The Heat Problem

In the semiconductor industry, you will often hear complaints about how ‘hot’ things are getting. No, that does not just mean the rising demand for chips, but the increasing temperature of the devices themselves, and how difficult it is becoming to keep them cool.

For the past half-century, the race has been to make transistors (devices capable of acting as logic components in integrated circuits) smaller, thereby packing more and more of them onto each chip. The more transistors we have, the more switching can exist on a single device. This strategy has powered exponential growth in computing, but it comes with two major challenges: manufacturing transistors at extremely small scales is increasingly difficult, and keeping these dense chips from overheating is even harder.

To understand why, think of integrated circuits, a primary component of chips, as amalgamations of many tiny components. These guide the flow of electrons, but, since no material is perfectly conductive, when current flows, resistance causes some energy to be lost as heat. The more current a device handles and the denser the circuitry, the more heat it produces.

But in modern chips, this is not even the biggest issue. Most heat comes from the act of switching–every time a transistor switches, it charges or discharges tiny capacitors. The dynamic power used in this process scales with the square of the voltage and the switching frequency. In the brief period that a switch is neither on nor off, voltage is still being applied, but now the element is significantly less conductive and electrons scatter more, causing heat to be generated. For one transistor, this is negligible, but when we scale, it can become a significant problem. The higher voltage we apply and the more we switch, the more heat is generated. With potentially billions of transistors switching billions of times per second, the little bits of heat generated compound to become a significant problem.

So, why is a lot of heat a problem? Transistors rely on precise control over whether current flows (‘on’) or not (‘off’). When the temperature is higher, electrons in a semiconductor acquire energy, making it more difficult to keep them fully off. Further, heat increases the tunneling probability of electrons (the odds they ‘flow’ randomly, even if they do not have enough energy and are not supposed to) and thereby can blur the distinction between states. On top of that, very hot chips suffer from reliability issues: interconnects can slowly degrade via electromigration, and insulating materials between individual devices can fail, causing short circuits.

Since today’s devices are built at the nanometer scale, even tiny amounts of leakage or variability can cause major failures, which is why temperature management has become a bottleneck. The most viable approaches to solving this issue can be understood as two distinct categories: either use new materials for integrated circuits or introduce heat reduction methods into systems.

An intriguing alternative to highly monocrystalline copper interconnects being explored is a class of materials dubbed ‘topological semimetals’, or TSMs. While they are not new materials, their potential as replacements for copper is a relatively recent development, as more is being understood regarding their unique electrical properties.

As chips have become smaller, and the number of transistors we can template per unit area has increased, the size scale of interconnects has reached the nanometer level. At this size scale (10’s of nanometers), the length of interconnections is comparable to the electron's mean free path in copper.[1] Thus, as opposed to reaching the end of the copper wire without scattering, the probability of an electron interacting with a wire surface or grain boundary begins to increase. This means that the dimensions of wires dictate the mean free path, not the material, and the resistivity of the interconnects is beholden to the size of the system. This is a problem if sub-nanometer systems are to be functional.

What’s unique about TSMs is that they somehow do not suffer from this scattering issue. At the nanoscale, many odd properties start manifesting, and in TSMs, we see a resistance to surface scattering.

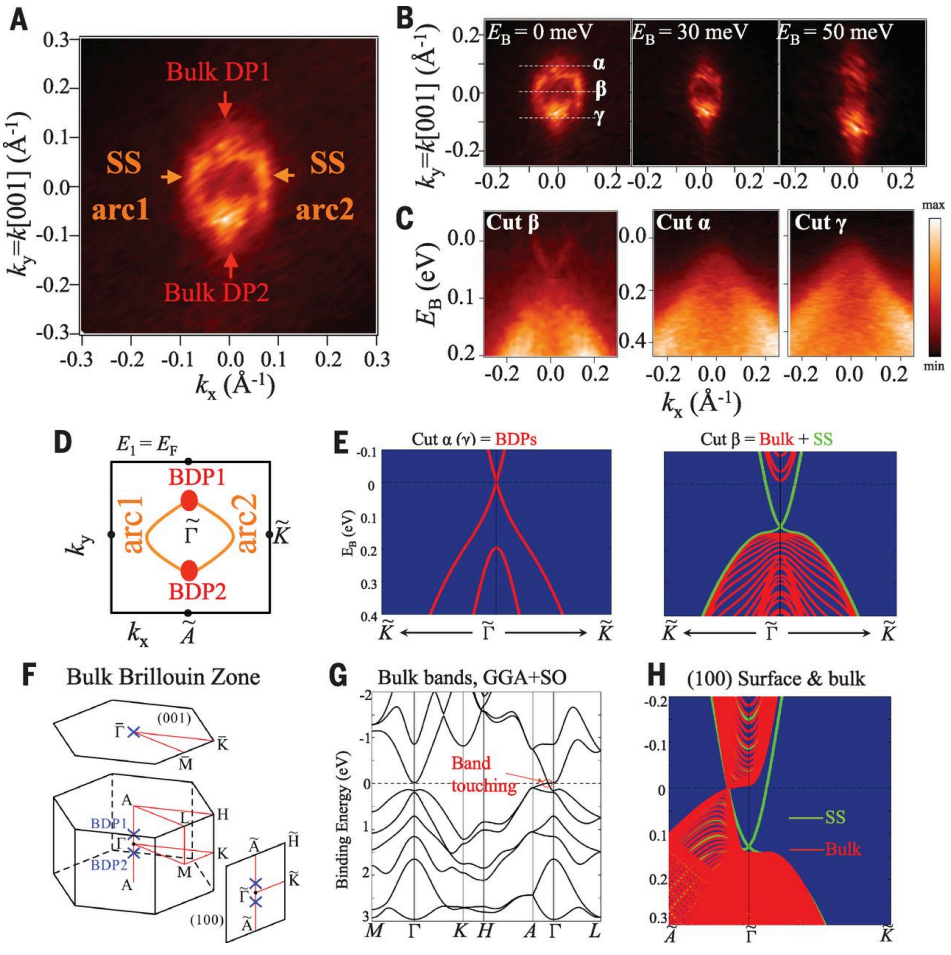

Figure 1: ‘Fermi Arc’ representation and proof in TSM[2]

As opposed to standard metals, which have no band gap, TSMs have several points at which their conduction and valence bands touch. These intersections, or 'Fermi arcs'(Figure 1), function as the so-called 'charge pathway' of the system, as opposed to the bulk of the material. As the name of the material type suggests, this conductivity path is a surface property(not bulk) of TSMs and thus is largely unaffected by changing the dimensions of the material. Since they function as the dominant conduction mode at the nanoscale(due to surface area becoming more significant in terms of cross-sectional volume), surface scattering is not a factor in TSMs. So the conductance of the material remains high while the resistance remains relatively low.

Of course, TSMs are far from a replacement at the moment; they are simply too difficult to synthesize consistently at high purity, and verifying whether a sample even displays the relevant properties requires extremely expensive and finicky equipment and experienced researchers.

For cooling, microfluidic systems are by far the most novel and effective method. The concept is simple: etch a channel through which coolant can flow at heat centers in integrated circuits, and you can dissipate the heat that accumulates. Further, as opposed to brute forcing channel locations, artificial intelligence is being implemented to analyze where heat tends to accumulate in GPUs, and cooling is focused on these hot spots. The relevant physics is certainly nothing complicated, but simple thermodynamics–a constant flow of some fluid that is at a lower temperature than its surroundings will absorb heat and then move out of the system. What's unique is the biomimetic nature of microfluidics; as opposed to creating new materials like with TSMs, the industry has focused on replicating cooling processes we see in our bodies every day and implementing them at nanoscale dimensions–and it works.

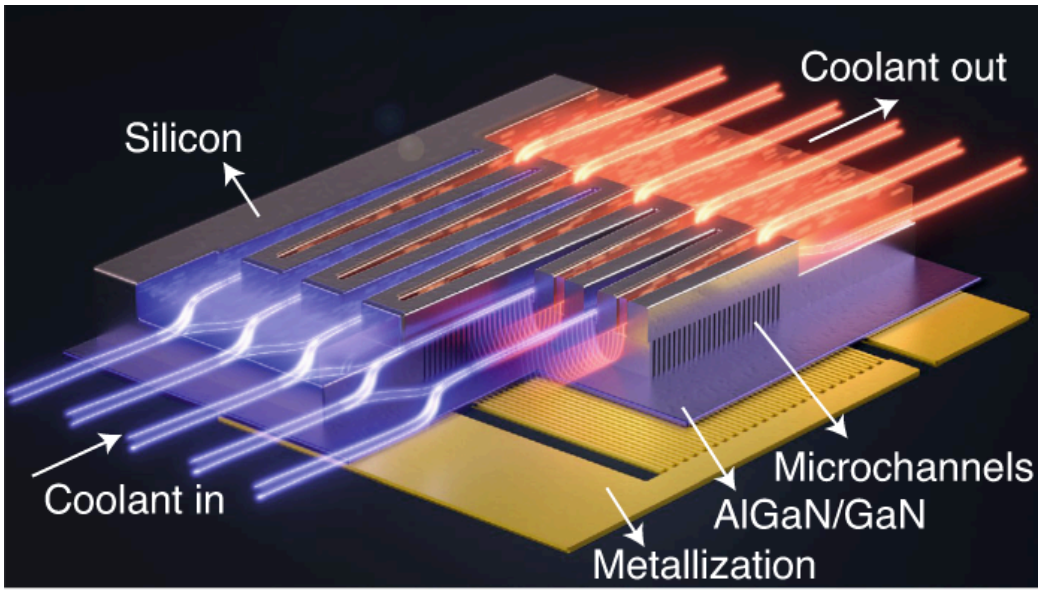

Figure 2: Diagram of microfluidics in a silicon chip[3]

Microsoft reportedly saw a decrease of 65 percent in the maximum temperature of silicon in a GPU.[4] Unfortunately, this incredible performance does not tell the whole story. Mass-producing accurately etched cooling pathways that are on the nanoscale will be just as difficult as creating transistors of that size, and that has taken decades of research to implement in industry. Further, cooling leaking from channels to the rest of a GPU could be catastrophic for a chip's function; thus, the systems are likely extremely delicate, which also makes them difficult to implement in industry. Still, for micrometer-scale devices, microfluidics is certainly viable and may be commonplace very soon.

So, engineers have been exploring topological semimetals as an alternative to monocrystalline copper, and microfluidics is being refined to be mass-produced and reproduced at the nano-scale. Still, there is no viable solution yet, and the semiconductor industry relies on inefficient cooling systems to keep stacks of GPUs within safe operating limits. But as demand for computing power keeps rising, finding new ways to keep chips cool is going to be at the forefront of technological progress.

Works Cited

1) Electron mean free path in elemental metals | Journal of Applied Physics | AIP Publishing. https://pubs.aip.org/aip/jap/article/119/8/085101/143910/Electron-mean-free-path-in-elemental-metals.
2) Xu S-Y, Liu C, Kushwaha SK, et al. Observation of fermi arc surface states in a topological metal. *Science*. 2015;347(6219):294-298. doi:10.1126/science.1256742
3) Microfluidic cooling: Microfluidics explained. Darwin Microfluidics. August 23, 2024. https://blog.darwin-microfluidics.com/glossary/microfluidic-cooling-microfluidics-explained/.
4) AI chips are getting hotter. A microfluidics breakthrough goes straight to the silicon to cool up to three times better. Source. September 24, 2025. https://news.microsoft.com/source/features/innovation/microfluidics-liquid-cooling-ai-chips/.