SEMICONDUCTOR ENGINEERING
DEEP INSIGHTS FOR THE TECH INDUSTRY

True 3D-IC Problems – April 27, 2023

True 3D-IC Problems

Stacking logic requires solving some hidden issues; concerns about thermal dissipation may be the least of them.

APRIL 27TH, 2023 - BY: **BRIAN BAILEY**

Placing logic on logic may sound like a small step, but several problems must be overcome to make it a reality.

True 3D involves wafers stacked on top of each other in a highly integrated manner. This is very different from 2.5D integration, where logic is placed side-by-side, connected by an interposer. And there are some intermediate solutions today where significant memory is stacked on logic, such as found in HBM stacks.

The first true 3D-ICs are close. “There will be some test chips happening toward the latter part of this year that involve full logic-on-logic,” says Vinay Patwardhan, product management group director in the Digital & Signoff Group at Cadence. “By the middle of next year, we can expect some realistic logic-on-logic chips, particularly something with multiple AI cores. Those companies are running out of area on a single die. Many of these designs are close to the reticle size limit, something over 600 or 700 millimeters squared. They are desperately trying to go full-3D stack for the next-generation design because it does not require too much architectural change. But cutting them and stacking them is a physical change.”

Other targets for 3D may be further out, such as the ability to stack heterogeneous dies. “That would require a true 3D placer and router that was working on a heterogeneous stack,” says Rob Aitken, a Synopsys fellow. “It would have to know that to build any logic path that crosses the dies, you’d need two separate libraries. They could quite well be two different technology nodes, but existing tools and flows make the assumption about consistent libraries, and those assumptions are baked quite deep. It’s not inconceivable that the tools can be modified to deal with that. But there are some base assumptions of the tools that would need to change.”

Physical issues
One of the key learnings from 2.5D integration is that significant mechanical issues exist at the junction of two dies. “Whenever two dies are integrated together, you introduce stress,” says

Thomas Uhrmann, director of business development at EV Group. “If you look at problems with interposers, most of the fracture points are at the connections, and that creates reliability issues. You shouldn’t underestimate the complexities that come from dealing with a mix of materials. In the middle of the die, you may have underfill. When you cure that, it shrinks. That creates stress, even though it stabilizes the connections. With 3D integration, that problem shifts to another dimension.”

Those problems are reasonably well understood at this point. “More interesting heterogenous stack questions come up when you start mixing different materials,” says Synopsys’ Aitken. “When you’ve got CMOS stacked on CMOS, even if it’s a different node, it’s probably going to behave mechanically in a way that makes sense. If you decided to stack a gallium nitride device on silicon, or put a layer of some other objects, there’s a bunch of cool things you could do. But you start to get some interesting mechanical questions that require a lot of thought.”

Another physical issue is heat. “Thermal is probably the biggest challenge today,” says Tony Mastroianni, advanced packaging solutions director at Siemens EDA. “While HBM is doing stacks of 12 die, that is a very different problem because it is memory, and you’re only going to enable one of those stacks at a time. They’re not all firing at the same time. They do not have to worry about thermal management. The practical limitation today is probably three die, and even that’s going to be a challenge.”

But it is not all bad news. “The 3D configuration helps a little bit by reducing dynamic power,” says Cadence’s Patwardhan. “With 2.5D, a signal has to travel across a large die and then travel on the interposer to another die, resulting in long wire lengths. When you have a stacked die, you can route in the Z direction, reducing the wire length. So the dynamic power, the switching power, is reduced in a 3D configuration. If the stacking happens correctly, where the switching elements on both dies are not switching at the same time, you can effectively use a 3D stack to reduce the power or thermal footprint. If there is too much switching happening on both tiers at the same time, thermal effects, the chimney effect, comes into play.”

That is one application for 3D technology. “If you believe in the concept of dark silicon, that not all of the device needs to be on all the time, then you could conceptually architect a 3D stack in a way that you were able to manage the thermal so that power and the heat were not a problem,” says Aitken. “You could take something that previously implement as a large 2D object and implement it as a smaller 3D object instead.”

In any case, it will require early analysis. “You must do some thermal characterization very early on, before place-and-route,” says John Park, product management group director in the Custom IC & PCB Group at Cadence. “You need the ability to put in parametric data describing the power dissipation of each chiplet, the molding compound to be used, parametrically describing what you expect a potential heatsink to look like, the size of the package it is going on, because that’s a natural part of the system that helps distribute the heat, and even the size of the PCB, which further helps. In the prototyping stage, you are starting to look at what things can be stacked, and even in a 2D world, how close they can be to one another, what types of chips or chiplets are best suited for stacking based on early knowledge about the design.”

Inter-die connectivity
When communicating between chips, and even within a chip across an interposer, high-speed communication requires sophisticated PHYs, SerDes, and communications protocols to ensure the reliable transmission of data. "You need PHYs for high-speed interfaces in 2.5D because you're driving up to two millimeters," says Siemens' Mastroianni. "You have to worry about timing and synchronization and dealing with signal integrity issues. But with true 3D, since the logic is nanometers or microns away, you can just use regular gates, the regular standard cells. They do have special cells that have a little bit of ESD built in, but essentially you don't need those PHYs. Instead, you just have those logical interfaces talking through regular logic. You have to do some synchronization for clock but that's normal STA logic type stuff and timing optimization."

That creates some different problems. "You have the opportunity for many more interconnects between the stacks, and putting tens of thousands of PHYs is a non-starter," says Aitken. "But you do have to be concerned about the testing of these things, the sign-off of these things. What are you actually going to drive? Will you have an inverter that drives up to a piece of metal and connects to a matching buffer on the other side? Or are you going to put in some kind of a MUX so that you can do some testing? Or are you going to try to contact them for wafer probe, or are you going to forget about the whole thing and not test it until you build it?"

The design community is attempting to answer those questions. "There is normally pre-bonding and post-bonding tests that OSATs perform," says Patwardhan. "Direct probing of these micro-bumps, which are less than 10 microns, may or may not be possible with today's testing techniques. A lot of the testing happens through test paths that are defined across the two dies. They insert programmable e-fuses that can run open circuit tests. We have to make sure that whatever test insertion we do, from an EDA point of view, we follow the emerging IEEE 1838 standard and make sure all those checks are available through the full EDA flow. Testing will evolve as these hybrid bonds become more mainstream."

Models and automation
3D-IC will require some significant upgrades to existing EDA tools and flows.

"We call them the three M's," says Marc Swinnen, director of product marketing at Ansys. "It is multi-physics, multi-scale, and multi-organizational challenges:

1. There is a multi-physics challenge with things like thermal, electromechanical, mechanical, and electromagnetic. These are issues that traditionally the chip designer didn't have to worry much about, except for the RF guys.
2. There are multi-scale challenges when you go from nanometers on a chip through millimeters on a package through centimeters on a 3D-IC interposer. That is six orders of magnitude you're encompassing. Traditionally, those have been handled by three different sets of tools. For 3D-IC, these all need to be consolidated into one.
3. It has become a multi-organizational issue. The skills for this do exist in industry, but they're separated across different teams, different companies sometimes. For 3D-IC companies, they will have to re-architect their organizations to bring together a single team that encapsulates all the expertise you need to solve this. You can't throw it over the wall to some remote team, and then to another team across the world, and back to the design team."

Many problems are extensions to ones being used today and may be introduced in phases. "The first phase will probably support homogeneous dies," says Mastroianni. "All the dies will be in the same technology. That's makes it a little bit easier, but ultimately to really leverage this technology, you want to be able to leverage different process technologies, different nodes. That will require common data models to be able to perform timing closure. In addition, when all devices are not on the same die, you can't assume that they're all fast fast, or slow slow. You have to deal with that. On-chip variation is a statistical technique where you make some assumptions about how much your timing can vary. It's basically overhead margin you're building into your design. But if you have different chips that are manufactured in separate runs, you can't assume any correlation, they're totally uncorrelated. So, you have to do a more extreme corner optimization analysis."

Not only do the place and route tools have to be rearchitected for the Z dimension, they also have to become a lot more thermally aware. "We already have 2D design tools that are activity aware," says Patwardhan. "The placer and other tools have the ability to take a VCD file, representing worst case activity. You can aggregate that from a simulation, and then the placement of cells happens in such a way that the hotspots are spread out. It is an iterative flow where we do the first placement, do the cropping, and after clock tree synthesis, we can refine the placement using some of the activity data. That is a power density-based flow. This can be expanded to 3D. We are working on this and have some early prototypes, where we can take activity information and then use early thermal analysis, either based on your static currents or full dynamic activity, and based on that decide 3D placement. We have enhanced our 2D placement engine now, we have to expand it to take the Z dimension and it's a multi-objective placer. Thermal effects can be directly modeled as an objective to the placer."

Another simplification likely to be seen in early tools is a restriction on where a Z dimension partition will be made. If macro cells or IP blocks are kept on a single die, they can be signed-off within a die and not have to wait until the entire stack is logically assembled.

"There are people talking about removing that restriction," says Park. "They call it macro-folding. In the analog world, they call it circuit folding. If you have a really small form factor in a planar sense, but they have some space to go vertically, there are people talking about folding macros on top of each other. I don't know of any design where that's actually in production, but there are certainly some of our customers talking about that capability. With folding you can make it half the size in a planar sense, and just a little bit thicker in a vertical sense."

Conclusion

When migrating from one technology node to another, there are invariably new effects that have to be taken care of, and new restrictions or limitations that are placed to ensure an easy path to sign-off. With that come new opportunities that can be exploited.

The migration from 2.5D to 3D makes those previous migrations look simple. 3D-IC is disrupting every aspect of the tools, models, flows, and even organizations. It would appear that the industry will tackle those problems when someone has the need and desire to solve it. But until that happens, it may be left as an interesting issue for the future.

https://semiengineering.com/true-3d-ic-problems/