Using Copper Water Loop Heat Pipes to Efficiently Cool CPUs and GPUs

Stephen Fried, Microway Inc.

As the amount of power being rejected by 1U servers starts to approach and exceed 2 KW, the question in HPC continues to be, how can we not only cool devices which reject this amount of heat, but also, how can we reject that heat efficiently.

The current technology that we use today came out of the commodity PC arena and has used forced air cooling since 1985. As the heat being rejected by CPUs approached 50 Watts and the format of server rack mount chassis went from 2U down to 1U, the problem became finding fans with enough capacity to cool CPUs mounted on a PCB, that had about an inch of headroom. The only solution to this problem turned out to be Copper or Aluminum based heat sinks that employed Copper or Aluminum fins which were in turn cooled by 1U fans that started out operating at 10,000 RPMs. As the power started to approach 100 Watts, the fan speed needed to increase to 20,000 RPM which in turn dramatically increased noise and power while reducing reliability. The reliability problem was partly solved by stacking a pair of fans, one in front of the other and making it easy to replace fans that failed. However, this did not solve the problem created by the horizontal stacking of GPUs into 1U chassis such that the effluent of the first GPU in the chassis pre-heated the air going into the second or the stacking of DIMM modules and CPUs. In all of these situations, the total air flow passing through the chassis had to be increased to deal with the fact that the trailing heat source needed to be cooled with faster moving air because it was pre-heated. In addition, it turns out that many chassis (especially those used in workstations) suffer from a problem we call recirculation (the mixing of hot and cold air), a problem which also arises inside of rack cabinets and data center rooms. At the end of the day, the air leaving a 1U chassis that contains stacked components leaves the chassis at a lower average temperature than a chassis that does not have stacked components for the simple reason that to transfer the heat to the flow, requires that the total flow velocity be increased. The reduction of the average temperature of the exhaust flow leaving a 1U chassis in combination with recirculation losses in the rack cabinet and data center room, results in a reduction in the Coefficient of

Performance (COP) of the water chillers that feed their cooling towers. The real purpose of water chillers turns out to be fixing all of the inefficiencies just described by raising the temperature of the thermal stream that it passes to the cooling tower. One the limiting performance constraints of cooling towers turns out to be the second law of thermodynamics, heat does not flow from a cold body to a hot one (the outside air). A discussion with a cooling tower company in Atlanta Georgia set one of the physical constraints that the system we are about to describe must meet. We asked them, on the hottest day of the year, what is the temperature of the water you can deliver to us in Atlanta. Their answer was we can give you 30 C water if you return 35C water for us to cool using outside air.

In a prior life, the author managed a research project for DOE that employed an Ammonia Rankine Bottoming Cycle to cool a 12 MW fuel cell power plant. The simulation used the waste heat to boil Ammonia and used the resulting vapor to drive a 1 MW electric turbo generator. One of the features of such a cycle, is the temperature clamping action of a working fluid that is changing phase: until all of the fluid has been converted to a gas or vice versa, the huge Enthalpy of Evaporation of many liquids essentially clamps the temperature of the process to that of the working fluid that is changing phase which in turn is controlled by the pressure of the working fluid. This suggested that one way to cool a CPU or GPU is simply to flow a working fluid like Ammonia through a heat exchanger in which boiling takes places (i.e. an evaporator) and then to reject this heat some place else inside of a 1U server chassis where there was plenty of room to mount a device called a condenser that in the case of an air cooled server would have much more fin area than the heat sinks that sit on top of a CPU. However, there were problems with this approach, that included finding a miniature pump as well as another device for holding the liquid prior to its entering the evaporator, which we discovered already had a name: the compensation chamber. Just when I was starting to think that this approach would never fly, along came Google and an invention called the Loop Heat Pipe used to cool electronics in space vehicles whose working fluid also turned out to be Ammonia. And, while Ammonia is the ideal working fluid for devices that need to operate in both hot and cold environments, we knew before we even

started we might have to change working fluids, after I contacted the inventor of the LHP, Professor Yury Maydanik of the Institute of Thermal Physics of the Urals. Yury just happens to also be the chairmen of the International Heat Pipe Symposium, and it did not take us long to get a grant from CRDF that applied Ammonia LHPs to cooling 1U servers.

We published the results in IEEE Components and Packaging. Here are the summaries of that investigation which employed Nickel Ammonia LHPs.

1. Miniature ammonia LHPs are efficient means for improving the cooling performance of high performance computers.

2. They make it possible to create a 1U computer with an LHP-based cooling system that has a thermal resistance of 0.13 to 0.33 °C/W in heat load ranges from 100 to 320W cooling a heat source that needs to operate between 40-70 °C.

3. It was discovered that the best results can be achieved during intensive heat rejection when the thermal resistance of the LHP condenser approaches that of the LHP evaporator.

4. The thermal resistance of water cooled LHPs are typically a factor of 2 to 3 lower than that of air cooled LHPs.

For air cooled servers, the initial motivation of our work was to double the size of the fins used to exchange heat with a cold air source as well as positioning an LHP condenser at a point in a chassis where the cooling air left it (to avoid recirculation within the chassis). The rule of thumb regarding fan cooling power is that it is inversely proportional to fin area cubed, although in reality doubling fin area probably only delivers a factor of four reduction in fan power.

There is a problem in general with air cooling when efficiency becomes a major issue, and that issue is the total distance that the heat stream has to pass through metal as it passes to the outside world when air cooling is employed. This distance is rather large when compared to water cooled devices. It turns out that in efficient cooling solutions, the overall thermal resistance is usually dominated by the transition points where the heat stream passes through a metal barrier that separates two working fluids. In the case of an

air cooled LHP condenser, the heat has to pass out of a condenser tube that has a .5 mm wall through a Copper heat sink whose base plate is 5 mm thick and then traveling across the base plate another cm or so before having to make its way to the center point of the fin, which is located another cm from the base plate. All together, the heat may have to pass through 2 or more cm of Copper or Copper and Aluminum. In the case of a water cooled LHP condenser, the thermal stream has to pass through the condenser tubing wall, which is only .5 mm thick and is the only thing that separates the working fluid that is condensing from a water stream that removes the heat. The final result is water cooled LHP condensers have thermal resistances that are much smaller than their air cooled cousins. Of course, one of the other problems that needs to be addressed, is the reduction in the number of times the thermal stream has to pass through a metal barrier to change working fluids. The solution we are about to describes reduces the number of transitions from the usual four for air cooling down to two, and only employs efficient transitions.

We mentioned above, the criterion for operating a cooling tower on the hottest day of the year in Atlanta, which was they can provide 30 C water if we can return 35 C water. We did an experiment with water cooled LHP condensers, and after a month of improvements that eliminated recirculation issues, arrived at a device that could produce 45 C water, given 30 C water cooling a 100 Watt Opteron. When combined with other items in a patent that we have been granted, this basically enables what is called “Free Cooling” in hot climates. By eliminating the need for water chillers and HVAC air blowers, this invention cuts out roughly 45% of the total power being consumed to run data centers. Put in slightly different language, where an air cooled data center typically will consume 1 MW to cool a 1 MW IT load, a cooling tower that is used to cool a 1 MW IT load will consume 50 KW or less!

The main problem with our first work, is nobody was interested in using Ammonia in data centers, even if the amount of Ammonia needed in an LHP was small. So, we spent the next three years fine tuning miniature Copper Water LHP designs and along the way discovered that one of the interesting features of this technology is that it is able to reject as much as 1,000 Watts/cm^2! This device demonstrated an important physical property of

LHP evaporator designs: that the heat rejection is not limited by liquid inflows to the escape channels within the evaporator. And, while the operating temperature that we reached at this power density might not be practical, it is possible for the same device to reject 1,000 Watts at much lower temperatures by simply distributing the heat over a larger area, making it possible to cool chips in the 500 to 1,000 Watt heat rejection range. What makes this performance possible is the fact that the boiling occurs on the inside of the escape channel walls and can not block the inflow of liquid that is replacing the vapor that is being produced. For those of you who are still wondering about our pump, it turns out that the distance between the sintered metal particles that make up our evaporator, set the dimension of the bubbles that form on the inside of the escape channel. It is this distance that sets the pressure at which the bubbles burst, which is large enough to drive the cooling loop. Thought about in thermodynamic terms, Loop Heat Pipes employ capillary pressure to drive a cooling loop, stealing the energy to drive the loop from the thermal stream that they move from the evaporator to their condenser. For those of you wondering about the tube that carries the vapor from the evaporator to the condenser, this turns out to be part of the magic: the temperature drop across what we call the adiabatic line is typically around 2 C. The length and diameter of this tube turn out to be much more important than that of the liquid return line that moves condensed working fluid back to the evaporator. As the line gets longer, the temperature losses along the vapor tube increase. As the vapor tube ID increases the losses become smaller, but problems arise threading the vapor tube between obstacles. The vapor tube OD in the case of Copper Water turned out to be 6 mm for an LHP that carries the heat 17 inches. In the case of Ammonia, it was only 2.5 mm, the big difference being that Copper Water runs at an internal pressure of .5 atmospheres where an Ammonia LHP runs at 20 atmospheres.

The performance characteristics of a 17 inch Copper Water LHP are shown in figure 1. The design point for this LHP was a 130 Watt CPU running near the front of a 1U chassis that contained four processors, two stacked in front of the other. The condensers were air cooled using 20 C air that was being pulled through heat sinks that had twice the area of a typical heat sink and were exhausted by a single fan: normally a pair of fans would be employed to cool one of these devices. This LHP was designed to cool the CPU pair at

the front of the chassis. Shorter LHPs were designed to cool the processors at the rear of the chassis. Cooling a device such as a GPGPU that can run at temperatures up to 80 C, makes the point that these LHPs could be used in such an application with devices that were rejecting 250 Watts. Increasing the vapor tube diameter would enable larger heat loads to be transferred. One of the constraints in this design was threading the vapor tube between DIMM modules in the chassis we chose as our design point. Figure 2 shows the same LHP without the fins, which clamp to the serpentine section of condenser tubing on the right hand side. The valve is used in the filling process and then removed. The evaporator is the rectangular object on the right hand side.

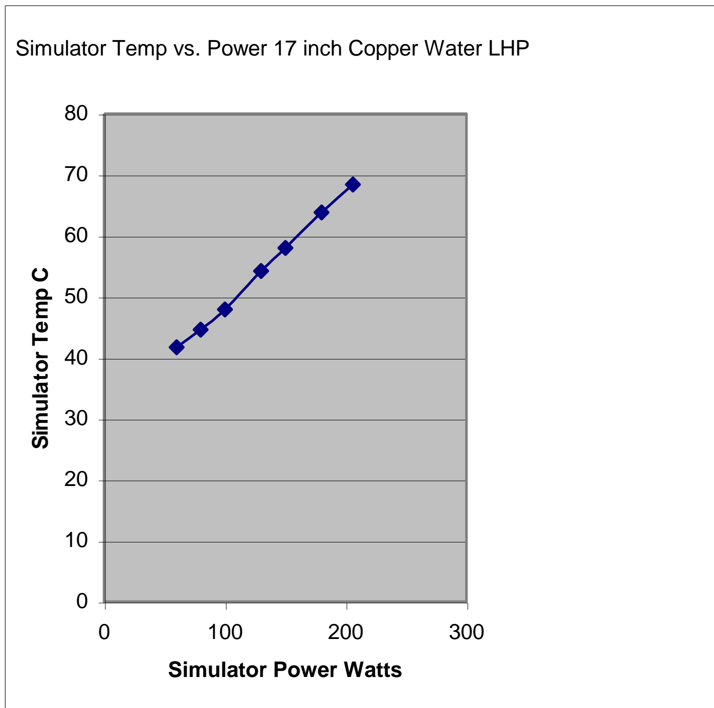

Figure 1

Figure 2