DF Tech Note 2512-01v1

3D Chip Thermal Solved with Edge Overhang Cooled Diamond Spacers

Chips packaged to place memory on top of the GPU in 3D enable much shorter wiring distances, resulting in distinct advantages in terms of performance (latency, memory bandwidth, etc.) and much – 6.7x according to TSMC – higher energy efficiency. The potential of true 3D is to avoid the energy lost in the long paths connecting compute and memory (which are far higher than just the Phy-to-Phy losses often cited when one properly includes the logic and GPU internal paths as well). Avoiding such energy loss not only saves energy but reduces heat management challenges in the first place.

But thermal challenges abound with 3D as HBMs are very sensitive to heat, and GPUs produce lots.

This Tech Note describes the surprising discovery that a spacer between memory and compute made out of a single-crystal diamond (SCD) – one which overhangs the compute die by a few mm's on each side and is cooled at such overhang only – is sufficient, even with standard cold plate cooling capacity, to solve the downside of heat concentration of 3D chip packaging.

Chip Packaging

In CoWoS/EMIB 2.5D chip packaging worlds, compute dies and memory stacks are placed side by side on top of an interposer that provides the connectivity between them. This means the wiring is generally longer than the chip size of close to 33x26mm.

3D packaging reduces wire length by effectively 10x. This improves the energy needed to move data with I/O energy per bit improvements as high as 10x. It also reduces interconnect delays and capacitance by placing memory on top of the logic.

But this comes at the detriment of memory, which is heat-sensitive and now being effectively fried in the logic frying pan.

It turns out SCD interposers can help -- in a surprising way: Placed between memory and compute, SCD is so thermally conductive that it can pull heat out laterally over the entire chip size -- just by cooling it in a standard way at its edges.

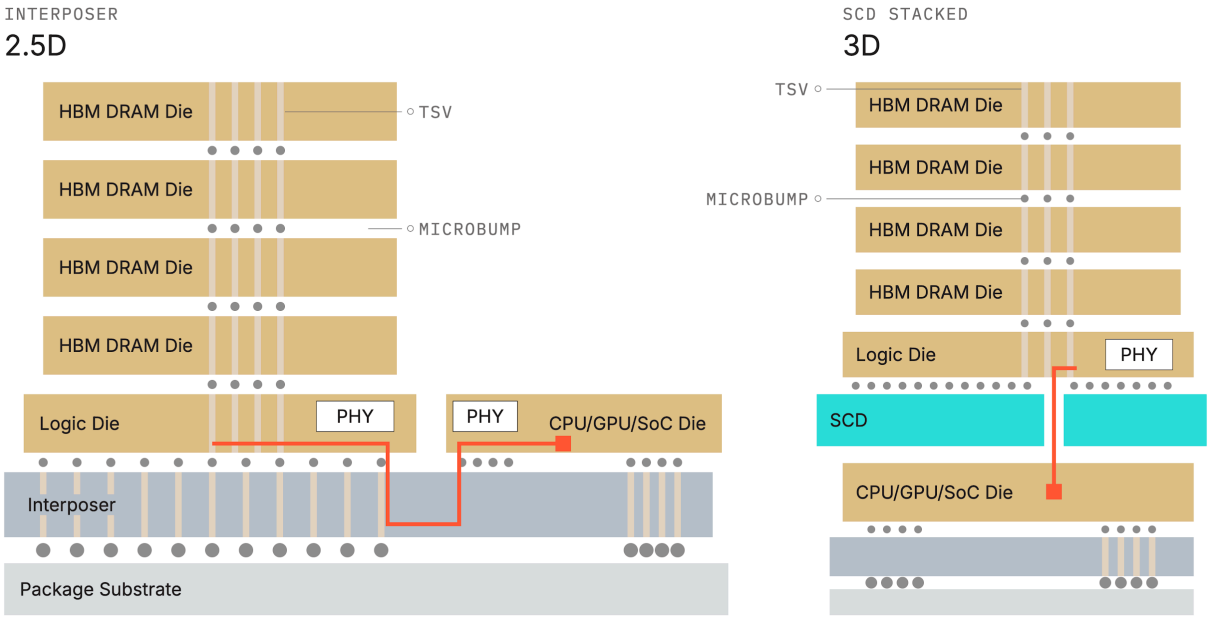

Figure 1: Moving from 2.5D (left) to true 3D (right) reduces wire lengths by as much as 10x. For a 10mm wide HBM and a 26mm wide GPU, while the Phy-to-Phy connection has been optimized to be very short, the path length can easily be 18mm in a 2.5D package whereas in a 3D package it can be less than 2mm even with a 1mm thick thermal spacer.

3D Chips with an Edge Overhang Cooled Diamond Spacer

Introducing a spacer made out of SCD with an edge overhang of several mm's at which cooling is applied might be a simple solution to enabling 3D stacking.

Is cooling a diamond spacer at a slight overhang enough to cool an entire 3D chip package? SCD is obviously superb at thermal conductivity – see Appendix – but still, is this enough to provide benefit for an ASML reticle-limit sized chip packaged in 3D?

If the cooling is just connected to the edge face, it turns out not so – but if there is an overhang of several mm's, it surprisingly turns out that yes, it is. (In a way, it evidences the >20x higher thermal conductivity of diamond vs silicon, which effectively allows heat to travel 20mm in diamond (half the chip size) when it would travel only 1mm in silicon.)

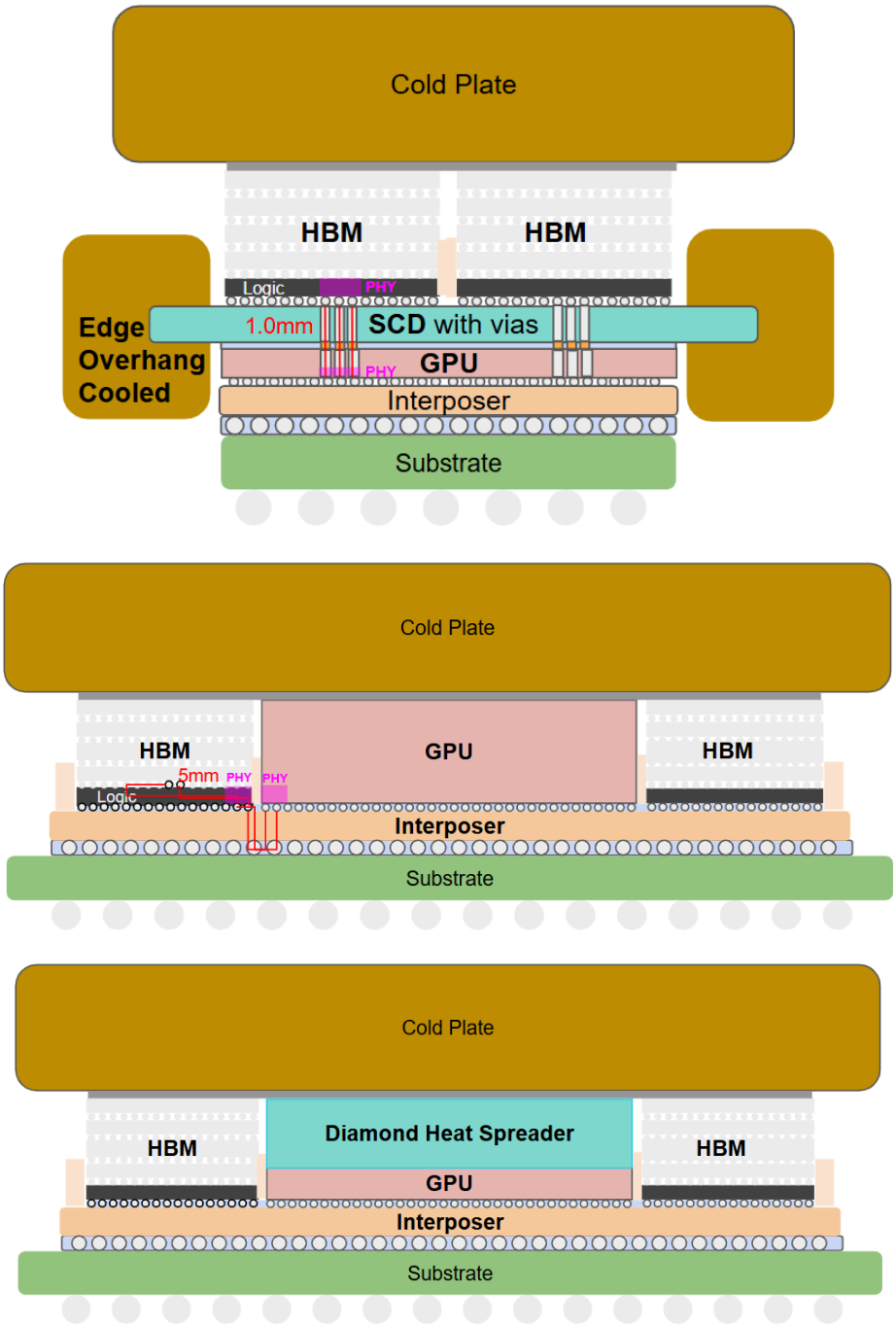

Figure 2: A 3D package with thru-vias SCD with edge overhang at which cooling is applied (top), a standard 2.5D CoWoS package with memory side-by-side the compute logic (middle), and a thermally optimized 2.5D package with thinned silicon and diamond heat spreader (bottom).

In other words, it turns out that SCD is so thermally conductive that it can laterally dissipate heat of ASML reticle limit-sized chips, provided there is an edge overhang that cooling can connect to. This overhang can be as little as a few mm's on each side. (Indeed, there is negligible benefit to an overhang much larger than 5mm.)

AI Chip Thermal Model

To understand the behavior of a diamond spacer's high thermal conductivity on the heat and power scaling of high-power chips, thermal simulations have been completed based on an Nvidia GH100-inspired power map with a hotspot for each tensor core and 80% of the chip power located in the hotspots and 20% of the chip power in the background.

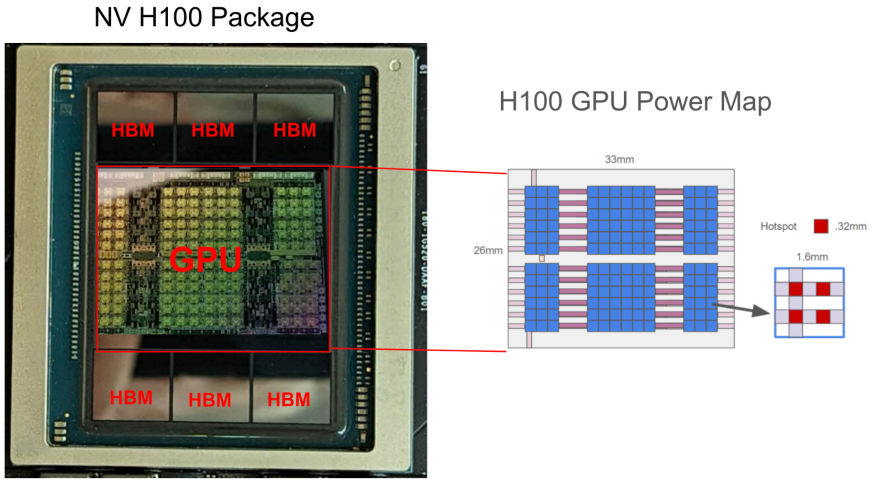

Figure 3: Hotspot map inspired by the Nvidia GH100 as used for COMSOL FEA simulations.

The following figure outlines the geometry in more detail as used for thermal simulations:

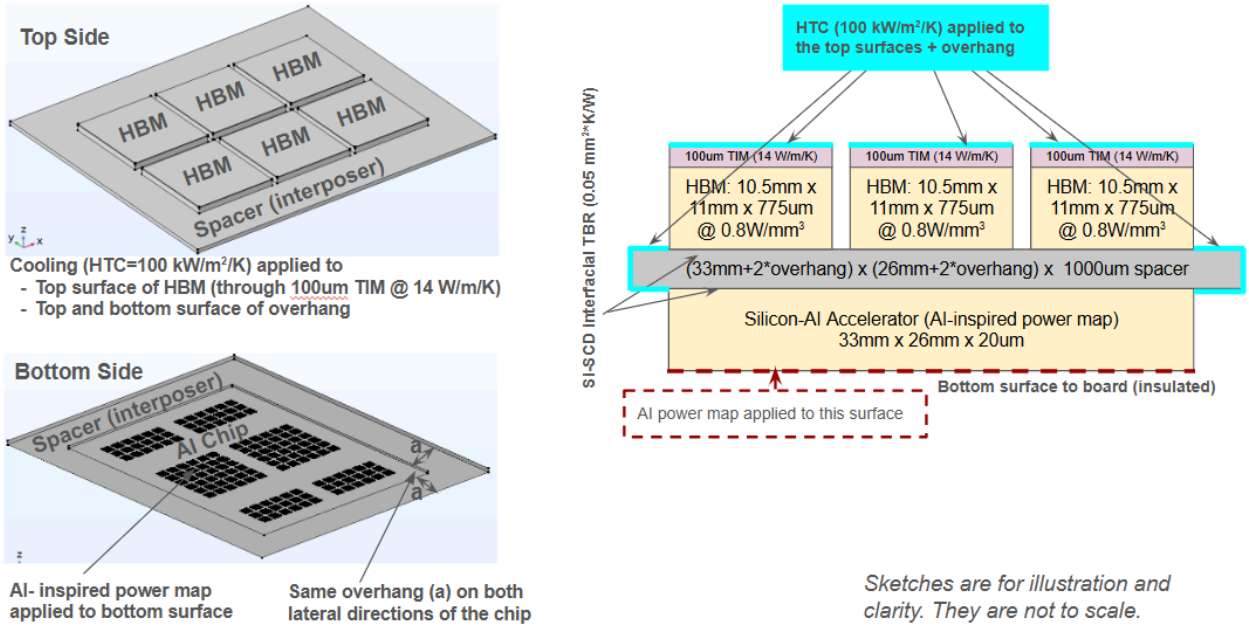

Figure 4: A 3D chip package connecting an AI accelerator chip to the HBM via an SCD spacer that overhangs by several mm and is cooled with standard cold plate cooling capacity at edge overhang. (Vertical interconnects were not simulated.)

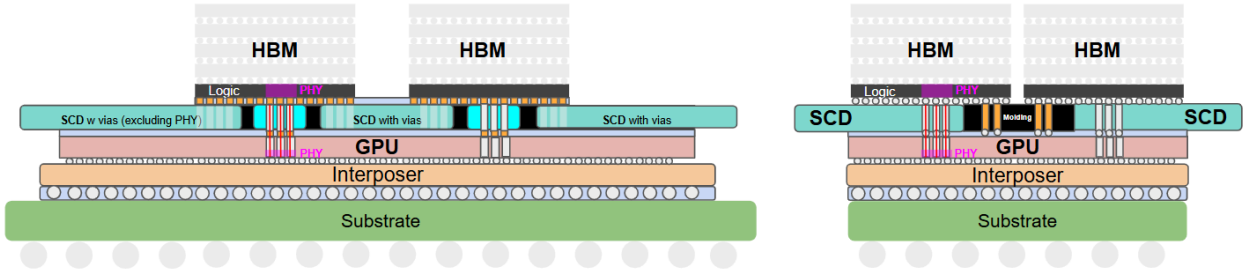

Figure 5: Variants of SCD spacers applied to 3D chips. The SCD spacer may be cut to have openings into which silicon or molding provides vertical interconnects – while still enabling lots of area for thermal dissipation.

Heat Behavior and Power Scaling: Thermal Simulation Results

Figure 6 below shows the simulations results:

Stacking memory on top of logic yields lower performance due to temperature rapidly exceeding 90ºC. Adding a diamond spacer without cooling helps spread heat but is not sufficient to keep the temperature low. It is also not sufficient to have a diamond spacer without overhang just cooled at its edge faces (equivalent to zero on the x-axis in the charts below).

However, by adding an edge overhang to the diamond spacer and cooling that overhang area turns out to be sufficient to cool the entire 3D stack of a chip package and completely eliminate the heat penalty of going 3D. Indeed, just several mm of overhang on all four sides suffice.

A diamond spacer does not just eliminate the heat penalty of 3D packaging an ASML reticle-limit sized chip – but it can support up to 1.6x the power.

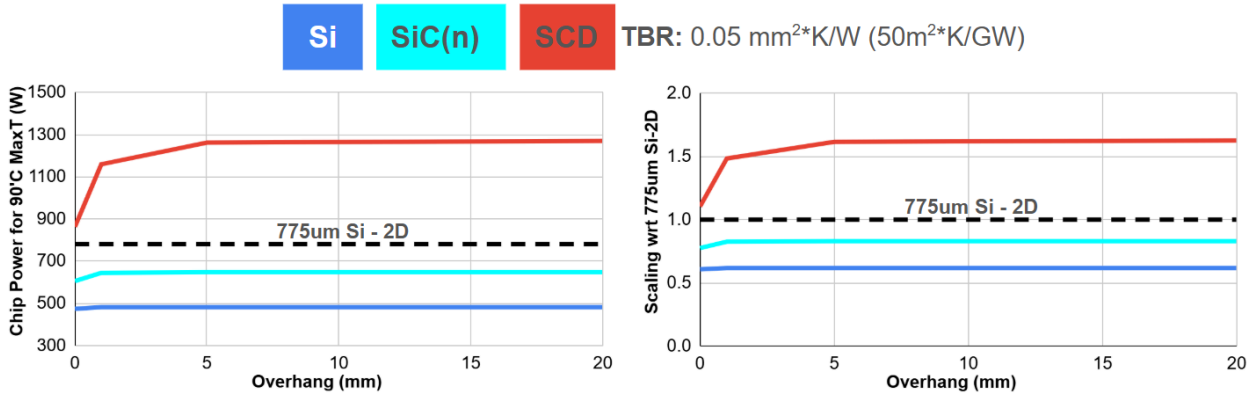

Figure 6: Thermal simulation results of a diamond (Si, or SiC) spacer with different amounts of edge overhang used for cooling for a stack of 20um-Si/1mm-spacer/HBM with a TBR = 0.05mm²-K/W. 775um Si - 2D represents the baseline without HBM on top of the compute chip. Absolute power is shown on the left, relative power scaling compared to the 775um Si - 2D baseline is shown on the right.

By reference, the use of silicon carbide (SiC) was also simulated. However, its much lower and anisotropic thermal conductivity does not allow a SiC spacer to work thermally in a 3D package.

Conclusion

A concept has been outlined and substantiated with thermal simulations by which the extreme thermal conductivity of diamond is sufficient to laterally dissipate heat as a spacer in a 3D chip package, with cooling applied at the edge overhang of the diamond. It allows any number of layers stacked in 3D.

Cooling only the edge face of the spacer does not deliver sufficient thermal bandwidth for heat sinking. But an edge overhang of just a few mm surprisingly can do the job.

As a solid-state solution, this approach promises to be very robust and reliable. Unlike microchannels, there is no possibility of fouling, clogging, flooding, or plaque build-up in arteries; and proven cold plate solutions can be applied with low cost.

The very substantial temperature reduction delivered by SCD – exceeding 50 degrees – extends chip lifetime materials-wise by as much as >25x, based on the rule of thumb that each 10^{o}C reduction doubles a chip's lifetime, and yields much improved energy efficiency.

Given extensive capital investment in datacenters, chips actually holding up over their lifetimes is elemental; and energy consumption is a key gating factor for growth.

In order to implement this concept, a diamond spacer with tens of thousands of small through-vias would need to be built with low cost.

Appendix: Single-Crystal Diamond Reference Info

SCD's isotropic thermal conductivity exceeds silicon's by up to 23x, copper's by up to 8x, and SiC's anisotropic thermal conductivity by up to 8x:

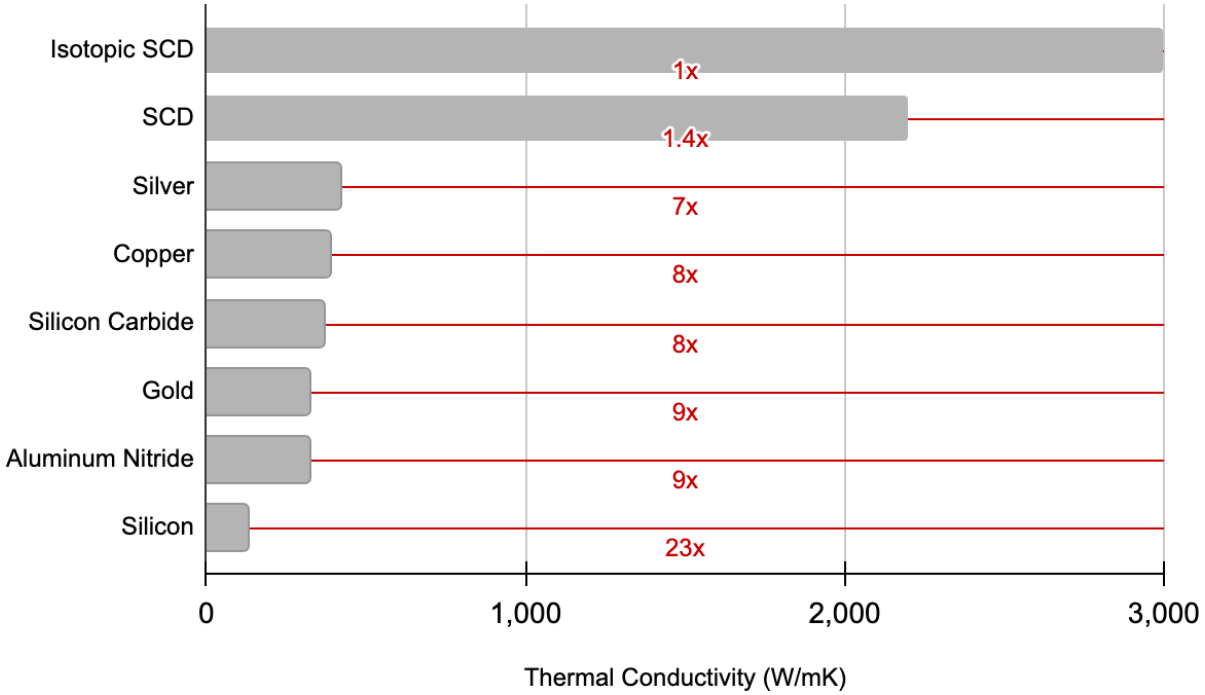

Figure 7: SCD is the highest-performing thermal conductor – by a large factor.

DF's SCD is electrically inert, mono-crystalline, and isotropic, which means that unlike polycrystalline diamond, graphene, and silicon carbide, it delivers equally in all directions (polycrystalline diamond is also more expensive to grow than single-crystal diamond – and very expensive to polish to the kind of precision-surface finish required for chip wafer bonding).

DF's SCD substrates have been measured by third parties with various thermal conductivity methods, consistently showing a thermal conductivity of >2,000W/m-K at room temperature. The methods used vary from laser-based techniques like thermoreflectance to an IR camera-based technique monitoring steady-state heat flow.

SCD's coefficient of thermal expansion (CTE) at 1ppm/°C is very similar to the CTE of silicon at 2.5ppm/°C (compared to 17ppm/°C for copper), which reduces thermal stress and further improves reliability. Silicon and SCD are chemically very compatible – they have the same crystal structure and are underneath each other in the periodic table of the elements. They form very strong bonds by Surface Activated Bonding (SAB). Furthermore, SCD has a high Young's Modulus (1050GPa), which allows both silicon and SCD to be thin and maintain sufficient stiffness.