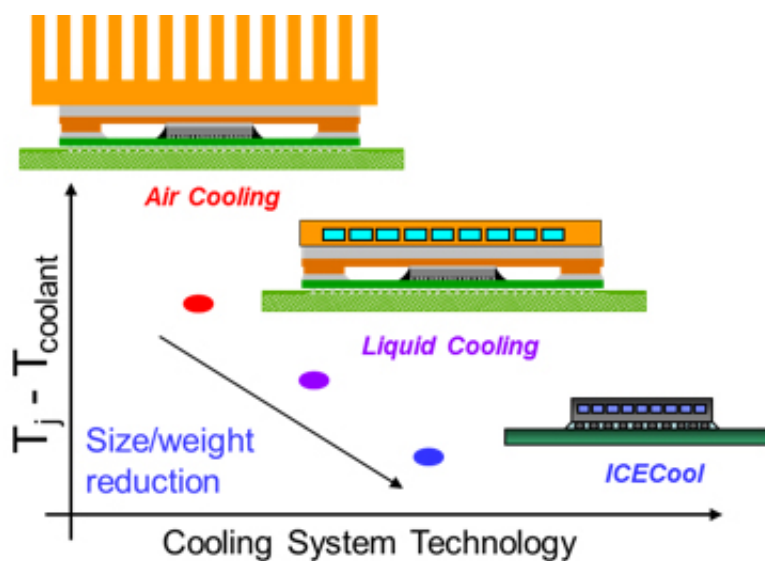# Beat the heat in 3-D chip stacks with ICECool
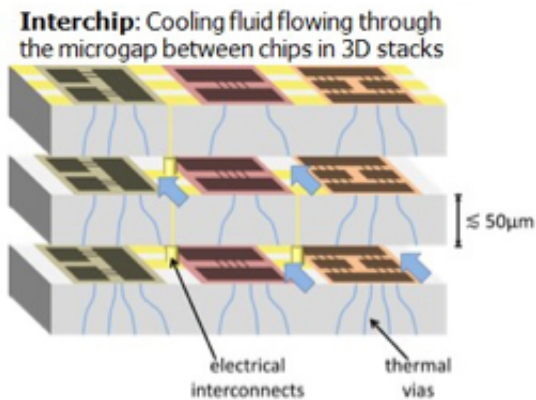
August 30 2017

Credit: IBM

In the Moore's Law race to keep improving computer performance, the IT industry has turned upward, stacking chips like nano-sized 3-D skyscrapers. But those stacks, like the law it's challenging, have their limits, due to overheating. So, our team in New York, alongside colleagues in Zurich, received a 2013 contract to tackle intra-chip cooling from the Defense Advanced Research Projects Agency (DARPA) in its ICECool program. For our part, we developed a new cooling technology to overcome the thermal barrier of stacking chips, an on-chip solution that could help to even cool off entire datacenters.

Today, chips are cooled by fans which push air through heatsinks that sit on top of the chips to carry away excess heat. Advanced water-cooling approaches, which are more effective than air-cooling approaches, replace the heatsink with a cold plate that is closer to the chip. But because of the electrical conductivity of water, this approach requires a barrier to protect the chip. ICECool uses a nonconductive fluid to take the next step of bringing the fluid into the chip (as shown in the image below). This does away with the need for a barrier between the chip and fluid. It not only delivers a lower device junction temperature (T_j), but also reduces system size, weight, and power consumption (SWaP). Our tests on IBM Power 7+ chips demonstrated junction temperature reduction by 25º C, and chip power usage reduction by 7 percent compared to traditional air cooling.

Today's chip stack "skyscrapers" in reality are more like chip stack "row houses." Using a heatsink or cold plate holds back 3-D chip-stacking height because of their inability to cool chips in the middle and bottom of the stack. IBM's ICECool technology circumvents that problem by pumping , a heat-extracting dielectric fluid right into microscopic gaps, some no thicker than a single strand of hair, between the chips at any level of the stack.

The dielectric fluid used in ICECool can come into contact with electrical connections, so is not limited to one part of a chip or stack. This "go anywhere" ability benefits chip stacks in terms of materials and architecture, such as putting memory directly on the stack, which improves the speed of everything from graphics rendering to deep learning algorithms.

Interchip: Cooling fluid flowing through the microgap between chips in 3D stacks

≤ 50μm

electrical interconnects

thermal vias

Credit: IBM

ICECool works much like coolant in a car's air conditioning. It's pumped into the chips, where it removes the heat from the chip by boiling from liquid-phase to vapor-phase. It then re-condenses, dumping the heat to the ambient environment where the process begins again. Cars, though, need a compressor to cool the air below the ambient temperature (because rolling down the window doesn't help much in rush hour traffic). Chips, unlike humans, can operate at 85º C or 185º F. So the outdoor ambient temperatures are already cooler than the chips. Therefore, our ICECool process doesn't need a compressor (one of many elements that contribute to lowering a datacenter's energy expenditure).

Datacenters chill out with ICECool, too

Datacenters in the US – often non-descript buildings spanning millions of square feet – full of servers that, among many things, power the internet, use about 70 million megawatts of electricity, annually. Those MWs translate to about 2 percent of the country's energy. Two percent may not sound like much, but that's more electricity than 29 states, as well as the District of Columbia use individually in a year.

IBM Research teams have been hard at work reducing the heat produced by datacenters – which accounts for a third of those 70 million MWs. While most data centers today are air cooled, IBM has developed warm-water cooling with projects such as a Department of Energy project (Economizer Based Data Center Liquid Cooling) and the SuperMUC hot water-cooled [data center](#) in Munich. While water is an effective coolant and shown to provide significant cooling energy savings, it requires isolation from the electronics. As ICECool uses a non-conductive dielectric fluid it can come in direct contact with electronics and remove heat by converting from liquid to vapor-phase as it flows through the electronics package.
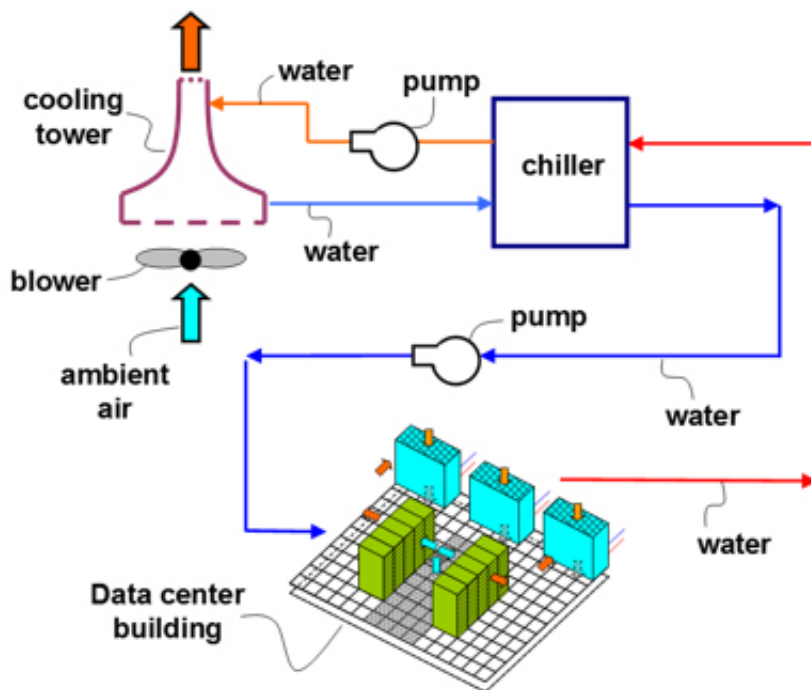
Credit: IBM

CRAC (Computer Room Air-Conditioning) and CRAH (Computer Room Air Handler) units, which are like "heatsinks" for today's

datacenters, blow chilled air across the rows and rows of servers. That chilled air is supported by a compressor-based chiller (like a car's AC). The chiller removes the heat via a tower on top of the exterior of the datacenter. Think of the tower as a giant radiator that dumps heat into the atmosphere – as shown on the top-left of the diagram, below. This is the loop that accounts for one-third of a datacenter's costs.

ICECool has the potential to eliminate the chiller, plus the CRAC unit, and most of the fans because it can be in direct contact with any and all electronic components. Based on our tests with IBM Power Systems, ICECool technology could reduce the cooling energy for a traditional air-cooled data center by more than 90 percent.

Read our *IEEE Transactions on Components, Packaging and Manufacturing Technology* paper, Improving Data Center Energy Efficiency with Advanced Thermal Management, to learn more about ICECool.

More information: Timothy J. Chainer et al. Improving Data Center Energy Efficiency With Advanced Thermal Management, *IEEE Transactions on Components, Packaging and Manufacturing Technology* (2017). DOI: 10.1109/TCPMT.2017.2661700

Provided by IBM