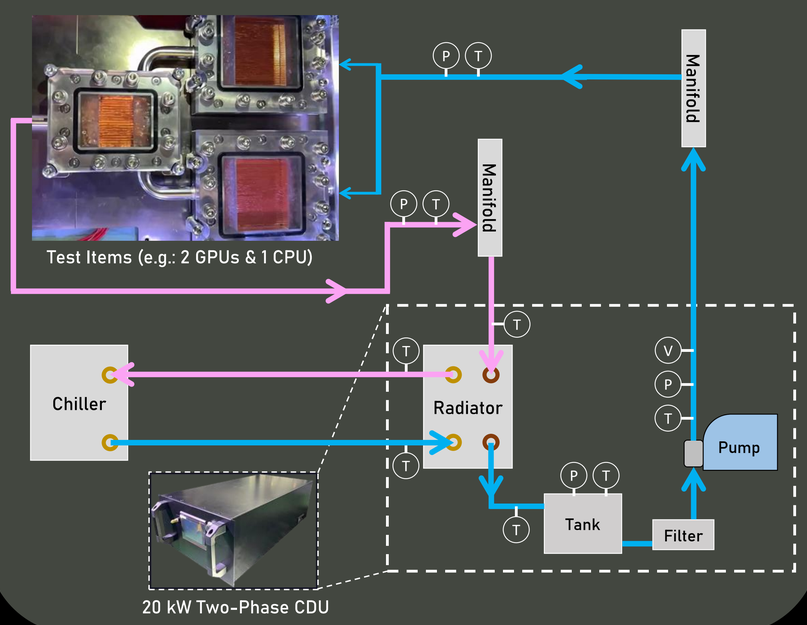
Pumped Two-Phase Cooling for the Next-Gen Chips with Heat Flux over 4 W/mm^2

Ken LC Cheng
Engineer / Wistron Corporation

Introduction

Pumped two-phase cooling technology is considered to be the much-anticipated solution to the incoming extremely-high heat flux of the next-generation GPU. Nowadays, the heat dissipation capacity of a two-phase cold plate has exceeded 3500 W. This presentation will showcase the current pumped two-phase CDU and pumped two-phase loop architecture. The test results of the two-phase cold plate sample is compared with those of the same sample filled with PG25 as the working fluid. Based on the comparison, the potential of the two-phase cold plates can be discovered. Without significantly altering the existing single-phase architecture, the two-phase coolant can be distributed freely to various racks, providing a solution for the chips with locally higher heat flux.

Methods



Results

Fig. 1 shows the standalone test results of a single GPU cold plate. The heating area of the dummy chip is around 8.5 cm^2, and the heat flux is distributed uniformly. The working fluids are R-1234yf (kept at a saturation temperature of approximate 20°C) and PG25 (kept at approximate 20°C), and the flow rate ranges from 0.75 LPM to 4.5 LPM. The TIM is PTM7900. Thermocouples are installed in the groove at the bottom of the cold plate pedestal, and the case temperatures are calculated based on the temperature values of those thermocouples. The case temperatures shown in Fig. 1 correspond to the center measuring points where maximum temperatures are measured. The junction temperatures are also measured by the thermocouples embedded in the heater. Repeatability verification shows that overall variation falls within the range of approximately ±5.0%, with the deviation at the vast majority of measurement points remaining below ±2.5%.

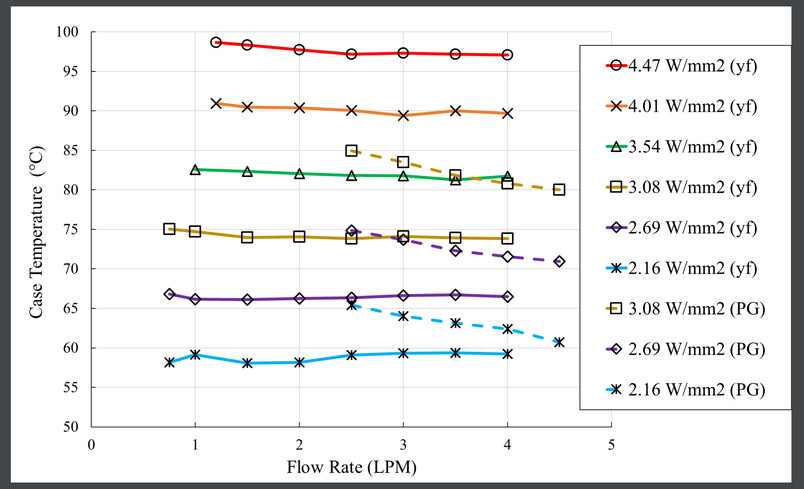


Fig. 1. Thermal performance of a GPU cold plate at various flow rates and heat fluxes using R-1234yf and PG25 working fluids.

Discussion

1. When using refrigerant as working fluid, the variance of the thermal performance is independent to the variance of the flow rate and shows only a weak correlations with the quality change.
2. Using refrigerant as working fluid, when the heat flux increases from 2.16 W/mm^2 to 4.47 W/mm^2, the effective heat transfer coefficient barely varies, which means the fluid is at the critical-heat-flux (CHF) stage. In this study, by optimizing the geometric design of the channel to facilitate a smoother bubble removal, the thermal performance is prevented from rapidly deteriorating due to heat flux exceeding the CHF.
3. Using PG25 as working fluid, the thermal performance improves as flow rate increases. Furthermore, the thermal performance is expected to surpass that of the refrigerant-based cold plate at flow rate exceeding 5 LPM.

Conclusions

1. For the next-gen GPU, the average heat flux may exceed 1 W/mm2, and the heat flux at hot spots may exceed 3 W/mm2. Based on the current heat dissipation capacities, using either refrigerant or PG25 as working fluid is feasible.
2. Whether using refrigerant or PG25, there is an upper limit to heat dissipation capacity, constrained by the thermodynamic properties of the fluid and the flow velocity within the system. However, based on the development trends of GPUs, although heat dissipation is becoming increasing challenging for the next two generations of chips, the problem is not unsolvable.
3. The next action might involve exploring using materials other than copper for the cold plate. The thermal performance may be further optimized by modifying the thermal conductivity.