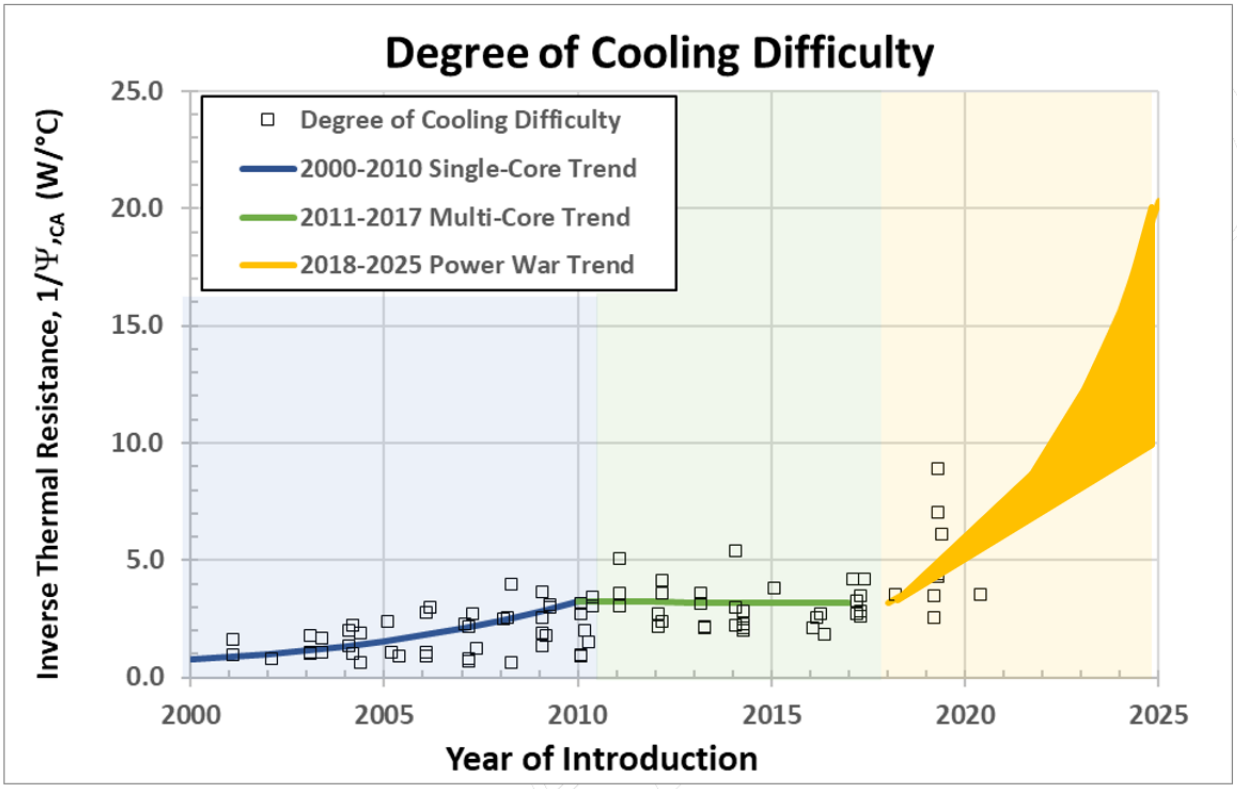

Figure 2 Degree of cooling difficulty for socket cooling, $1/\Psi_{ca}$.

of cores while spreading the same total power amongst more sources within the chip. From around 2018 to the present, the benefits and advantages of this multicore period diminish. Devices are no longer spreading power amongst more cores. Device power must increase to enable further performance increases. In addition, many software workloads can use multithread processing that allows for further use of socket thermal design power (TDP). Add to that, related case temperature decreases only serve to steepen the slope as the denominator (temperature) trends down and the numerator (wattage) trends up. This period might be thought of as the “power wars,” as device vendors try to outperform their competitors.

Figure 3 presents the same degree of cooling difficulty as in Figure 2, but now in a log scale. The exponential trends will appear linear in this view and the three period trends can be more easily identified. As noted previously, only data that is released and/or publicly available is represented in this figure. Any data that would be deemed unreleased is used in formulating the trendline but not represented in the graph. The current trendline again increases at a faster rate than during the first period. The trend is clear that higher socket powers with lower Tcase requirements are present in the upcoming generations of processors and will hasten the adoption or expansion of liquid cooling.

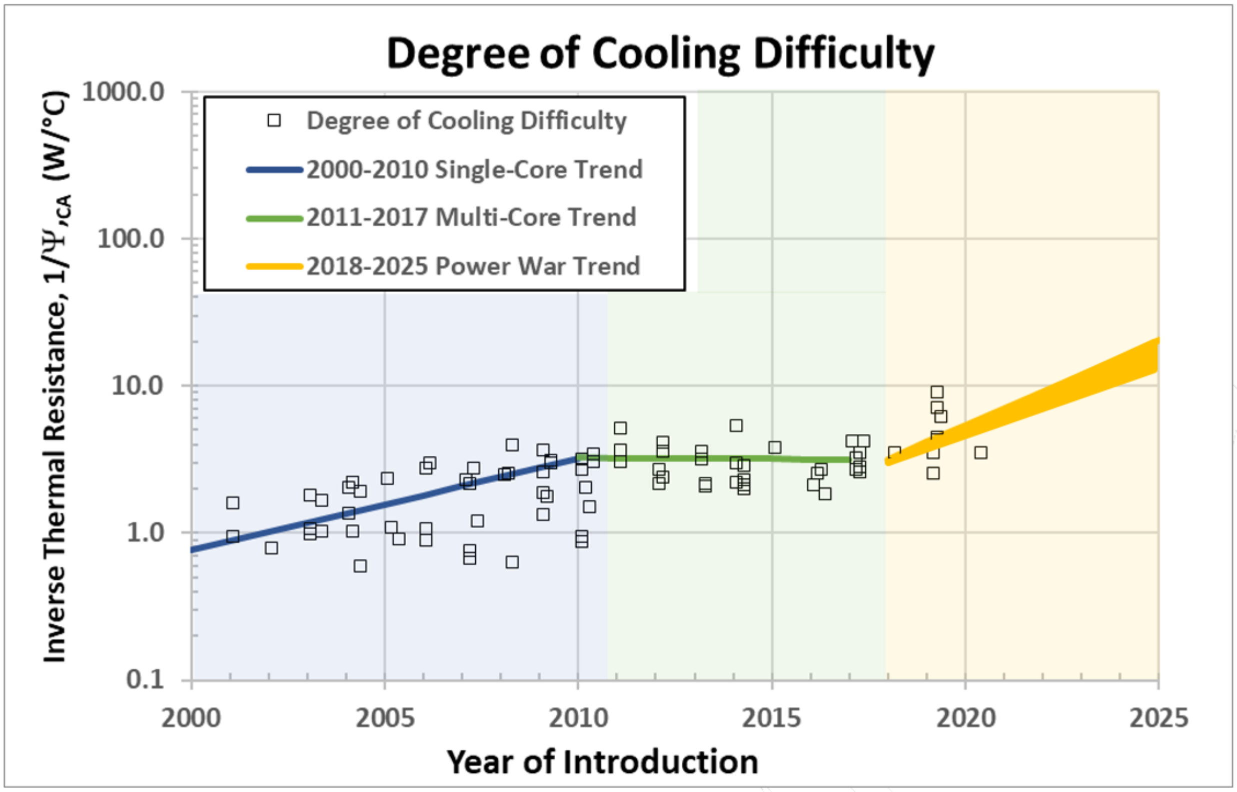

Figure 3 Degree of cooling difficulty for socket cooling, 1/Ψ_{ca}, in log scale.

Memory Power Increases

The memory subsystem is rapidly emerging as a cooling challenge as well. As the memory speeds increase and the number of dual in-line memory modules (DIMMs) per socket increase, the cooling challenge is compounded by an increase in the power density of components on the DIMM as well as an increase in the total power of the memory subsystem. Depending upon the server layout, in some cases where memory is upstream of processors or GPUs, the increased power also causes a high degree of preheat, increasing the cooling demands placed upon the processor. Liquid cooling is moving from an enhancement to a requirement for high-capacity memory given the significant increase in power for fifth generation double data rate memory technology (DDR5), expected to be introduced in 2021, as frequency scales to meet customer workload requirements. For previous memory technology transitions the memory voltage could be reduced to offset frequency increases, but this is no longer possible.

Data Center Impacts

Server manufacturers will stretch air cooling as far as possible, but that has implications. To keep advancing air cooling, the server simply needs more airflow. If done within the same size enclosure, that will be an increase in airflow per U, which has already surpassed the ability of providing conditioned air to the cold aisle in many cases. Increases in flow rate often are accompanied by increases in acoustical noise emissions. This has already been noted as problematic in the ASHRAE white paper “Hard Disk Drive Performance Degradation Susceptibility to Acoustics” (ASHRAE 2019). Increase in raw fan power may be objectionable. The server industry has done a great job reducing fan power as a percentage of server power, from levels as high as 20% down to as low as 2% in some cases. Continuing to use air cooling driven by these higher power components will reverse this positive trend of lower fan power.

Impact of Increased Airflow Per Rack

First mentioned in ASHRAE Datacom Book 13, *IT Equipment Design Impact on Data Center Solutions* (ASHRAE 2016), it was said that best-of-breed data centers could deliver 1900 cfm per floor tile. There are servers on the market today that require 100 cfm or more per U. If relying solely on a single tile in the raised floor, that would translate into only occupying 19U within a rack. Of course, there is supplemental cooling in the form of overhead or row cooling and one could consider replacing or supplementing raised-floor cooling with a closely coupled coil such as a rear door heat exchanger. Many data centers, however, gradually depopulate IT equipment as they refresh their racks to fit within legacy power or cooling levels.

Impact of Specialty Servers with Reduced Inlet Temperature Requirements

Many of the 100 cfm per U servers referenced above are extremely dense, resulting in temperature limitations of ASHRAE classes A1 or A2. To meet these environmental temperature requirements, these products may limit the number and type of components designed into the server. For example, the server might be an A2 product with very few drives, without a full GPU complement, or with lower power CPUs; however, when one purchases this type of server because of its density capability, it can be rated to no more than 25°C or 30°C. This might increase the challenge of positioning these servers and requiring them to go into a colder portion of the data center. As part of *Thermal Guidelines for Data Processing Environments*, *Fifth Edition* (ASHRAE 2021), a new H1 class is created for high density, air cooled IT equipment with an allowable range of 15°C to 25°C (59°F to 77°F)and a recommended range of 18°C to 22°C (64.4°F to 71.6°F). To be clear, the intent of the H1 class is to cover existing and future products that, when modestly to fully loaded, cannot meet the ASHRAE A2 temperature range and are sold today by the server original equipment manufacturers (OEMs) as lower-temperature products.

Fan Power as a Percentage of Server Power

The IT manufacturers represented within TC9.9 have spent considerable time over the last decade repeatedly saying that fan power is not 20% of a server's power. At one point, it was. Particularly when servers used constant-speed fans, it was common to waste 20% of the total server power on fan power. The industry evolved and began using variable-speed fans with rudimentary control, offering some relief in fan power. Servers have progressed with further optimization using very complex algorithms acting on feedback from many sensors that tune fans individually to no more airflow than is needed. During the multicore period, fan power decreased to below 2% in many cases. Responding to the steep challenge seen in the "power wars" shown in Figure 3, fan power is on the rise again—exponentially. A fan power percentage of 10% to 20% is not uncommon for some of the denser servers. Interestingly, it is not always a hot chip triggering higher fan speed. Quite often, the fans must speed up due to excessive preheat violating rear rack air temperature design. Like the components on the circuit boards, rear-mount electronics such as power supplies and active optical connectors also have a maximum inlet temperature. Those trying to hold on to air cooling, putting off a move to liquid, might consider total cost of ownership (TCO) implications of 10% to 20% fan power inside the server. In a 50 kW rack, the fan power translates to be at least 5 kW. Additionally, it is important to acknowledge server cooling fans are powered by the same uninterruptible power supply (UPS) source as the server itself. As such, retaining air-cooled IT equipment while server fan power increases from 2% to 10% of the total server power equates to reducing the data center UPS capacity by 8%.

Acoustical Impact of Increased Power and Airflow

Acoustical output accompanies airflow increases and thus has repercussions for human interaction in addition to performance of rotational media. Data center and individual product sound pressure levels already reach levels that require U.S. Occupational Safety and Health Administration (OSHA) consideration for hearing protection. Further airflow increases will only exacerbate this safety issue. Rotational media continue to be affected by vibrations transmitting along chassis from air movers at a magnitude that scales with about the second power of air mover speed, but the acoustical impact is now dominating. As capacity of rotational media increases, their targets for writing data are becoming smaller and are thus more susceptible to both the wider frequency range of acoustic energy (i.e., up to 20 kHz that reflects multiple blade-pass harmonics) and the sound pressure that scales at about the fifth power of air mover speed.

Hot Aisle Temperature and Worker Safety

There are multiple workplace safety standards globally that deal with heat stress on workers in a given environment. Within the United States, OSHA is often referenced as the authority, with research input from the National Institutes of Health (NIH) and other health organizations. Most use wet-bulb globe temperature (WBGT) as the key metric to determine the exposure conditions, exposure time, and risk of heat stress. WGBT is based on dry-bulb temperature measured with a black globe thermometer T_g, natural wet-bulb temperature T_{nw}, wind speed, solar radiance, and cloud cover. In the case of data center hot aisles, solar radiation and cloud cover do not apply. As such, the WBGT is calculated as follows (OSHA 2020):

$$\text{WBGT} = 0.7 \cdot T_{nw} + 0.3 \cdot T_g$$

One byproduct of increased inlet rack temperatures and humidity is a corresponding increase in the temperature and moisture content in the hot aisles. Typical work performed in the hot aisles of a data center, including but not limited to cabling and component replacement, can be categorized by OSHA as Easy (component replacement) or Moderate (cabling).

In the case of air-cooled IT equipment operating at the upper limit of the ASHRAE recommended range of 27°C (80.6°F), and maximum recommended dew point of 15°C (59°F) with a rack ΔT of 15°C (27°F), there are no limitations to Easy work, but Moderate work requires a 30 minute break every hour and a prescribed minimum amount of water intake to avoid heat stress.

Taking the case above and moving to the upper dry bulb of the A1 allowable range (32°C [89.6°F]) with a 15°C (27°F) ΔT across the rack, and no work of any kind in the hot aisles meets the safety guidelines without risk of heat stress per the OSHA standard. However, localized cooling could be provided while the service personnel are working in the high-temperature area.

Workload-Dependent Regression of Air and Water Inlet Conditions

As shown in Figure 3, there was an entire decade where server power hardly grew at all. Servers could be refreshed by new versions with good performance improvements and maybe slightly higher power levels due to an increase of similar power memory DIMMs. The industry became accustomed to a trajectory where air and water inlet temperatures were continually expanding, increasing opportunities for economization, heat recovery and reuse, and other similar highly efficient solutions. While this is still forecast to be present in many markets, it is critical to understand this trend is dependent upon IT equipment workload. There are plenty of data centers who are already seeing servers produced from the “power wars” period shown in Figure 3. While they had been used to only minor power increases, they may be faced with accommodating a server with 25% or more power and cooling requirements than the one it replaces. This much of an increase is typically not easy to absorb. The next challenge may be the temperature set point. Those operating within a cold data center, using chillers year-round, probably won’t have an issue. With data centers trying to operate more efficiently, they may already be seeing issues due to their higher air set point, especially if deploying the new H1 class servers mentioned in the section on Impact of Specialty Servers with Reduced Inlet Temperature Requirements.

Facility Temperatures May Need to Decrease

Hyper-dense systems that require a reduced ambient temperature supply due to difficulties with chip temperatures and/or preheat issues already exist. Obtaining a server with ASHRAE air cooling class A4 (max of 45°C [113°F]) support was once fairly easy; however, many of these platforms have already slipped back to only supporting the A3 or even A2 class. The hyper-dense systems may only support A2 with limited configurations. As mentioned previously, ASHRAE has just defined the H1 category with the suggestion that they be segregated into a more controlled portion of the data center. Similarly, supported water temperatures will soon be regressing due to increased power and lowered case temperatures.

Figure 4 attempts to show a transition from air to liquid cooling based on socket power. It is not exact, but some important trends should be recognized. As socket power moves through 300 W toward 400 W, standard 1U and 2U servers become more difficult from an air-cooled standpoint. It is likely that the data center may have transitioned to a maximum of A2 temperature or lower by this point but could easily be operating at W45, had it adopted liquid cooling. As chips fur-

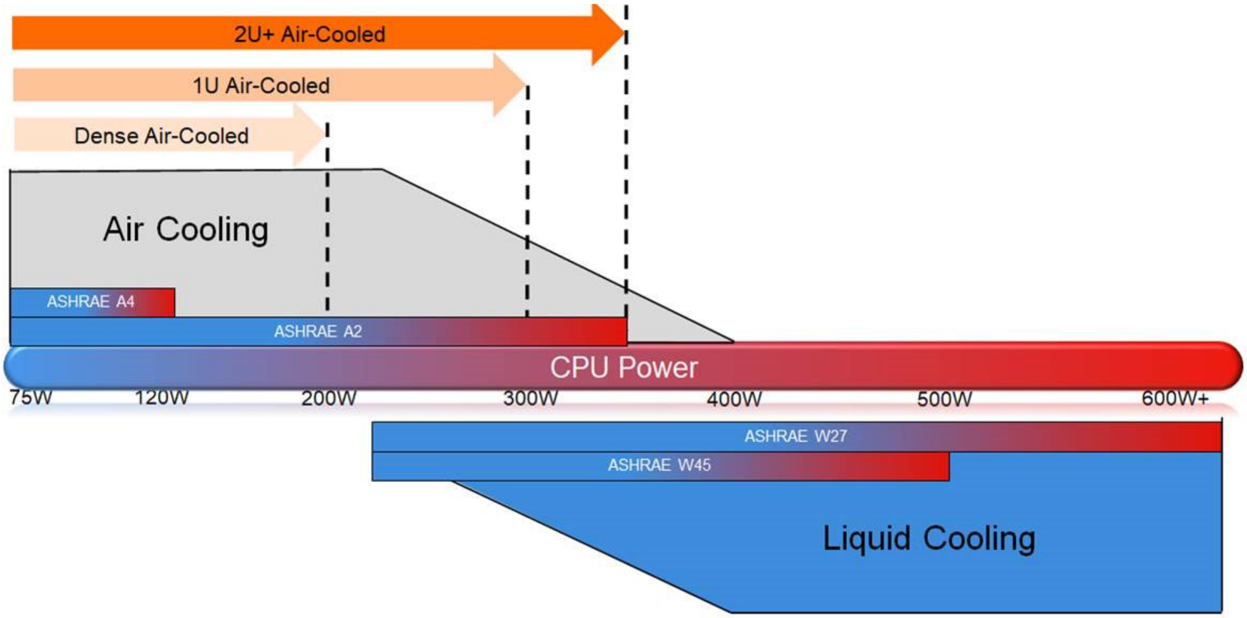

Figure 4 Air cooling versus liquid cooling, transitions, and temperatures.

ther increase in power, it is expected that the facility water temperature will likely have to be reduced. The exact numbers are somewhat speculative. IT manufacturers only have specific detail on one to two future generations of chips. Socket size plays an important role as well. A shift to a larger package size might help to alleviate the future socket cooling demands.

Multiple Components Driving Difficulty to Cool

The degree of cooling difficulty metric was introduced in Figure 2. This figure dealt specifically with CPU and GPU components. Once liquid cooling is considered, however, the entire liquid ecosystem must be taken into account. Memory, for instance, can become a dominant driver in the degree of difficulty. The combination of CPU and memory cooling requirements, along with the preheat within a common loop, will establish the ultimate limit in supportable facility water temperature. Future designs will use combinations of serial and parallel loops within the server to either minimize flow rate or to eliminate preheat. Some designs may be a bit more optimal than others. It is highly likely, however, that the DLC products of today might have the ability to operate on W45 (W4). In the future, they may be replaced by newer, higher-powered products only capable of W32 (W3) operation, and eventually cooling requirements may drive facility water temperatures down into the W27 (W2) range. Facilities should plan accordingly. This might seem like the industry is moving backward; in terms of data center cooling efficiency, it might be. At some point, the data center owners will need to decide whether compute performance is more important than data center energy use. The ultra-high water temperatures of W45 and beyond will become increasingly difficult to use as the heat flux at the socket increases.

While this section and the previous one identify the components driving designs toward liquid cooling, it is also clear that air-cooled products will still exist. *IT Equipment Power Trends, Third Edition* (ASHRAE 2018) presents the power trends of several form factors and workloads. In that book, only two products, scientific 1U-1/2 width and analytics 1U-2S, have compound annual growth rates (CAGRs) of greater than 4%. One other visualization, 2U-2s, has a CAGR of 2.9%. All the other 21 products listed in *IT Equipment Power Trends* have 2% or less CAGR; business processing servers show only 1% CAGR. This white paper is directed at the imminent need for liquid cooling for those in the scientific, analytics, and visualization categories, or those that use GPUs. Categories that have workloads with lesser power requirements may be progressively delayed so that liquid cooling is not required for a few years. As a data center operator, this progression of when to retrofit or design a new data center for liquid cooling is an essential consideration.

Decreased Hours of Economization

If the facility is built primarily to be operated in economized mode, but does have chiller capability, then it is positioned to handle the future. If it has no mechanical cooling, the facility should plan to add it in the future. For example, see the graph in Figure 5, provided by Oak Ridge National Laboratory (ORNL).

Through the approach temperature of the cooling tower and heat exchanger, this graph shows the facility water temperature throughout the year as reflected by the wet-bulb temperature in Tennessee where it is located. Most cold plate solutions today can live easily on W32 water with case temperatures in the 80°C–90°C range. It won’t be long before case temperature requirements drop into the low 70°C temperatures, especially for CPUs with high-bandwidth memory. The facility water temperature will need to be reduced or regressed even further to meet the cooling demands for dense systems with high-power memory and CPU. A case study for a dense high-performance computer (HPC) system with shadowed processors and CPU in a 1U half-width configuration is described in the Appendix. It is clear from Figure 6 that the combination of a 500W CPU and 12W DIMM could push the facility water to W27 and start eliminating the upper end of the economization graph shown in Figure 5.

This will impact the operating costs associated with running these types of HPC or high-density systems. Additional capital will also be needed to prepare a facility because of the cost of additional piping, heat exchangers, and controls required to trim the W32 (W3) water down to the required inlet temperatures. These cost impacts are compounded further if the facility must install new chilled-water equipment to support the lower required inlet water temperatures.

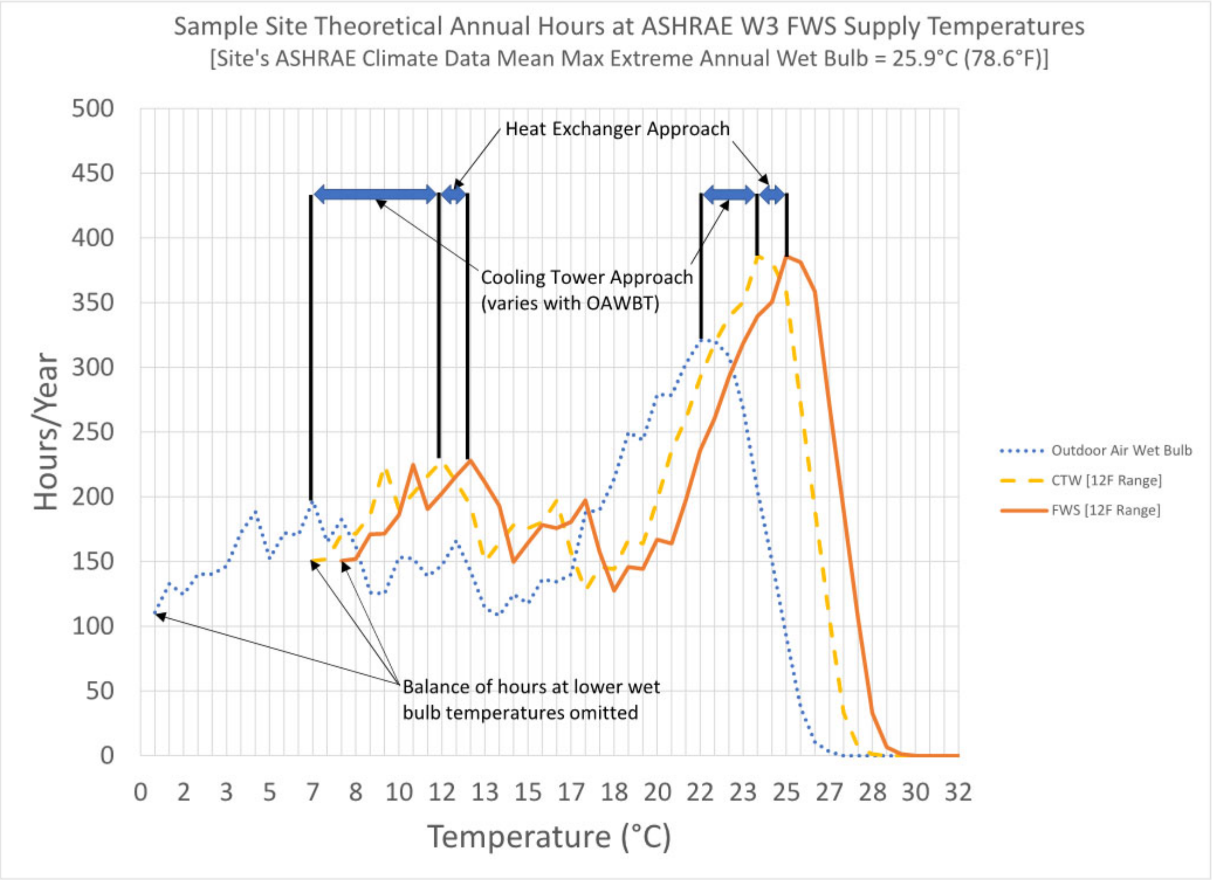

Figure 5 Economization example at ORNL.

No Water to the Rack, What to Do?

As the industry begins a more holistic shift toward liquid cooling, there may be data centers where infrastructure is not in place to easily enable these solutions. There may be data centers reticent to bring water close to the racks. One class of data centers where this has been most prevalent is colocation facilities, where service level agreements (SLAs) may be specifically written to exclude water close to the racks. It's unclear whether this is a hesitancy to bring primary water close to the racks or if it is a fear to expose aqueous-based secondary fluid systems above the raised floor that, if failed, might expose adjacent racks to water. If it is the latter, dielectric cooling could be considered. If it is the former, then clearly the facility is left with only computer room air conditioner/computer room air handler (CRAC/CRAH) cooling; the density advantages of row or door cooling, not to mention internal liquid cooling, are off the table.

This hesitancy for bringing water to the rack must change in the future or the facility will suffer performance limitations. For an interim period, these data centers might get by with liquid-to-air heat exchangers, either in the rack or at the row level, that provide the IT coolant while rejecting the heat back into the data center. Though this is less efficient than rejecting heat directly to facility water, this can provide a bridge in the interim for those that are set up with advanced air-cooled solutions. In the future, facility water headers may be installed, and the liquid-to-air heat exchangers can be replaced with liquid-to-liquid heat exchangers, whether they be an aqueous-based coolant or a dielectric.

It should be emphasized: this paper assumes that facility water temperatures will decrease over time due to increasing power and lowered case temperatures. Because of the substantially lower heat capacity of air, liquid/air heat exchangers have much higher approach temperatures than liquid/liquid. As a result, the use of liquid/air heat exchangers as an interim solution will accelerate the required reduction in facility water temperatures, as well as all impacts associated with such reductions.

Appendix: Additional Information and Other Considerations

This section contains additional information on some of the earlier portions of the paper. It also points out other considerations, including details from case studies of existing liquid-cooled data centers.

Energy Effectiveness

There is ongoing pressure to reduce the energy consumption of data centers. Energy costs (kWh) vary widely throughout the world and in locations where the cost is very high (such as some countries in Europe), it is hugely beneficial to design data centers with maximum energy efficiency and a low power usage effectiveness (PUE). Energy capture of waste heat to heat buildings, gardens, pools, and so on; generating cold water using adsorption chillers; or contributing heat back to the utility company are examples of further improving energy effectiveness. Liquid cooled solutions allows a reduction of PUE to less than 1.1. Warm-water cooling minimizes (or eliminates) the need for chillers, which helps energy effectiveness.

A good example of an energy-efficient HPC data center is the Super MUC-NG in the Leibniz Supercomputing Center (LRZ) at the Bavarian Academy of Science and Humanities in Garching, Germany. Super MUC-NG is a cluster of 6480 thin and fat nodes (Intel SkyLake Processors, Lenovo ThinkSystem SD 650) with main memory of 719 TB and peak performance of 26.9 PFLOPs. The cooling is direct warm-water cooling (40°C–45°C) with the waste heat used to generate cold water using adsorption chillers. The chilled water is used for rack door heat exchangers to capture heat from other IT equipment in the data center that does not use direct water cooling. The energy efficiency (30% savings) is achieved by lower server power consumption (no fans in server), reduced cooling power consumption (warm water at 40°C), Energy-Aware Scheduling, and a need for less mechanical refrigeration to cool the system. The use of warm-water cooling and absorption chillers results in a partial PUE (cooling) of less than 0.07.

TCO Discussion Energy Effectiveness

When evaluating the cost of adding liquid cooling to data centers, owners, engineers, and contractors often focus on the cost increases typically observed in mechanical systems due to increased piping materials, insulation, installation labor, and so on. However, a proper total installation cost analysis also factors in the cost differences for the supporting systems.

When transitioning from air-cooled to liquid-cooled IT equipment, this almost always is accompanied by densification. By placing the same electrical load in fewer racks, there are fewer connections. In HPC applications, higher density applications use 480 V distribution to the server cabinets, eliminating low voltage transformers and power distribution units (PDUs). The end result of this densification is typically a reduction in the cost of the electrical installation due to reductions in distribution requirements.

The increased densification of IT equipment compute also increases the network bandwidth requirements at the rack level. This is offset again, however, by fewer racks to support the same IT equipment load. Significant decreases in physical network distribution and terminations, such as in the case of the electrical distribution and terminations, typically more than offsets the increased specification of network cabling to each rack, resulting in a net reduction in installed network cost.

Such densification can also reduce the required building footprint to support the IT equipment deployment and, in particular, high-density or high-load applications can reduce the facility site space requirements. A proper first-cost comparison takes all of these items into account as they are a package deal with liquid cooling.

Recent studies on such comparisons have shown that as rack density increases, the installed first cost per MW decreases due to the factors mentioned above. There is a rack density point, which varies per cooling solution, at which the first cost to build and deploy liquid cooling is actually lower than air-cooling the same IT equipment load, making the payback period instantaneous. A comprehensive evaluation of both solutions must be performed, otherwise the resultant financial analysis will be incomplete.

Waste Heat Reuse

A unique feature of the Super MUC-NG Cluster at LRZ in Germany mentioned previously is their focus on waste heat reuse. Super MUC-NG waste heat is sufficient to reheat about 40,000 m^2 (430,500 ft^2) of office space. LRZ also uses the waste heat to drive an adsorption chiller for rear-door heat exchangers, capturing the non-liquid-cooled heat. This eliminates the need to add a CRAC/CRAH to the data center. The racks are insulated to reduce radiation to the environment given the high supported water temperatures. The adsorption chiller uses a combination of an absorber and an evaporator to generate 600 kW of chilled-water capacity, helping the drive toward better total energy efficiency.

Pure Performance

The HPC community is quite competitive and supercomputers are ranked regularly on their performance capability in PFLOPs. The highest-placed players in the TOP500 HPC data centers (published twice annually) are all full or hybrid liquid-cooled data centers. These data centers have extremely dense racks (10s to 100s kW) with high-power nodes using a combination of CPUs and GPUs. Full or hybrid liquid cooling enables dense, high-power racks, essential to achieve the performance of these data centers using a reasonable number of racks within the real estate available in a typical data center.

Frontera, located at the Texas Advanced Computing Center (TACC) at University of Texas, Austin, is the 9th most powerful supercomputer in the world (TOP500 2020) and the most powerful supercomputer on any academic campus. The peak performance of Frontera is 39 PFLOPS. The primary computing system for Frontera is provided by Dell EMC and powered by Intel processors, interconnected by a Mellanox HDR and HDR-100 interconnect. The system has 8008 available compute nodes. The nodes have hybrid cooling with the CPUs, direct liquid cooled, and rest of the system air cooled with a rear door heat exchanger (RDHX) in the rack.

Immersion

Immersion cooling is becoming a popular option for cooling IT equipment used in cryptocurrency mining/blockchain, oil and gas applications, and HPC. Immersion cooling has the benefits of broad temperature support, high heat capture, high density, and flexible hardware and deployment options. There is a great interest in immersion cooling, with a large number of startup and established players enabling large integrated solutions and proof of concepts in all key market segments.

Immersion cooling can create additional challenges and complexities, specifically with regard to deployment and service. For tank applications, a crane or two-man lift is often required to remove IT equipment hardware for service. For rack-based solutions, sealing the immersion fluid can be challenging. Interoperability with an immersion fluid with IT equipment hardware can also be an issue that may impact warranty. It is recommended that a materials compatibility assessment and warranty impact evaluation are performed before deploying an immersion cooling solution.

For these solutions, high heat dissipation devices such as GPUs and CPUs require increased surface areas for heat transfer relative to a package with an integrated heat spreader or bare die, which is generally enabled by using heat sinks designed for forced air cooling. As component heat increases limits for single-phase natural convection, immersion cooling limits may be exceeded requiring a move to combine other cooling technologies or a wholesale shift to forced convection or two-phase immersion cooling altogether.

Memory Cooling

System-level memory has become an increasingly important focus area for server cooling. Over the course of the development of double data rate synchronous dynamic random-access memory (DDR SDRAM), the increasing power levels have reflected the need for larger capacity and higher frequency to support bandwidth. The first generation of this memory type is referred to as *DDR* and was introduced in 1998. Subsequent generations of the DDR SDRAM memory are denoted using the labels DDR2, DDR3, DDR4, and DDR5.

There are several flavors of the DDR DIMM. For the purpose of clarity, this section will focus on the registered memory DIMM, or *RDIMM*. The RDIMM is a common version to use as server memory as it is buffered from the system memory controller and often provides error correction. The CPU will support memory in assigned channels and can support multiple DIMMs in that assigned channel. For simplicity, this section will focus on a single DIMM populated in CPU memory channel and referred to as one DIMM per channel (1 DPC). The 1 DPC configuration has the DIMM reach its highest power state.

Investigating the 1 DPC configuration over the course of DDR generations shows an interesting trend. First, the largest-capacity DIMM running the highest frequency has shown a steady increase in power consumption over time. Second, as the CPU performance and capability to support more memory has increased. The memory subsystem rapidly increases in power. Table 1 shows the historical trend in the DDR memory over four generations of memory. The power at max frequency for an individual RDIMM has grown from 12 to 20 W. The CPU has evolved to support four DIMM channels in the DDR2 time frame to the projected twelve DIMM channels in the DDR5 time frame. The associated memory subsystem power has grown from 48 to 240 W per CPU. That memory subsystem power is now on par with the CPU power consumption.

One could consider this a more difficult challenge to cool relative to CPU because the power is distributed in multiple devices of varying height and the DIMMs must remain serviceable in the system. Furthermore DDR5, in current development, will contain additional discrete devices that require cooling such as the power management integrated circuit (PMIC).

Air cooling may be extended up to DIMM power levels of approximately 15 W; however, airflow balancing between the CPU heat sink and the DIMMs becomes more difficult as both the DIMM and CPU power increase.

Liquid cooling with direct on-chip cold plates can provide cooling relief, especially if the cold plates are placed on both sides of each DIMM card. While two-sided cooling can be expected to cool upwards of 20 W per DIMM, the design, manufacture, and servicing is more complex. A case study with one-sided DIMM

Table 1 Memory Generations Power at Highest Capacity

Memory Generation	Highest Capacity	Power at Max Frequency	1 DPC Count	Total Memory Subsystem per Socket
DDR2	16 GB	12 W	4	48 W
DDR3	32 GB	16 W	4	65 W
DDR4	128 GB	17 W	8	144 W
DDR5	256 GB	20 W	12	240 W

Table 2 Cold-Plate Water Cooling System Configuration Assumptions

Component	Details
Coolant distribution unit (CDU)	Commercial off-the-shelf nominally sized for 750 kW
Number of racks per CDU	8
Rack height	40U for liquid cooled blades + 2U top-of rack (TOR) air-cooled switch
Liquid-cooled blades	2U consisting of 4 half-width 1U nodes
1U nodes	2 CPUs with 8 DIMMs per CPU
CPU cooling	Direct on-chip cold plate
DIMM cooling	One-sided cooling with cold plate

cooling in a dense 1U half-width node with two CPU in series is shown in Figure 6. The design configuration used for this case study is described in Table 2.

The ASHRAE water cooling class W27 can support up to 500 W for each CPU and up to 12 W per DIMM with single-sided DIMM cooling. On the other hand, ASHRAE water cooling class W45 reduces the cooling capability to a point that it may be comparable to a conventional air-cooled system with a low cold aisle temperature for this dense 1U half-width shadowed configuration. Several underlying design choices modulate the cooling capability shown in Figure 6. For example, a more capable pump and heat exchanger in the CDU, fewer blades per rack, fewer racks per CDU and/or a double-sided memory cooler would contribute to greater cooling capability. This exercise is simply intended to show that careful analysis is necessary to architect the cold plate-based water-cooling system. Another option with direct on-chip cold plates is to cool the CPU with liquid while the memory subsystem is cooled by air. This hybrid approach also has its downsides. System fan power remains high, especially because the CPU cold plates contribute to airflow impedance, the heat capture to liquid is decreased, and therefore the overall TCO benefit of liquid cooling is eroded.

Single and two-phase immersion may provide sufficient cooling for high-power DIMMs while not introducing the serviceability challenges with direct on-chip cold plate cooling.

Acoustics

Increase in air movers speeds used to cool IT equipment has significant airborne noise impacts to human health and annoyance and also to hard disk drive performance degradation. Some respective references are ECMA-74 (ECMA 2019) and a previous ASHRAE white paper (ASHRAE 2019). Empirical observations documented in literature show that sound pressure amplitudes from air mov-

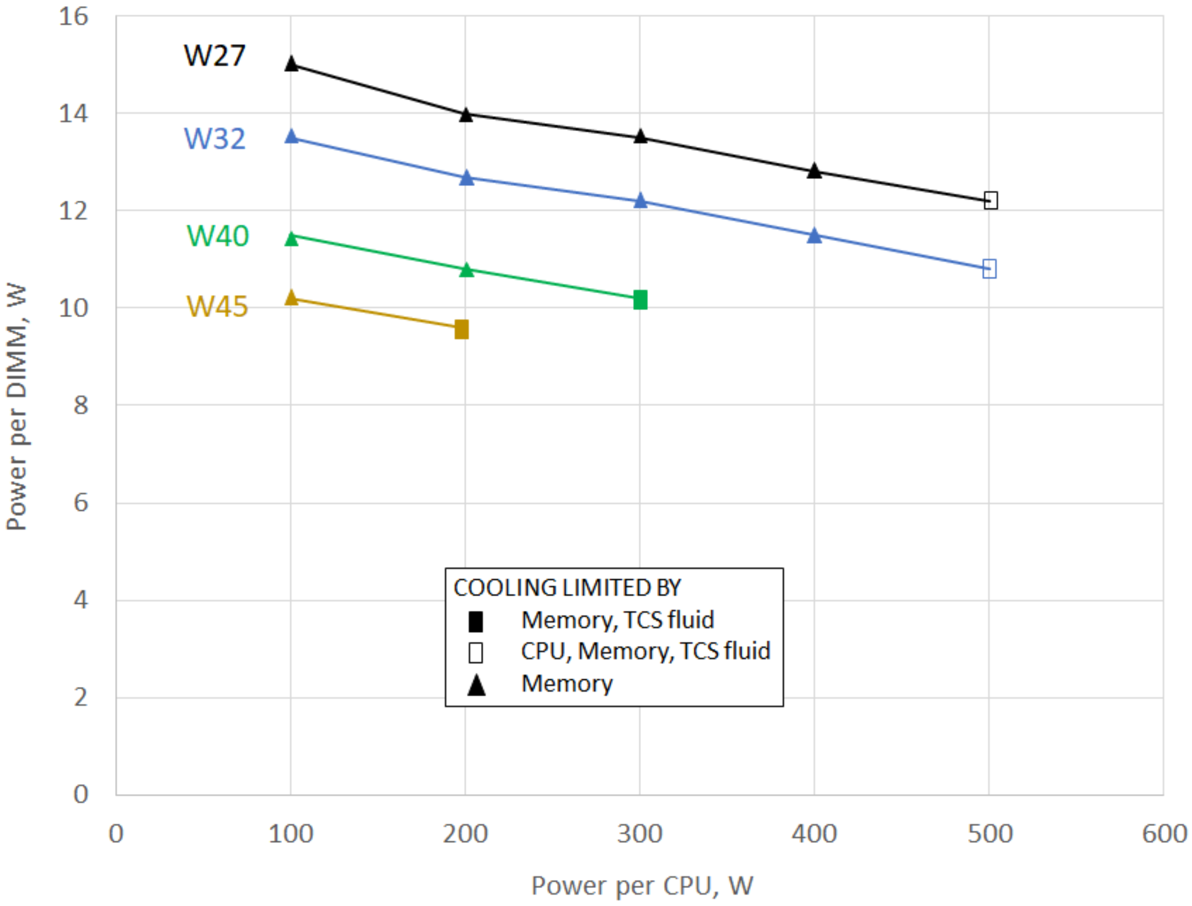

Figure 6 Reduction of cooling capability with cold-plate-based liquid cooling of CPU and memory.

ers increase with ~5th power of their operating speeds and vibration with ~2nd power.

Beyond the annoyance from such characteristics as blade-pass harmonics, maximum time-weighted average A-weighted sound pressure level thresholds per exposure time for hearing protection have existed for many years, for example by the National Institute for Occupational Safety and Health (NIOSH 1998). This document states that exposure beyond an 8-hour time-weighted average at or above 85 dBA is considered hazardous. Or, per Table 1-1 in this document, “no worker exposure shall equal or exceed 90 dBA for 2.5 hours, 100 dBA for 15 minutes, 110 dBA for 1.5 minutes,” and so on. That is to say, the allowable exposure time dramatically decreases with the experienced A-weighted sound pressure level.

Ambient A-weighted sound pressure levels of data centers regularly meet or exceed 80 dBA. In fact, such an A-weighted sound pressure level may be recorded at a standing position 1 m (3.3 ft) from a single server. If we assume a 30%

increase in air mover speed, then an increase of 6 dB is expected. With a 50% increase in air mover speed comes a 9 dB increase. Clearly, the conflict with safety exposure thresholds becomes more stark and greater precautions must be taken.

Sensitivity of hard disk drives is increasing with their storage capacity. In a general sense, because hard disk drive heads must target ever-smaller locations on the platters, the same amount of disturbance can more easily move the head off track and disturb write operations. Although vibration transmission has been the historical concern and remains a risk, disturbance from acoustic energy in the range of 1 to 20kHz has become the dominant mechanism over the past decade.

Unweighted overall sound pressure levels at the hard disk drives in servers regularly reach 100 dB and higher. Each hard disk drive model has its own nonlinear response to frequency content of the disturbances, but in a very general sense, unweighted overall sound pressure level at the hard disk drive in excess of 110 dB may substantially degrade throughput performance. Similar to the argument for human perception of airborne noise, increases of 30% to 50% in air mover speed result in jumps of 6 to 9 dB sound pressure level and thus higher hard disk drive throughput performance risk.

Disruptive Workload Types

There has been a paradigm shift regarding workload types and their associated market segments. This shift is driven by insatiable demand for more data, the need for real-time data consumption, and real-time analytics. High-performance analytics and artificial intelligence (AI) are crossing market segments from pure HPC play to more central, edge, cloud, HPC in cloud, and colocation data center solutions.

Enterprises embracing AI applications for enhancing market data analyses, predictive model building, and so on are primary customers of colocation data centers. Machine learning and AI functions optimize demand-side advertising platforms. Online gaming introduces intensive time-of-day loading based on gamer behavior; gaming uses typically involve nonvirtualized GPU-based server allocations to enhance user experience. Aircraft manufacturers, global industrial machinery conglomerates, and others are increasing their use of HPC systems with the latest processor designs. Emerging autonomous-vehicle-focused applications also use AI capabilities, and will do so increasingly at the edge. All of these put a strain on data center cooling systems. Almost all these applications demand dense compute/memory/networking clusters enabling low-latency fabric, pushing the limits of air-based cooling, but more easily accommodated by liquid cooling.

Data Center Infrastructure Limitations

The critical path to support a specific IT equipment mission for any data center will rely on its ability to cater to that mission’s power, space, or cooling needs. A recent article on ComputerWeekly.com titled “Datacentre liquid cooling: What needs to happen for it to become commonplace in colocation?” also describes these challenges (Donnelly 2020). A balance must be found between the space utilization for the IT equipment and the supporting infrastructure such as switchboards, PDUs, CRACs/CRAHs, and CDUs. Some older data centers shifting to denser racks may find themselves with more empty space than a casual observer may expect simply due to floor loading limitations. Design variation within the world's stock of data centers is high, but this section attempts to describe common hurdles that one may encounter when pursuing liquid cooling in an existing data center for the first time.

Power

If a liquid cooling solution is needed at a rack, the physics of heat removal necessitate it. Once a liquid-cooled system is introduced, data centers that may have previously been limited by the facility's ability to deliver sufficient air cooling may now have the opportunity for significant rack density increases. One cannot ignore comparable compliments of power delivery that might come in the future. Given a new liquid-cooled mission, the quantity and size of existing circuits may suffice for smaller systems. As systems grow, there is an inflection point at which increasing the circuit voltages from 208 to 480 V makes sense given the smaller conductor sizes, circuit and in-rack PDU quantity, and operational costs associated with voltage losses when distributing lower voltages.

Arch flash modeling and fuse and circuit breaker coordination must be considered in any system, but are even more important as large electrical distribution gear gets closer and closer to the electrical power's point of use.

Whether electrical power is overhead or underfloor, space coordination must identify conflicts between other electrical circuits, electrical power trays, network cable trays, cooling infrastructure, fire protection systems, and anything else that could compromise the performance, accessibility, or maintainability of any system.

Space

Data center whitespace must be able to accommodate the dimensional aspects of the racks being cooled as well as their associated weights. Rack aisle pitches may need to accommodate additional spacing for RDHXs. For rack densities that are challenged in getting adequate airflow from a raised-floor system, designers must understand the aisle space requirements to know if the number of available floor tiles would actually provide enough airflow from the raised floor or not. Air flow consumption of high rack powers should be carefully analyzed. Air flows of a 40 to 50 kW rack could be up to 5000 cfm. Floor tile best-in-class is 1900 cfm.

For large system installations, the orientation and spacing of rows (or pitch) must be analyzed and coordinated with existing obstructions such as other rows of cabinets, columns, and electrical and cooling equipment and the area they require by code and for maintenance activities. In looking at a metric of cooling capacity per unit area taken by the cooling equipment, one can determine some value of conserving more space for IT equipment.

Load paths for heavy racks must have the appropriate rating for rolling loads. This could include elevators, ramps, thresholds, corridors, and so on.

Cooling

Small-scale liquid cooling can be accomplished in a straightforward manner. A one-rack system can use internal to the rack liquid-to-air heat exchangers if facility cooling water is not available. Even at a small scale, a fully populated liquid-cooled rack could generate significant heat into the data center. Best practice air management must still be used to prevent hot spots. These high-density racks can quickly use any remaining capacity of existing air handlers. Or, if facility cooling water is available, a rack-mounted cooling distribution unit can be used. If needed, these units can be installed in a neighboring rack or under a raised floor if space is available.

Large-scale liquid cooling involves bringing more piping deeper into the data center. If done overhead, piping should be kept above aisles and branch piping should be sized to serve rows on either side. With piping connections passing over critical or costly equipment, drip pans with leak detection and piped drains routed to the floor should be installed. Coordination between power, cooling, and network routing is of the utmost importance.

Efficiency Expectations

To take advantage of the efficiency benefit enabled by liquid cooling, one must translate the elevated fluid temperatures released from the data center to the heat rejection equipment designed for low kW/ton efficiencies. This largely means eliminating mechanical cooling equipment such as chillers. Existing large-scale data centers expecting significant future adoption of liquid cooling with extensive low-temperature chilled-water distribution piping could consider converting the existing piping over to warm water and then installing economization equipment at the centralized level. Then smaller diameter piping could be installed to continue providing chilled water to those loads that require it.

Water Quality and Maintenance

Water quality and ongoing maintenance is an issue of great concern to many facilities. A closed system's long-term success often is determined in the design and installation phase. During design, a wetted materials list should be created and subsequently maintained to identify that all materials that come in contact with the liquid are, in fact, approved. Materials with properties that inherently keep the closed system clean should be considered. Installers should abide by clean construction techniques. A clean and flush plan should be considered during design to incorporate features that will enable an effective clean and flush.

Controls

Additional power use by IT equipment can cause problems with cooling system controls and equipment stability. If elevated temperatures are used, any headroom for temperature or flow excursions becomes limited. IT equipment OEMs outline the temperature and flow magnitude, duration, and rate of change envelopes in which stable operation can be performed. Matching the cooling exactly with the heat load results in perfectly consistent supply temperatures. Temperature fluctuations in the supply water from facilities can occur if the cooling load is erratic, there is low cooling load diversity on the system, large equipment is staged

ON/OFF, control system tuning is off, or there is a mismatch between the cooling load and the cooling being provided.

Nomenclature

CAGR = Compound annual growth rate

CDU = Coolant distribution unit

CRAC = Computer room air conditioner

CRAH = Computer room air handler

CPU = Central processing unit

CTW = Cooling tower water

DDR = Double data rate. A type of memory; a number behind DDR indicates the generation

DIMM = Dual in-line memory module

DPC = DIMMs per channel

FPGA = field programmable gate array

FWS = Facility water system

HPC = High-performance computer; also called a *supercomputer*

GPU = Graphic processing unit

NIH = National Institutes of Health

NIOSH = National Institute for Occupational Safety and Health

OEM = original equipment manufacturer

ON/OFF roadmap = Silicon vendors have a list of CPU SKUs that are advertised that are considered *on roadmap* and generally available to the public; there also can be *off roadmap* SKUs that can be made available upon request by the customer and/or server vendor

OSHA = Occupational Safety and Health Administration

PFLOPs = petaFLOPS. A measure of computer performance equal to 1015 floating point operations per second

PUE and partial PUE = Power usage effectiveness. A metric used to describe facility efficiency defined simply as the energy entering the facility divided by the energy used by the IT equipment. A facility with equal amounts of overhead and IT energy would have a PUE of 2.0 because only half of the energy is used to power the IT equipment. Overhead can be attributed to many things such as cooling, power conversion, lighting, office equipment not performing the functions of the IT equipment, etc. A partial PUE would an example of a breakdown of the overhead into a specific function such as the cooling partial PUE consisting of all energy associated with the cooling function.

RDIMM	=	Registered memory DIMM
SDRAM	=	Synchronous dynamic random-access memory
SKU	=	Stock keeping unit. Units within the same family of processors with differing attributes such as frequency, number of cores, wattage, and case temperature limits
SLA	=	Service level agreements
TCO	=	Total cost of ownership
TDP	=	Thermal design power. The maximum amount of heat generated by a component
TOR	=	Top of rack (network switch product)
UPS	=	Uninterruptible power supply
WGBT	=	Wet-bulb globe temperature. A measure of the heat stress in direct sunlight, which takes into account temperature, humidity, wind speed, sun angle, and cloud cover (solar radiation)

References

ASHRAE. 2013. *Liquid cooling guidelines for datacom equipment centers, second ed.* Peachtree Corners, GA: ASHRAE.

ASHRAE. 2016. *IT Equipment Design Impact on Data Center Solutions*. Peachtree Corners, GA: ASHRAE.

ASHRAE. 2018. *IT Equipment Power Trends, Third Edition.* Peachtree Corners: ASHRAE.

ASHRAE. 2019. “Hard Disk Drive Performance Degradation Susceptibility to Acoustics.” Whitepaper. Peachtree Corners, GA: ASHRAE. https://www.ashrae.org/file%20library/technical%20resources/bookstore/hard-disk-drive-performance-degradation-susceptibility-to-acoustics.pdf.

ASHRAE. 2021. *Thermal guidelines for data processing environments, fifth ed*. Peachtree Corners, GA: ASHRAE.

Donnelly, C. 2020. Datacentre liquid cooling: What needs to happen for it to become commonplace in colocation? *ComputerWeekly.com*. https://www.computerweekly.com/feature/Datacentre-liquid-cooling-What-needs-to-happen-for-it-to-become-commonplace-in-colocation.

ECMA. 2019. EMCA-74, *Measurement of Airborne Noise Emitted by Information Technology and Telecommunications Equipment, 17th Edition*. Geneva, Switzerland: European Computer Manufacturers Association International.

NIOSH. 1998. *Occupational noise exposure.* Washington, D.C.: United States Department of Health and Human Services.

OSHA. 2020. O*SHA Technical Manual TED 1-0.15A*. Washington, D.C.: United States Department of Labor.

TOP500. 2020. TOP500 LIST-NOVEMBER 2020. Sinsheim, Germany: Prometeus GmbH. https://www.top500.org/lists/top500/list/2020/11/.