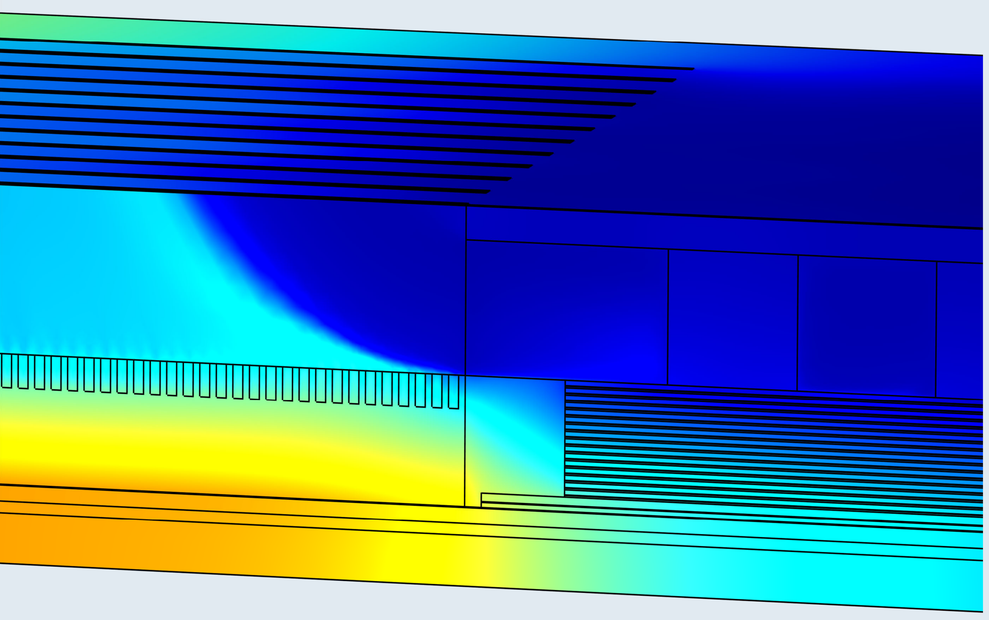


Integrated Microfluidic Cooling of Heterogeneous HBM–GPU Package Exploiting Die Height Mismatch

An embedded microfluidic cooling solution for an HBM–GPU package was simulated using COMSOL Multiphysics. By exploiting the die-height mismatch between the GPU and HBM, the maximum GPU temperature was maintained below 70 °C at a flow rate of only 1.8 L/min.

S. Ahn and M. Nomura

IIS, The University of Tokyo

Summary

The increasing power density of heterogeneous HBM–GPU packages presents significant thermal management challenges. This study proposes an integrated manifold–microchannel cooling structure that exploits the die-height mismatch between the GPU and HBM stacks. Pin-fin structures above the HBM stacks and active microchannels replacing the conventional structural silicon above the GPU bring the coolant closer to the primary heat sources. The manifold shortens the coolant flow path and reduces pumping power without increasing the package height. Using COMSOL Multiphysics, a three-dimensional conjugate heat transfer model was developed for a 2,228 W package comprising a 2,000 W GPU and six 38 W HBM stacks. At a total flow rate of 1.8 L/min, the proposed design maintained the maximum GPU temperature at 342 K (69 °C) with a pumping power of only 0.1656 W. These results demonstrate that a cooling architecture designed around the die-height mismatch can achieve high thermal performance with low hydraulic power consumption in high-power 2.5D packages.

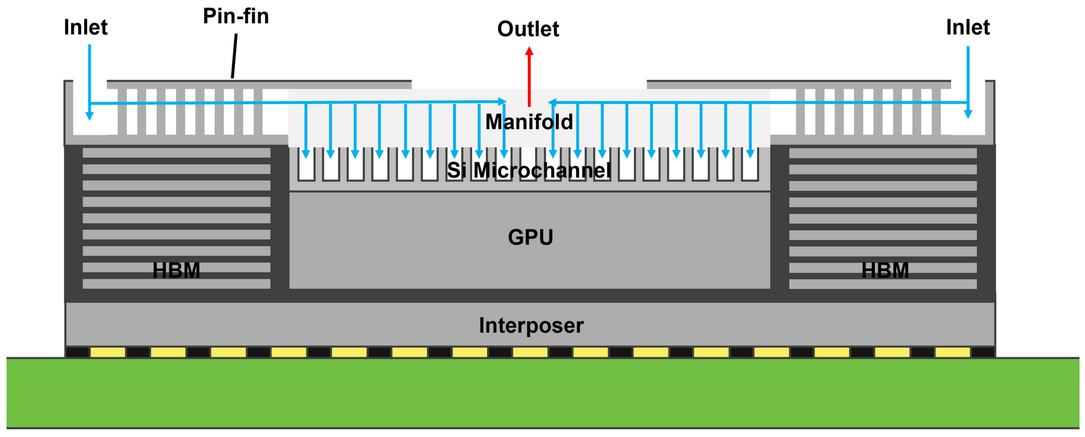


Figure 1. Proposed microfluidic cooling solution for heterogeneously integrated HBM-GPU module with step height variance.

Numerical model

A three-dimensional steady-state conjugate heat-transfer model was developed using the CFD and Heat Transfer Modules in COMSOL Multiphysics®. To reduce the computational cost, a representative unit cell was simulated using symmetry boundary conditions. The package dissipated 2228 W in total, comprising 2000 W from the GPU and 38 W from each of six HBM stacks. The maximum GPU temperature and inlet-to-outlet pressure drop were evaluated at different flow rates. Pumping power was calculated as the product of the volumetric flow rate and pressure drop.

Results

The manifold structure distributed the coolant across the microchannels embedded above the GPU and effectively removed heat from the high-power region. Increasing the flow rate continuously reduced the maximum GPU temperature while increasing the required pumping power. At a total flow rate of 1.8 L/min, the maximum GPU temperature was maintained at 342 K (69 °C), with a pumping power of only 0.1656 W. These results demonstrate the potential of die-height-mismatch-based microfluidic cooling for high-power 2.5D packages.

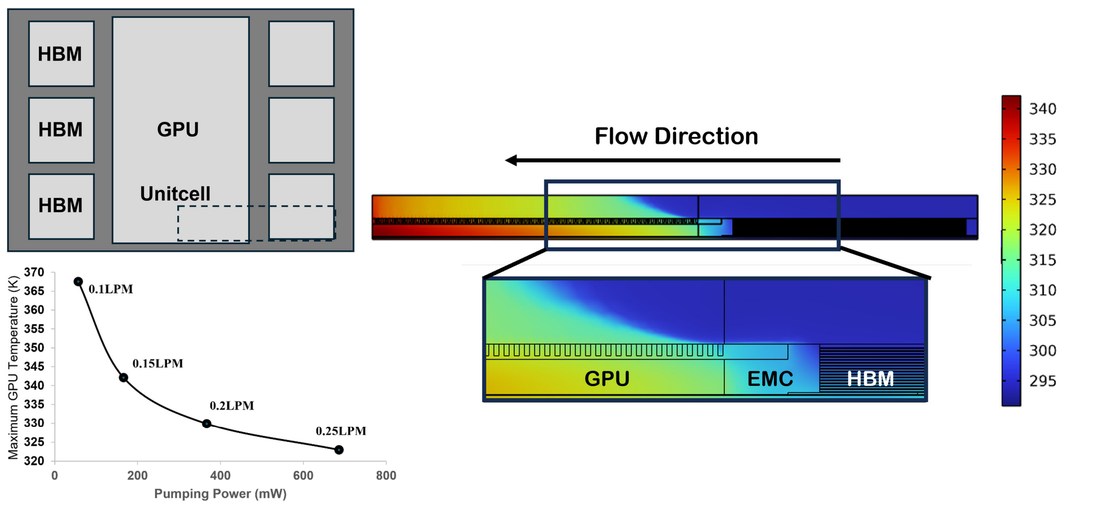


Figure 2. Simulation of the HBM-GPU module based on proposed cooling solution.

参考文献

E.Chung, M.s. Bakir et al, “Integrated Microfluidic Cooling of Heterogeneous HBM–GPU Package With Step Height Variance,” ***IEEE TCPMT***, 2025

東京大学 生産技術研究所
Institute of Industrial Science, The University of Tokyo