Moore's law is the observation that the number of transistors in a dense integrated circuit (IC) doubles about every two years. Moore's law is an observation and projection of a historical trend. Rather than a law of physics, it is an empirical relationship linked to gains from experience in production.

Moore’s law is effectively dead because of thermal limitations *i.e.* the constraints of how much power can be dissipated from a chip die.

Note that that this picture is confused by focusing on the number of transistors only, as opposed to the effective computing power for typical instruction mixes and workloads. What good are hundreds of millions and billions of additional transistors if their utilization keeps dropping?
Since the dawn of the super-scalar out-of-order Intel architecture age, way back in 1995, more and more parallelism in form of additional transistors was thrown at the problem, with exponentially less utilization of those additional marginal transistors. Nowadays most of those additional transistors sit idle with ever-decreasing fraction doing work for typical sequential workloads. For it would be physically impossible for all of them to be running at once as that would burn up the die.
Note that the thermal envelope (TDP) of contemporary processors of about 120 or so watts was reached as early as 2004 with Pentium 4, which incidentally ran at 3.8GHz. It is no coincidence that 3.8GHz is near the top now, 15 years later.
It is now largely forgotten that clock scaling was the earliest and most well-known form of Moore’s law. It was truly stunning to keep hearing about chips doubling their FREQUENCY every 18 months.

But that stopped about 15 years ago because CMOS transistors, which were rock-solid in terms of low power consumption (this is how they won over competing transistor technologies) became very leaky, dissipating much more power.

In the picture above we can clearly see how frequency scaling stopped, together with single thread performance, which is key for most instruction mixes and workloads. This is why people can run 5–10 year old computers nowadays fine for common tasks, with no need nor urge to upgrade. That was impossible until about 2004. This is why the magic of Moore’s law truly stopped.
To compensate for this downer, new cores were invented as basically a marketing ploy to get masses to believe their new shiny processors were continuing to be supposedly exponentially better. Well, try to examine utilization of those additional cores in modern CPUs.
Keep in mind that multi-core Intel CPUs are hyper-threaded which actually means every core has the capability to run TWO hardware threads in the best case. And each of those threads is super-scalar out-of-order, meaning they can issue FIVE or more instructions in parallel simultaneously, in the best case.
It is kind of shocking to know that in a typical sequential instruction mix in a super-scalar CPU with five execution units the average utilization is about 1.7 meaning that out of five only one additional is used with 70% efficiency! All the others are needed to just to reach the 70%, on average, on the second one.
Of course, critics will say that there are parallel tasks and programs where those additional resources can be used more effectively. And that is true, but still for a very limited set of programs and tasks. The compiler technology, even after many decades of intense research is unable to extract more parallelism from typical

instruction streams. And it is not for the lack of trying as we went trough mainframes, minicomputers, PCs, CISC, RISC, VLIW and so on.
BTW should you think that you can run all those transistors on full speed try doing it on a PC or on a server for a task optimized for it *e.g.* some video processing. You will quickly hear and see how your fans spin up and then the processor starts throttling internally to prevent itself from burning up.
One recent fresh direction is advent of alternate parallel architectures for AI and cryptocurrencies in form of specialized graphics cards used for other such tasks. Their model of computation is massively parallel, with literally thousands of parallel processing units running in very simple pipelines. Those pipelines are filled by massive AI and crypto computing tasks running rather straighforward algorithms, such as gradient descent in AI. But thermal constraints very much apply there too.
In summary, Moore’s law has been effectively dead for more than 10 years now, but that has been obscured and obfuscated by marginal parallelization. New computing paradigms such as AI and crypto offer some hope but there will be no free lunch until we find some new technology comparable to CMOS efficiency in the early days. Quantum computers are another such hope but the difficulty of their programming and just understanding what they do is truly unprecedented and off the charts.