Packaging v0.90

In a way few of us expected perhaps five years ago, packaging has become sexy and interesting! I've tried to cover some aspects of it below, to my limited ability.
To motivate things somewhat, I start at the sexy end with the M1 Ultra, then to put that in context, we retreat all the way back to the A7 and move forward from that noting each year's interesting new step.

2.5D (side by side) packaging

wafer reconstitution+wafer stitching (M1 Ultra?)

Fairly early in 2021 I knew that Apple had patents on various types of chip packaging techniques, but, like everyone, I had no idea of their relevance. With M1 Ultra the (initial) story is a little more clear, so let's examine some of these.

Think about chip packaging. You probably remember the days of 70s packaging where the silicon chip was embedded in a (much larger) black organic resin package with a few large pins on the side of the package. Wire bonding was used to solder small wires from appropriate metal pads on the actual chip to the large wires on the side of the package.

This was followed by flip-chip packaging, also known as C4, also known as BGA. All these are slightly different in details, but they're all the same sort of idea. You can see a description of the process here https://en.wikipedia.org/wiki/Flip_chip.

The flip chip transition achieves four goals:
- the package becomes smaller (much less space take up by resin, now the board area of the package is not much larger than the area of the actual silicon)
- many more pins are available (because the contacts are smaller, and we have the whole bottom face of the package available, not just the sides)
- we waste much less power (because larger contacts require more energy to transition between states)
- we can toggle smaller contacts faster than larger contacts so can support higher frequency communication through the pins.

For years flip chip was state of the art, but continue to people want to improve these four dimensions. The next step was various forms of 2.5D packaging.
Imagine that we
- remove the solder balls and the rest of the flip chip package
- replace the PCB on which chips are mounted with silicon in which we have embedded an RDL (redistri-

bution layer)
- we mount the chip directly on that silicon (called an interposer)
This gives us various improvements because we lose the volume of the solder balls, the RDL can be printed much more finely (thinner lines, smaller contacts) on the silicon interposer, and two or more chips mounted in this way can be placed even closer together.

At this point things vary out into many different technical options, but we'll follow the strand that is most relevant to Apple.
(If you want to learn about *all* the options in extreme detail:
https://www.circuitinsight.com/pdf/status_outlooks_flip_chip_technology_ipc.pdf *Status and Outlooks of Flip Chip Technology*)

The Apple strand starts with something called Fan-Out packaging. The idea here is we want to give a small chip a lot of connections to the outside world. So we mount the chip on an interposer that can fan out the very dense set of connections at the base of the chip, through the RDL, to a less dense set of connections that can be directed to the outside world via flip chip or something similar.
https://en.wikipedia.org/wiki/Fan-out_wafer-level_packaging

Now the point of interest is that the way FOWLP is implemented is somewhat remarkable. The initial silicon wafer is diced, the chips are tested, the good chips are precisely positioned on a carrier wafer, and a molding compound is flowed around them to lock them into place, creating a "fake" wafer. This fake wafer can then be treated like a real wafer in that it can be passed through the BEOL (back end of line) stages of a fab, which place successive layers of metal on the fake wafer just like a normal wafer! This was state of the art as of around 2016 (A10 shipped with FOWLP as a big deal).

So for the purposes of FOWLP the idea is to put each good chip on the carrier wafer fairly widely separated (so there's a lot of molding compound between each chip) and then build the RDL on top the fake wafer, with a via layer (like M0, dense vertical connections) sitting directly above the chip, connection to M1, M2, etc as the metal routing layers which are mostly over the molding compound, and which deliver the signals to a much less dense set of connections covering the area of both the chip and the molding compound.
In other words we have as essential ideas
- make a a fake wafer
- with widely separated chips
- use standard BEOL on top of that fake wafer
- with the end result of "fanning out" the density of connections.

At this point we are ready to understand the cleverness of (2018) https://patents.google.com/patent/US11158607B2 *Wafer reconstitution and die-stitching*!

Suppose you have all the elements of this scheme in place, but imagine a new use for it.

We still create a synthetic wafer by precisely positioning known good dies. But now we position them extremely close together, we are no longer interested in fan-out.
We again have a single reconstituted wafer on which we can use all our BEOL technology to create a set of metal routing layers.
But if the chips are very close together, we can now create routing layers that communicate information from one chip to another!
We can now achieve a few different things.
- firstly to some extent we improve yield. Rather than creating a single large chip, we can tie together some number of known-good chips, and the connection is more or less the same quality of wiring (the same standard TSMC BEOL) as on a single large chip!
Silicon interposers, EMIB, and such like are nice, but having the genuine BEOL is as good as it gets for creating dense, low-power wiring.
- secondly (I think this is true) you don't have to line up each chip in its exact "source" position when the chip was fabricated, you can stagger them by half a chip size vertically and/or horizontally. This means even starting with fairly large chips, you can now create connections between then to tie together as many as you want. As long as you get the geometry correct, and can fit your RDL as a repeating unit within the reticle limit, you can go wild and create Cerebras-level insanity!

Note also that there is nothing in this scheme that requires the different dies to be identical (as in the M1 Ultra). They could, in principle, be different chiplets from different fabs using different processes. You get the benefits of chiplets (eg use of different processes optimized for different tasks) while paying almost none of the costs (either economic, in the interposer/EMIB packaging, or technical, in higher energy/lower frequency connections between the chiplets)!

TSMC has a number of technologies that, once you understand them, seem somewhat similar to this. These have names like InFO-oS and InFO-LSI. There's some coverage of these all here: https://www.anandtech.com/show/16031/tsmcs-version-of-emib-lsi-3dfabric
To my eyes the scheme described by Apple looks closest to what's being called InFO-R.
EMIB and the InFO-LSI scheme try to provide a denser network of connections from the base silicon to the RDL. The Apple patent seems uninterested in that, suggesting instead that the required dense network can be formed in the base chips that are joined together in the fake wafer. This difference may be an issue of mix-and-match. If Apple control the entire design, they can create the necessary dense network on their base chiplets, whereas if chiplets from different companies are being used together this is probably less feasible.

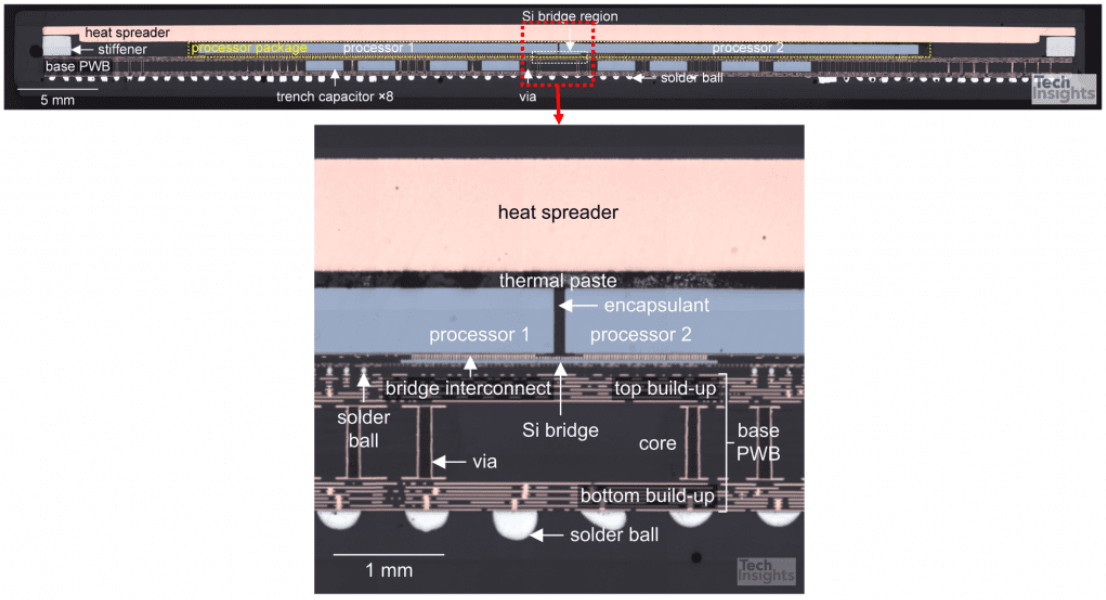
An interesting point is that if the RDL is not especially demanding, it can be fabricated in the back-end of an older fab, maybe a 65 or 90μm fab. Even if the RDL is demanding, as I just mentioned, these most demanding parts can be built in a leading edge fab, while the base chips are being constructed, with the rest of the RDL done later, at the fake wafer stage in an older fab.

The patent includes one further technical detail, which is the actual real content of the patent, all this

previous material was background.
The issue is how the space between the separate chips is filled in when the fake wafer is created. I've said that the most basic solution, when fan-out is the goal, is to fill the space with molding compound (ie an organic resin). Now, when the chips are to be tightly packed together, as I understand the patent, the current state of the art appears to be silicon oxide; so I assume something like oxygen (perhaps also with silane in it or something?) is flowed over the reconstituted wafer and at the edges between the chips to form silicon oxide which somewhat bonds them together?

This is considered sub-optimal in various ways, being less robust to subsequent high temperature processing, and having a somewhat different coefficient of thermal expansion compared to bulk silicon.
And so the patent describes ways in which silicon can be laid down (via various schemes like CVD). This silicon fill is, of course, not going to form a perfect crystalline bond with the chips, and so it will present a scattering interface for the purposes of electrical conduction; but that's fine because it matches for the purposes of mechanical strength and thermal expansion.

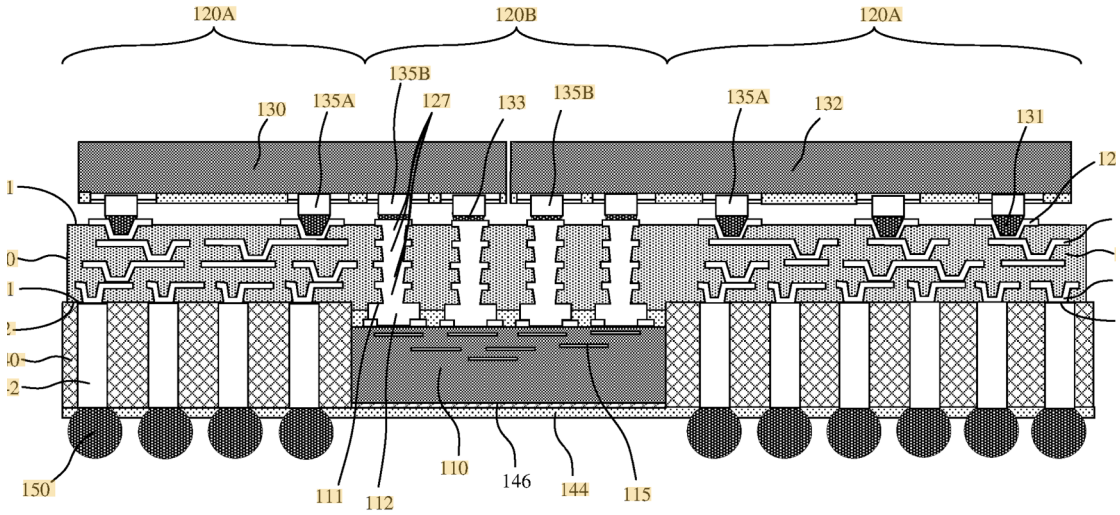
Now that's all great. Is it what the M1 Ultra is using?
Apparently not! According to
https://www.semiconductor-digest.com/apples-m1-ultra-does-use-info_lsi-or-is-it-cowos-l M1 Ultra uses an EMIB style bridge, as can be seen in the image below.



So was this all a waste of time? Not necessarily!
Apple have an apparent EMIB/Silicon Bridge patent is dated 2014, https://patents.google.com/patent/US20150364422A1 *Fan out wafer level package using silicon bridge*.
On the other hand, with these patents, you need to note the details. That 2014 patent was for a logic to memory bridge...

Unlike the other patents I've explored, the packaging patents seem to have a much more tenuous relationship to actually shipped products! So many of them look like ideas Apple was contemplating as possibly one day useful, but all grounded in "you know, we could do things this way" rather than in "let's write up what we have already implemented in the latest SoC".
So, while interesting, some skepticism is warranted (or at least an acceptance that the packaging idea may only become a product in five years or so).

However if we look more carefully, then more relevant is (2017) https://patents.google.com/patent/US10943869B2 *High density interconnection using fanout interposer chiplet*, which includes this diagram, almost identical to the above photo!



The patent with which I began this discussion is dated 2018, a year after this 2017 patent. So it seems plausible that Apple began with an EMIB-style solution (known to work) even as they investigate solutions that are likely cheaper (no bridge needed, simpler assembly) and perhaps even more performant (fewer extra levels of routing layers to get from one chip to the other)? Maybe we'll see a bridge-free design with the next version of the Ultra?
(Lest you be too cynical about these sorts of patents, which may seem like just putting lego blocks together, the real innovations are in the precise process steps used to build the final product. Ordering the steps is not trivial for a variety of reasons; for example some components may be sensitive to heat, so we have to ensure that they are only connected to the assembly once all subsequent stages operate at lower temperatures.)

If I might venture cautiously into an area I know very little about, I think the 2018 patent also clarifies the practical distinction between TSMC's InFO family and the CoWoS family.

To my eye it looks like the salient detail is that InFO begins with the synthetic wafer, on which is constructed a secondary RDL based on BEOL processing. This limits the "added layer" to RDL (metal wires) functionality.
Conversely CoWoS begins with a wafer which you construct as you like, with capacitors, logic, wiring, whatever. This forms the base layer, on top of which are placed a layer of additional chips. So, in

principle, you could form the base layer as a pure RDL metal layer, stick the known good chips on top, and having something very similar to the outcome I've described above.
On the plus side CoWoS gives you a second layer that doesn't have to be pure RDL, it can also have logic and whatever else you want.
But on the minus side I can see a few issues. One is that the mechanical stiffening that can be provided by the InFO scheme (whether by molding, by oxide, or by silicon fill) is absent, making strength and thermal expansion more of a concern.
You could try to add the filling (molding compound, oxide, or silicon) but doing that once the chips are already bonded to the underlying wafer might be problematic? So much of the processing details seem determined by what chemicals and what temperatures can be handled by the components at any given stage.
Secondly the scheme is more expensive because you're essentially constructing each package from two wafers and two full passes through a fab, whereas the InFo scheme requires only one wafer and the second pass uses only the BEOL of a (possibly trailing edge) fab.

earlier version (die stitching, but no wafer reconstitution)

This 2018 patent builds on (2017) https://patents.google.com/patent/US20180294230A1 *Systems and methods for interconnecting dies*. This patent includes the basic idea of using the BEOL to create connections between two die on a wafer, but what's missing is the idea of a synthetic wafer.
So you create a wafer of dies, pairs side by side, then create the BEOL connections between pairs, then scribe the die pairs and test. If a pair fails, you can split them in half and hopefully one half will work and you can still sell that as an M1 Max rather than as an Ultra pair.
The 2017 scheme is less flexible (good for pairs, but will scale sub-optimally to quads or larger), but obviously also a little easier (doesn't require a reconstituted wafer).
In a way you can think of this 2017 scheme as being a variant of the same idea as Cerebras. The 2018 version is then an optimization, based on the fact that we can form synthetic wafers at an alignment that is good enough for our needs (this may not [or may?] also be true for Cerebras).

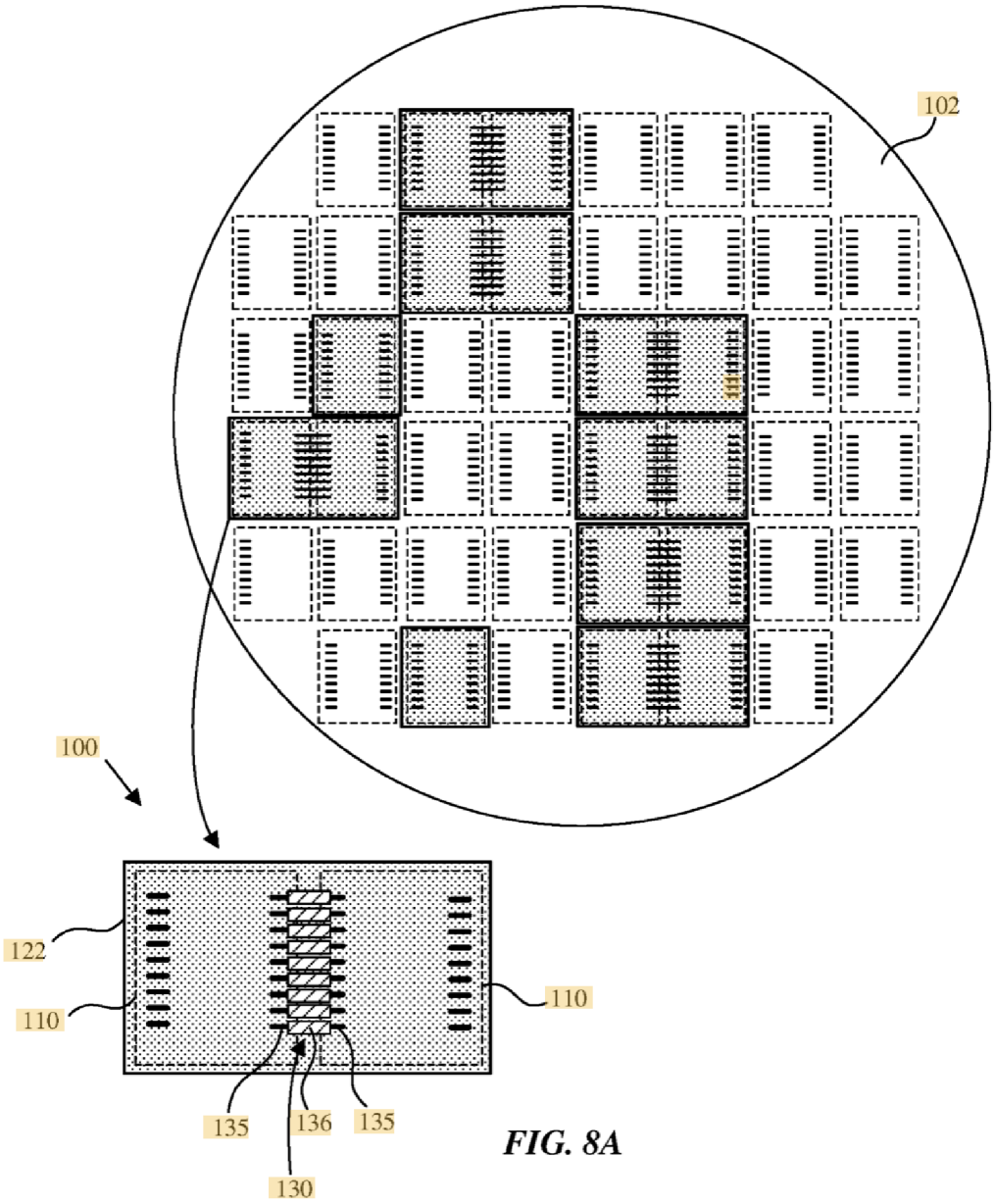
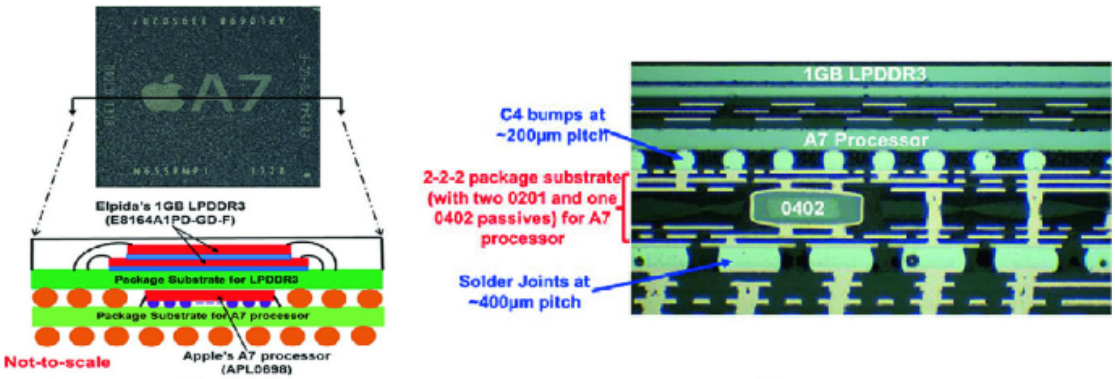


FIG. 8A

History of iPhone packaging

Let’s try to understand earlier packaging.

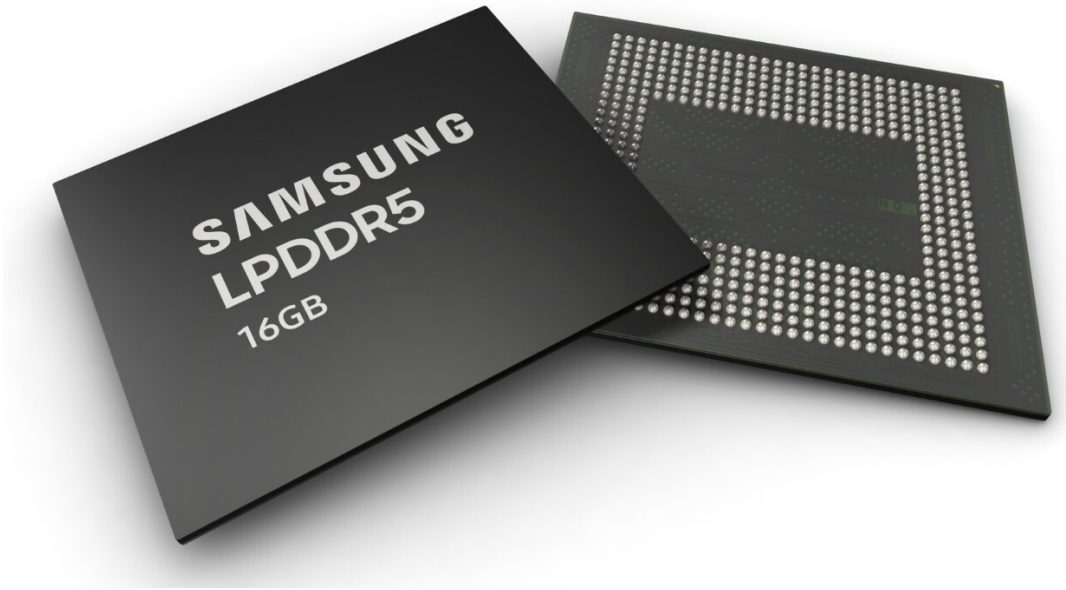
The images available are not great (and I think some start to become misleading/mistaken as Apple becomes less traditional) but let’s do what we can.



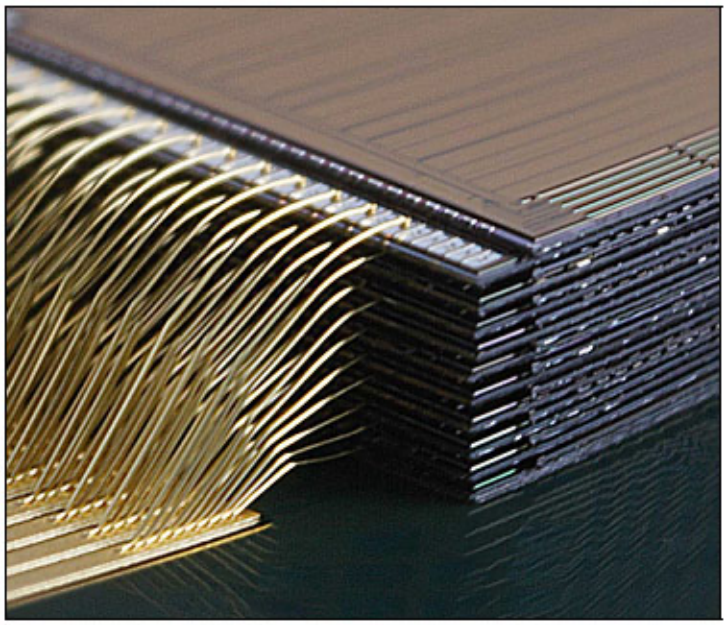
(a) (b)

This is traditional (for phones) DRAM wirebonding. The two DRAM chips are wired bonded to a large'ish substrate, the A7 SoC is mounted via bumps onto a second substrate, and the two communicate via large bumps that connect upper with lower substrate.

First thing to note above is that the A7 is built out of two separate units.
The top, memory, module consists of the DRAM wirebonded to the upper substrate, which has pins around the sides to get the signal to the SoC, as in



This upper memory module comes from a company like Hynix or Samsung or Elpida, and is not Apple's problem. The actual manufacturing is astonishing: here's a stack from Elpida:



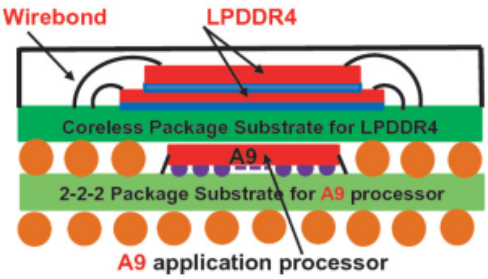
It's hard to believe that his wire-bonding not only can be performed successfully and robustly, but is cheaper than what would seem simpler alternatives like using a laser to create TSVs through the stack of DRAM chips!
Note the distinctive alternating stacking pattern of the DRAM chips, called *cross-stacking*. We will occasionally see it in later cross-section photos.
To be honest, the above 20 die stack is not (yet?) relevant to us. Phones still seem to use 2-die stacks, and the M1 shots I have seen use 4-die stacks. (But this may reflect people destroying the cheapest devices for teardowns! Perhaps larger memory capacities use 8-die stacks?)

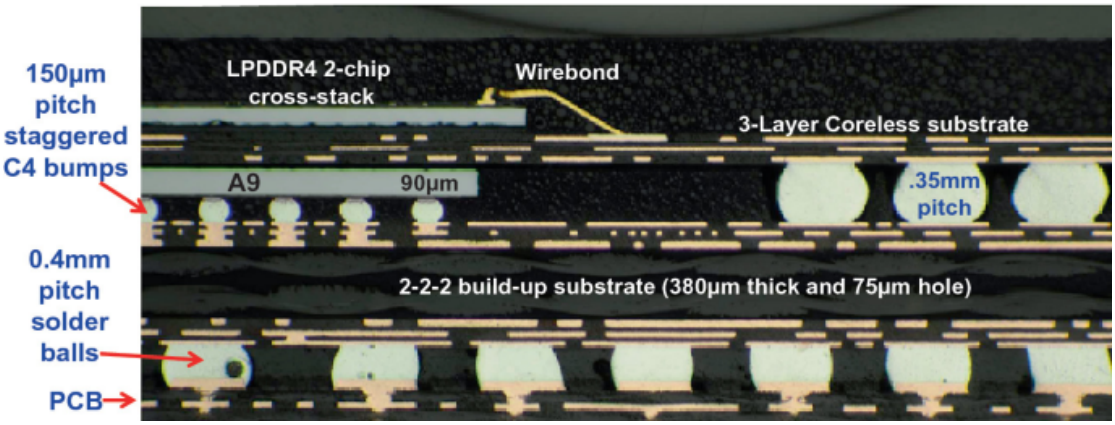
So the DRAM sits on a substrate whose job is to route signals from the termination points of each little wire to the appropriate bump on the BGA (Ball Grid Array) of the final package.

The second point of interest is that the A7 SoC itself is mounted on a second substrate, whose job once again is to spread a dense set of signals from the A7 BGA out various bumps, one set of bumps connecting to the upper memory module, the second set of bumps connecting to the lower PCB and thence to the rest of the phone.

An interesting (non-obvious) third point is that the size (x×y) of the final package is ultimately determined by the size of DRAM package! And the SoC must be able to fit into the empty middle area away from the outer BGA of the memory module.

The A9 image below makes things a little more clear:





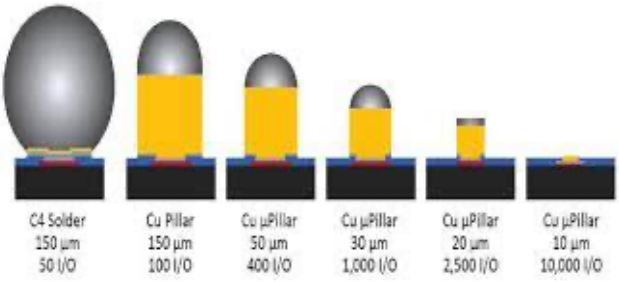
You should be able to easily identify
- the upper module (DRAM, wirebonded to routing substrate)
- the lower module (A9 SoC sitting on a multi-layer substrate, with one type of solder ball connecting to the memory module, and a second type of solder ball connecting to the PCB)
- it's not obvious but I assume at a different area of the lower substrate, there must be some vertical connections. Of course the primary (highest speed) connections are required in the upper layer, from SoC to DRAM; whereas the lower layer to the PCB supplies power and ground, but also various IO from radios to flash.

Note that the finest solder ball connectors (from the SoC to the lower substrate) have shrunk from 200µm pitch to 150µm pitch.

So the above give us the "old" state of the art.
What's new is when TSMC introduces Integrated FanOut (InFO). FanOut is easy enough to understand: a chip is small but may need many pins, so you need some sort of layer (carrier plus some wiring) to distribute the very dense pinout at the bottom of the chip (as it comes off the assembly line) to a less dense collection of bumps that will appear below the chip package.
This is part of a world where what I buy from a third party is something encapsulated in some sort of molding compound or ceramic, already fairly robust, and with fairly large, not especially dense, solder balls as pins. What if I change those assumptions?
The first change InFO makes is to assume that I have access to bare dies. On a carrier wafer, I create an RDL (the same pattern of wiring that connects the dense pin out below the chip to a less dense package

pinout), then put the bare die on top, then I add molding compound. By changing the order of operations I can get the RDL "closer" to the chip, and it can be a lot thinner (more like the wiring layers of a chip than like a thin PCB).

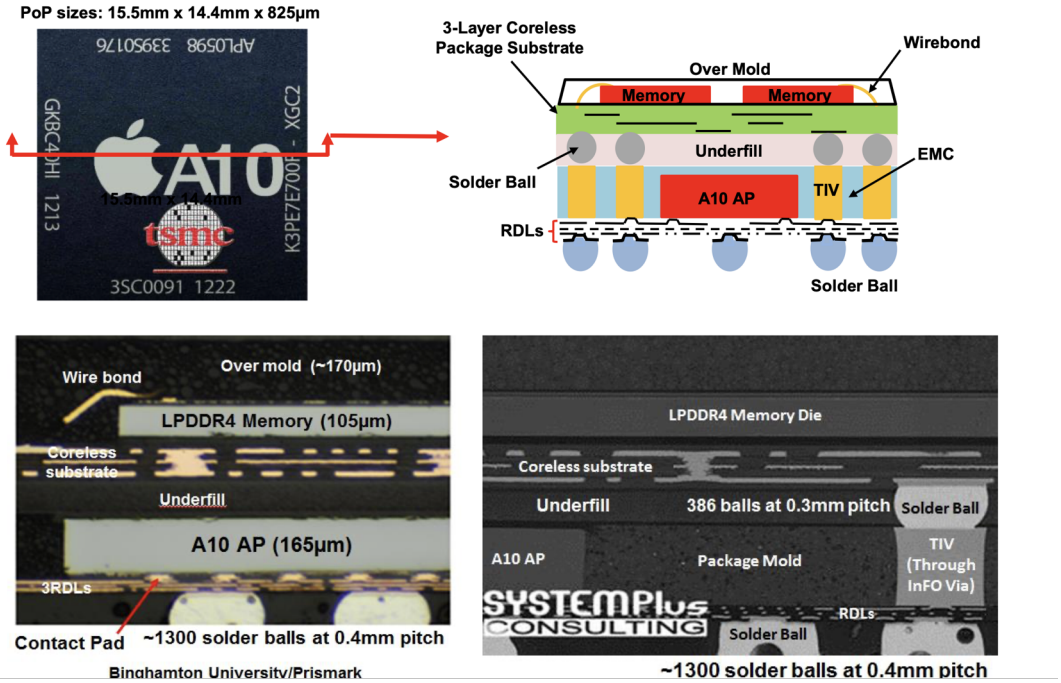
The next change InFO allows for is that I have the option for more dense connections. The old style connections are the solder bumps you've seen above. These can be shrunk smaller, but there are limits. The alternative is that I use copper pillars. Imagine that my chip and or my RDL expose minute copper surfaces as the pinouts. As long as I can get the alignment correct, these can be tiny, and if they are clean enough, simply placing them against each other and squeezing (perhaps with a thermal assist) will bond them.



We can use these either to mate the chip with the RDL or to mate the new package (chip plus RDL) with other devices. One possibility is to use a small EMIB-style chiplet to bridge two such RDLs; another possibility is to simply aligned the two devices of interest against each other and create a new RDL on top of the prior RDL's, connecting to the chiplets. Ultimately what is opened up is a packaging world that's more like chip manufacturing than old-style PCB manufacturing.

So look at how this evolves going forward.

What's new with the A10 in 2016? InFO!



Note that we, along with the basic InFo layer, we have added “Underfill” and “EMC”.
I think these are essentially about this patent:
(2012) https://patents.google.com/patent/US8963311B2 *PoP structure with electrically insulating material between packages*.
The issue is easily explained: compare

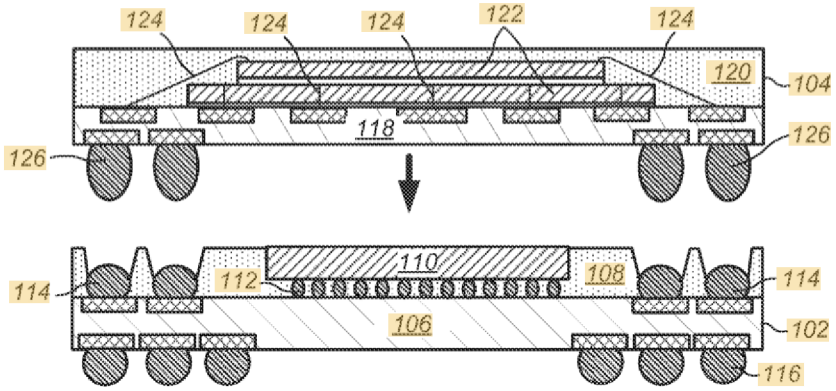


FIG. 1B
Prior Art

to

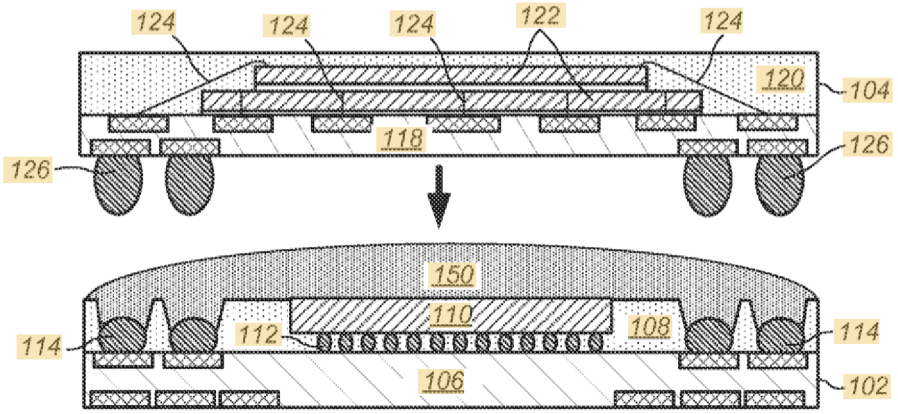


FIG. 2B

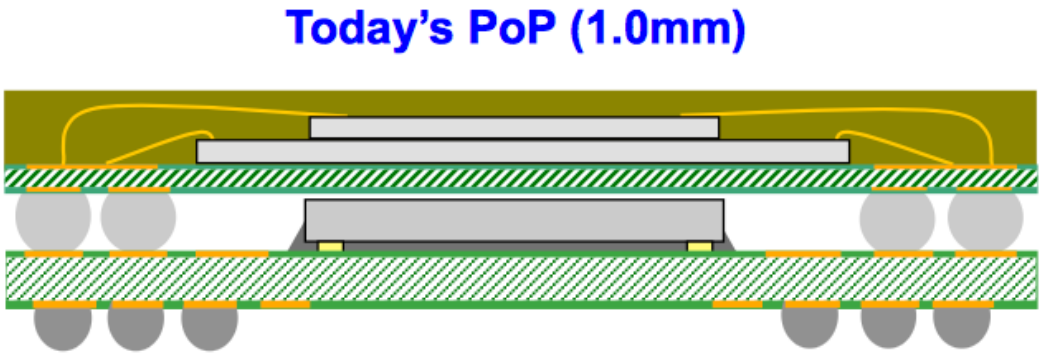
The prior art shows manufacturing say an A7. You have the memory module on to, DRAM wire-bonded to the substrate. You align it against the SoC module below, squeeze and heat, and the aligned solder bumps melt and fuse.
Sounds good; the problem is that the relevant modules are very thin and on cooling tend to warp/twist (think of changing from say a flat rectangle to something saddle-shaped, like a pringles chip). In the process the solder bump joints tend to crack.

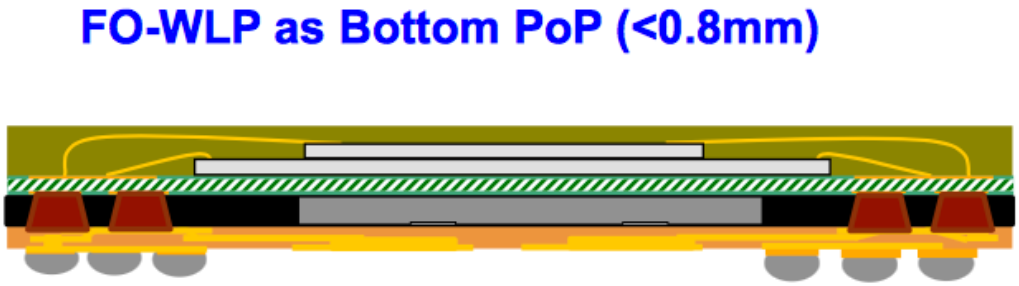
The patent scheme deals with that via a simple fix: smear the lower module with some sort of non-conductive material that’s somewhat thermosetting. Now once they cool down, the non-conducting resin will hold the stress of each module trying to twist, rather than the solder joints having to hold that

stress, and we get much less solder cracking.

This seems to be what we're doing here. We add non-insulating "structural" filler to both the memory module (where it is called Underfill) and the SoC module (where it is called EMC – Embedded Molding Compound), along with holes drilled through the EMC, the so-called TIV's, to electrically connect the routing layer (below the EMC) to the memory module. Note how much thinner the routing layer now is. We have removed the structural core that was giving it strength and warp-resistance (it has becomes "coreless"), and we have moved that job to the EMC.
Among other things this allows us to shave a small amount of height from the package.

This step, of replacing a PCB-like substrate with a thin RDL seems to be what is meant by InFO (Integrated FanOut), as in this diagram. (The terms InFO and FO-WLP are slightly different, but not in ways we care about right now.)





Not only does this give us a slightly thinner package, it is in fact cheaper (though less obvious) to manufacture. The final manufacturing path was probably a collaboration between Apple and TSMC, but (2013) https://patents.google.com/patent/US9305853B2 *Ultra fine pitch PoP coreless package* describes one possible way to do it. (One important aspect is that you build it "upside down" and flip it at the last stage; another aspect is that you build it on a carrier wafer to provide strength until the final stage.)

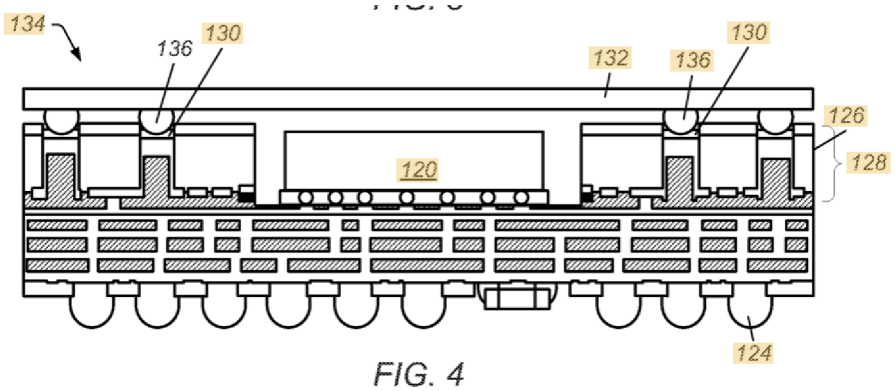
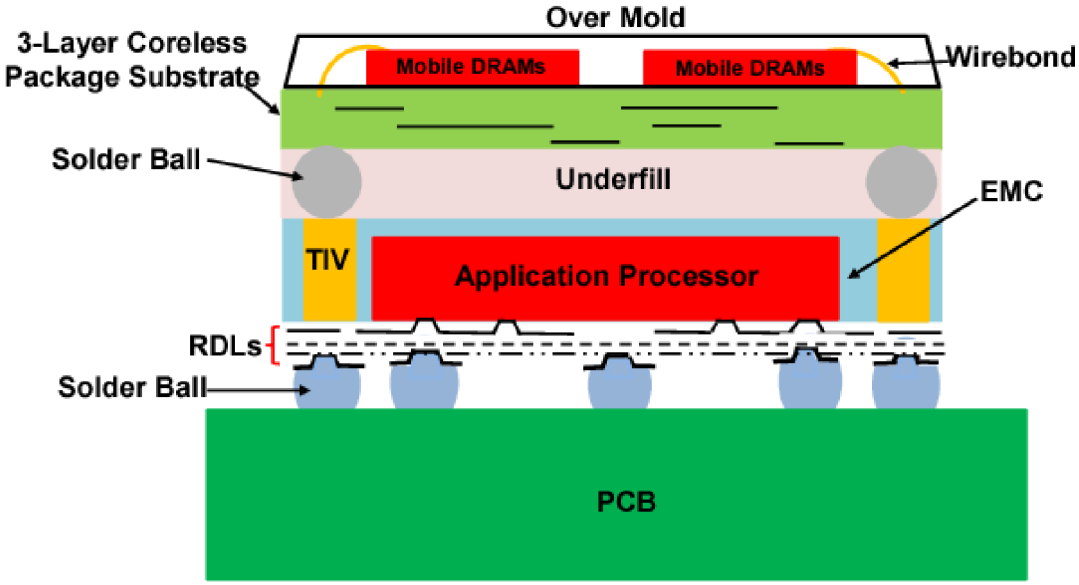


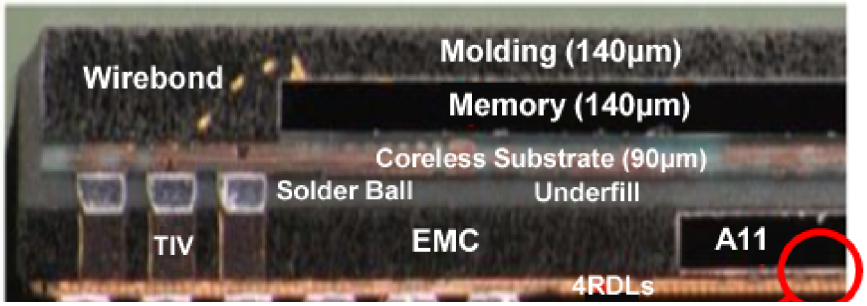
FIG. 4

The patent suggests a few different ways to assemble the final package, but the A10 looks like it matches the version above.
132 is the DRAM package, 120 is the SoC, the layer 128 is the layer of EMC, while below SoC 120 is the RDL eventually spreading out to large solder balls that can mate to the PCB.
Once again we see a delay. The patent is date 2013, the A10 ships in 2016.

The DRAM solder balls pitch has now shrunk, from 400 to 300µm, and by eye it looks like the pitch of the bumps from the SoC to the RDL has shrunk from 150 to 100µm.
Strangely the A10 SoC is thicker than the A9 SoC. This may correspond to packing more passive components (capacitors and inductors directly below the silicon chip of the SoC.

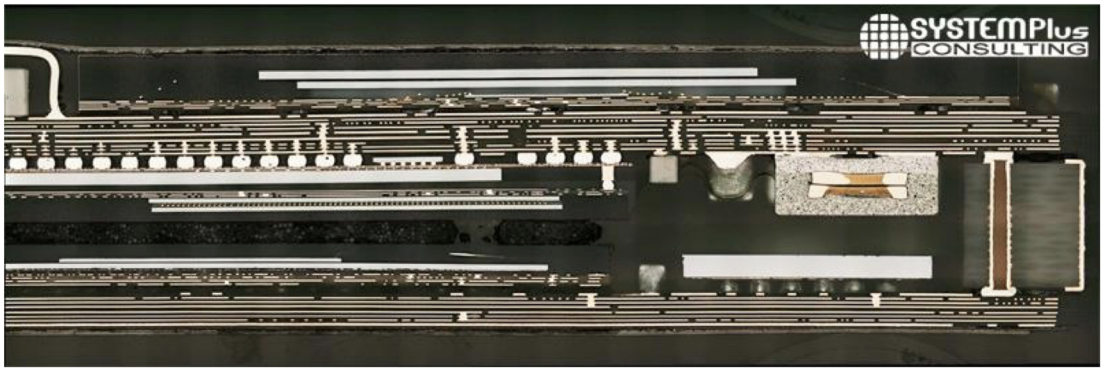
What's new with the A11?





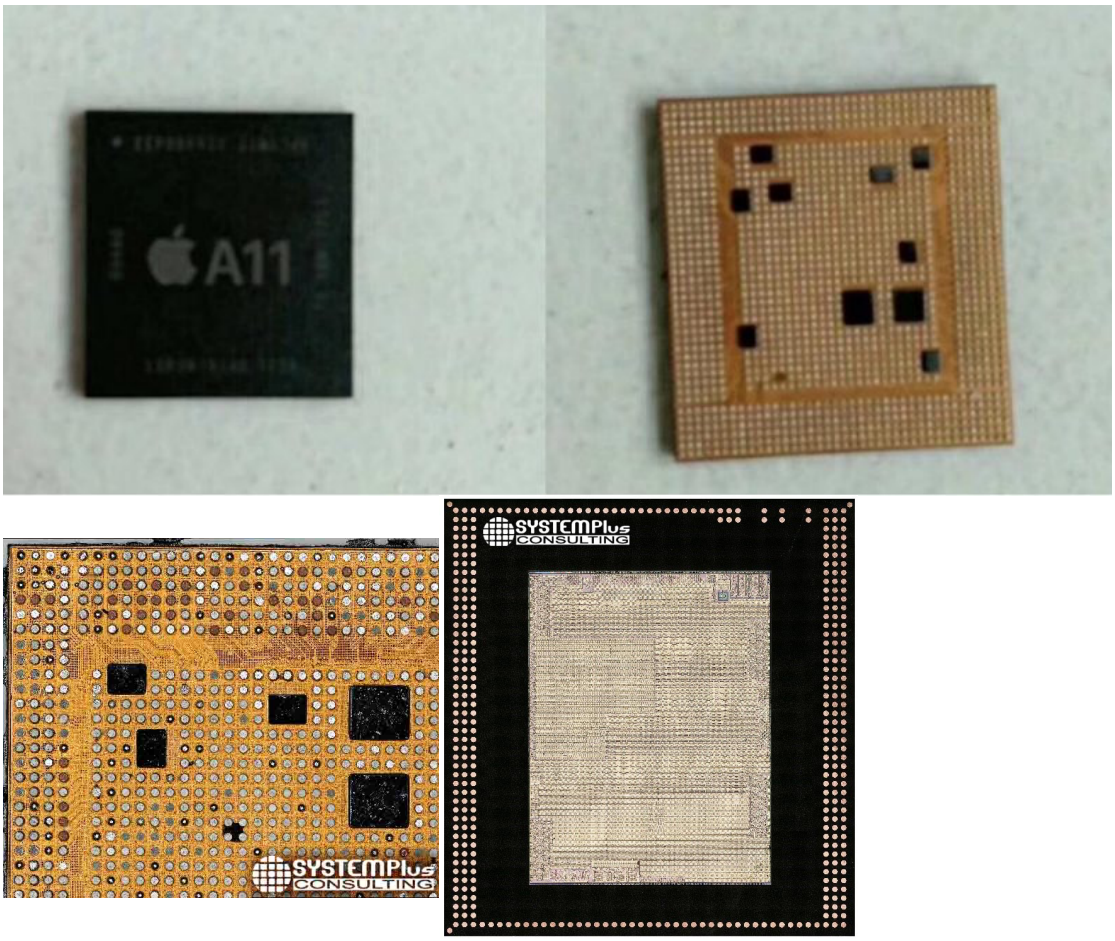
That all seems fine and understandable.
The primary new item is the addition of a large (relatively) capacitor mounted right under the SoC (a so-called Land-Side capacitor). We see more of this going forward. In a later section we will discuss this evolution.

However there is an alternative image available that is rather more difficult to interpret: the version below makes it tough to see precise details. I'd recommend you look at the original, at https://s3.i-micronews.com/uploads/2018/02/Yole_SP18373-Apple-A11-inFO-Packaging_Flyer.pdf .



I *think* that this image actually barely shows the A11. As best I can tell, it is a cross-section through the iPhone X motherboard.
The points of interest include
- there are two PCB layers to the motherboard
- the upper PCB has the A11 mounted on it, the reverse side of that PCB opposite the A11 has various RF chips (see https://www.ifixit.com/Teardown/iPhone+X+Teardown/98975)
- the lower side of the PCB has the flash in that area.Flash is wire-bonded, and you can see what looks like a wire-bond trace, somewhat like the DRAM wire-bonding.
I think the cross section is such that it doesn't even cut through the actual A11 chip.

Here's a different image of the BGA side of the A11, removed from the PCB.
We start with the first two images (what you would see if you were given an A11 at the factory). The black items are components involved in voltage regulation and stabilization.
The third image shows what you get if you (very carefully!) pry an A11 off an iPhone X logic board; and the fourth one shows, I believe, what we get if we remove the RDL from the package. On doing that we are left with various connectors from the DRAM (around the outside) and the dense set of connectors of the actual SoC that would be connected via InFO to the (now-removed) RDL.



You can imagine that if you cut through the black area of this fourth image, you would cut the DRAM chips, wire-bond, and the DRAM substrate; but you would miss both the large solder balls and the dense connections in the central area, so you would not (as in the PCB image above) actually see any of the A11 in such a cross-section.

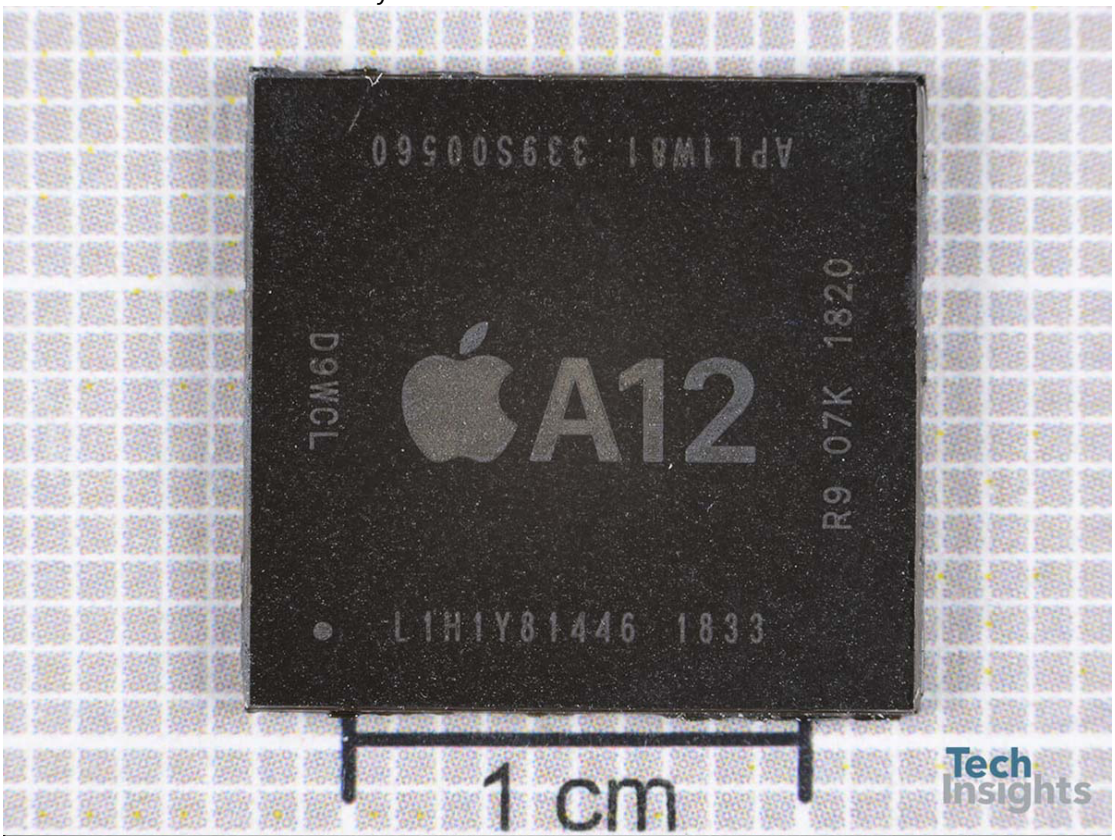
So let's take stock.
As the system has advanced we've seen
- (A10) the replacement of the SoC module substrate with an RDL (better electrical properties, thinner, easier to manufacture).
- (A11) an additional layer of RDL, and land-side capacitors (capacitors mounted right below the RDL, as close as possible to the SoC).

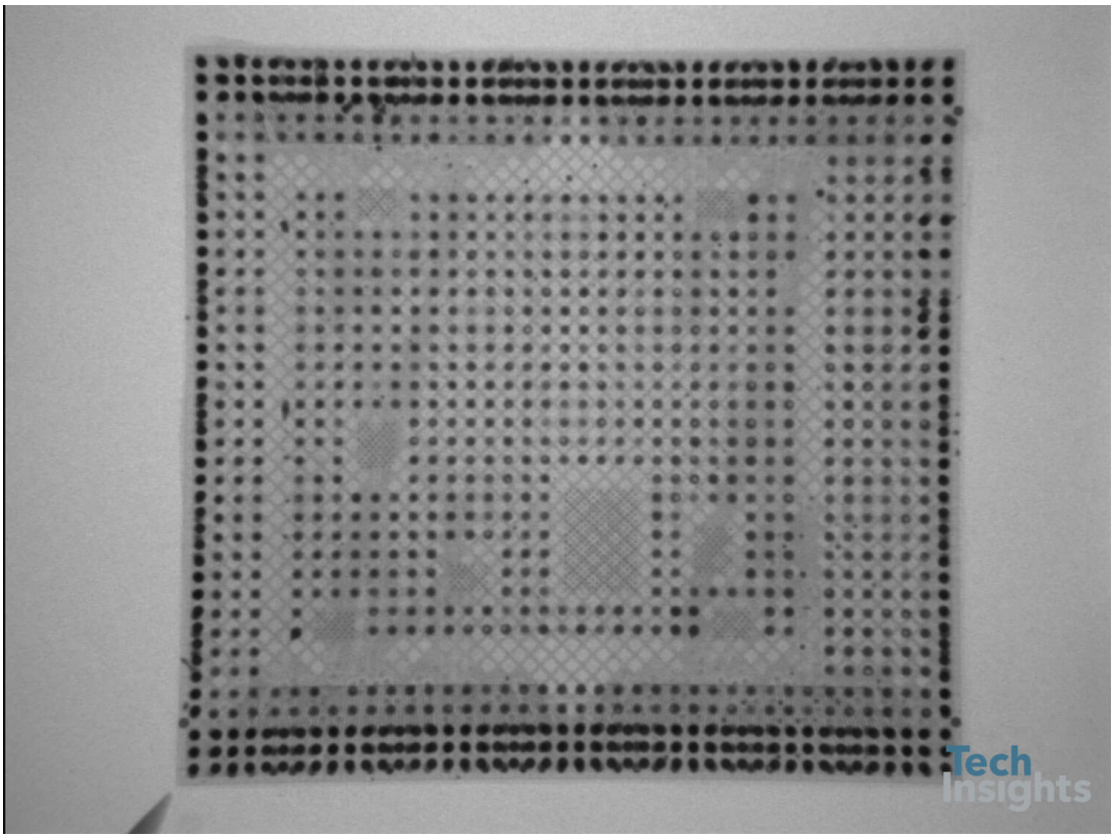
There's always some energy efficiency to having the capacitors as close to the SoC as possible, but they gain more importance as you reduce the overall functional voltage of your SoC (it goes down a little with every improved process) and as you increase frequency (these both mean you need a cleaner power supply for the SoC). As the designs evolve you see a constant attempt to increase capacitance, from large capacitors on the PCB to medium-sized capacitors mounted on the back of the SoC package, to small capacitors placed inside the molding compound to make use of that space:
https://www.realworldtech.com/power-delivery/4/
At the extreme, there are even in-silicon capacitors placed within every line connecting an external

power/ground pin to the SoC itself; all in an attempt to dampen noise in the power supply as much as possible.

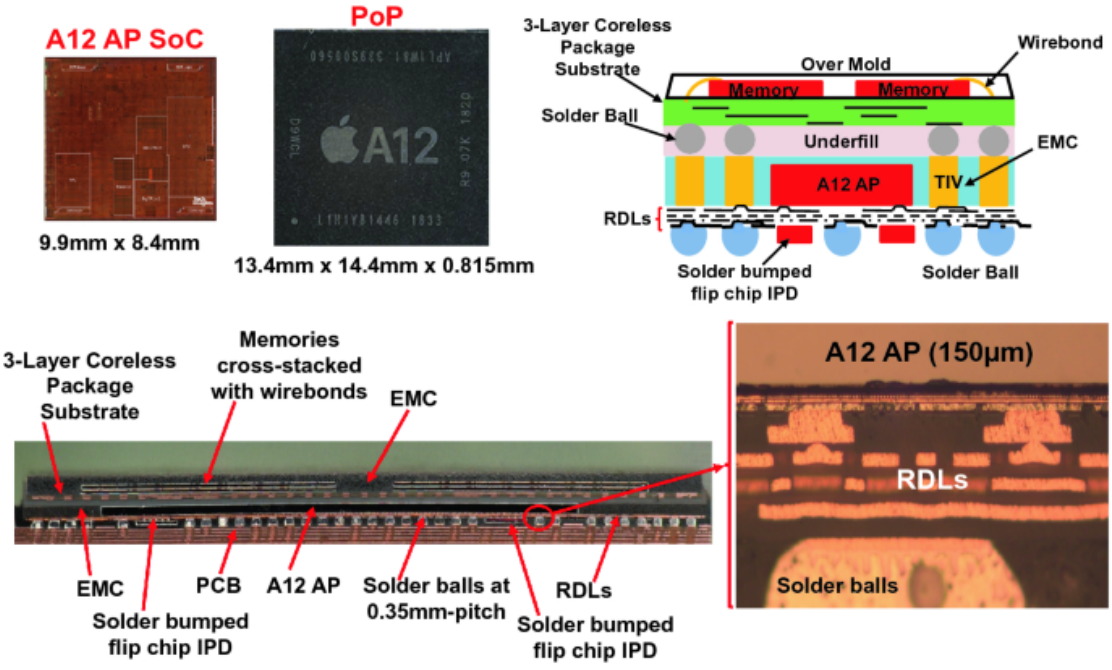
Let’s move on to the A12.

From the outside it looks mostly the same:





The structure is similar (perhaps with more IPDs [Integrated Passive Devices] mounted below the RDL. On thing that starts to become striking in the images below is how much thinner the actual SoC (plus RDL) is compared to the DRAM package. I assume Apple are constantly thinking about ways to change this situation, but this would presumably require intense collaboration with a DRAM vendor.



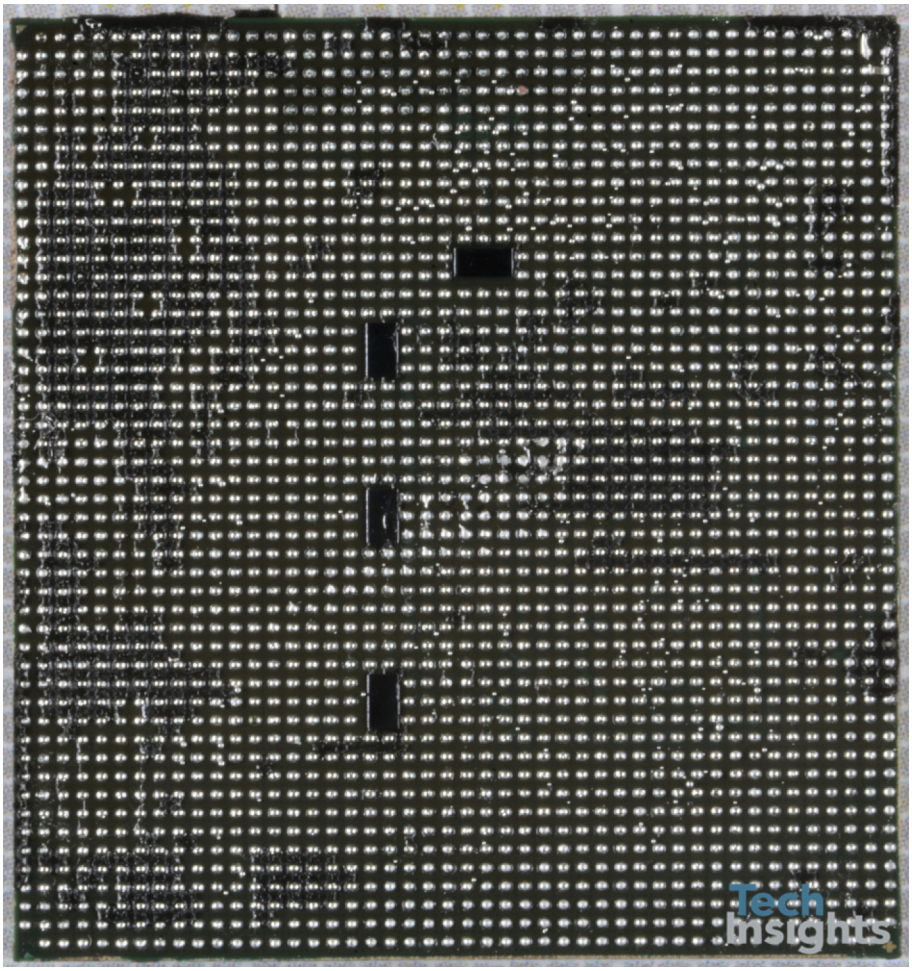
The interesting change at this point is the A12X: https://electroiq.com/the-packaging-of-apples-a12x-is-weird/#

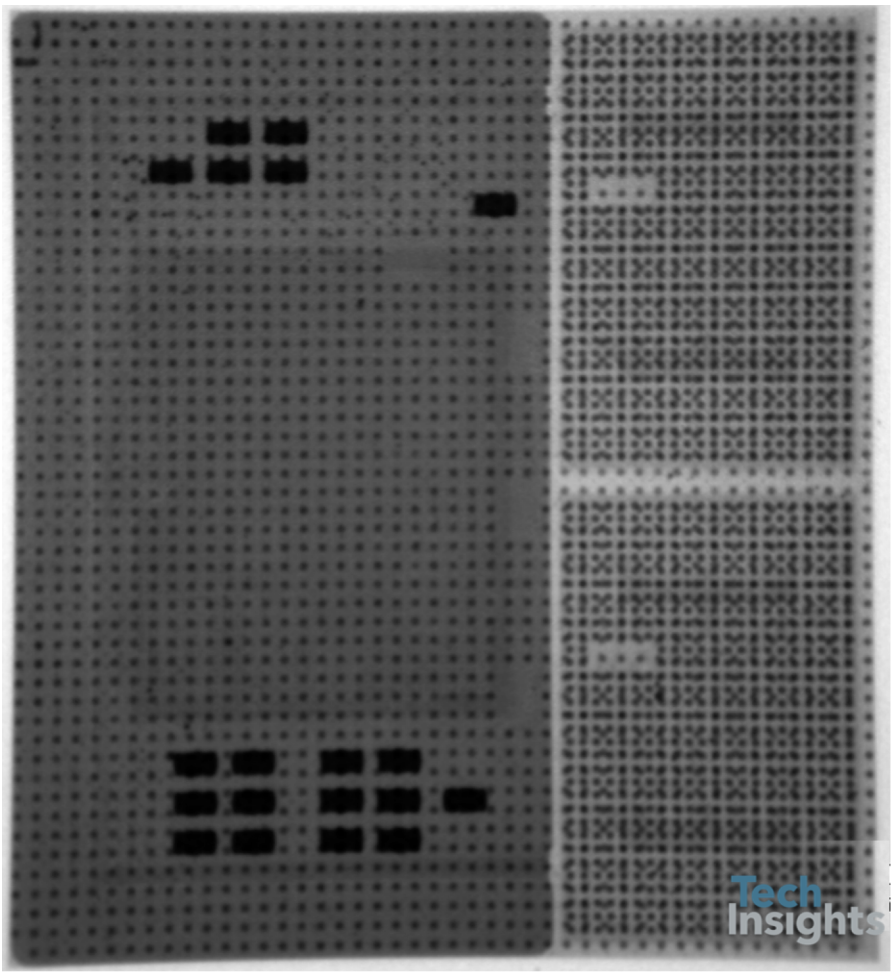
Earlier we had iPads mounting their DRAM on the PCB as in:

With the A12X we move the DRAM onto the A12X package. (Note the capacitors everywhere, on the above PCB, and in the various packages above and below.)

The DRAMs are presumably the usual commercial structure (cross-stacked, wire-bonded, on an internal substrate).

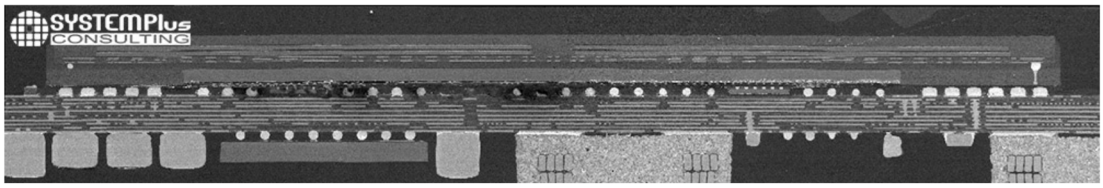
The flip side is as expected, along with an unsurprising X-ray image:


Tech
Insights

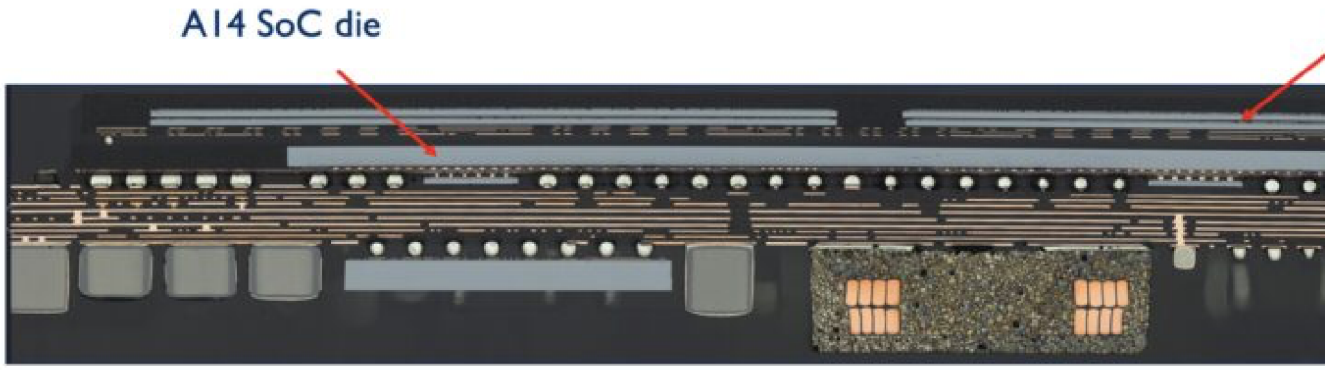


The general belief is that this is manufactured with the A12X SoC placed on a substrate which extends “beyond” the SoC, then the DRAM chips are subsequently mounted onto that same SoC substrate. Note that DRAM doesn’t run at exceptionally high speed, or use exceptionally fine bumps (everything follows specs which change slowly), and this in turn means that the mounting of the DRAM on the SoC substrate requires nothing fancy. No silicon bridges, no micro-bumps or pillars. But we do get the DRAM very close to the CPU (lower energy to move bits), while also seeing slightly better thermals (CPU heat doesn’t rise upward through the DRAM). The cost is about twice the area compared to, say an A12. Of interest in the X-ray image, we see the black rectangles as various passives (inductors or capacitors) but also four slightly lighter grey rectangles around the edge of the dark grey SoC. these are likely capacitors manufactured into the substrate.

Nothing interesting seems to be available about the A13. We have one A14 photo which looks like more of what we’ve seen multiple times, only neater.
https://www.i-micronews.com/products/apples-a14-bionic-system-on-chip/

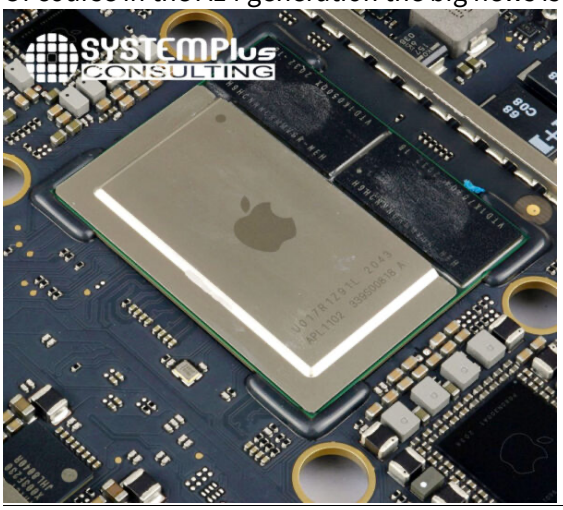


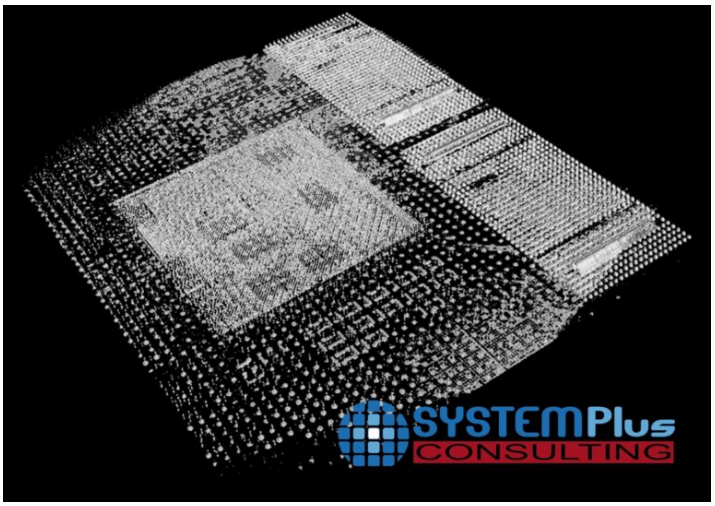
A different version is here:



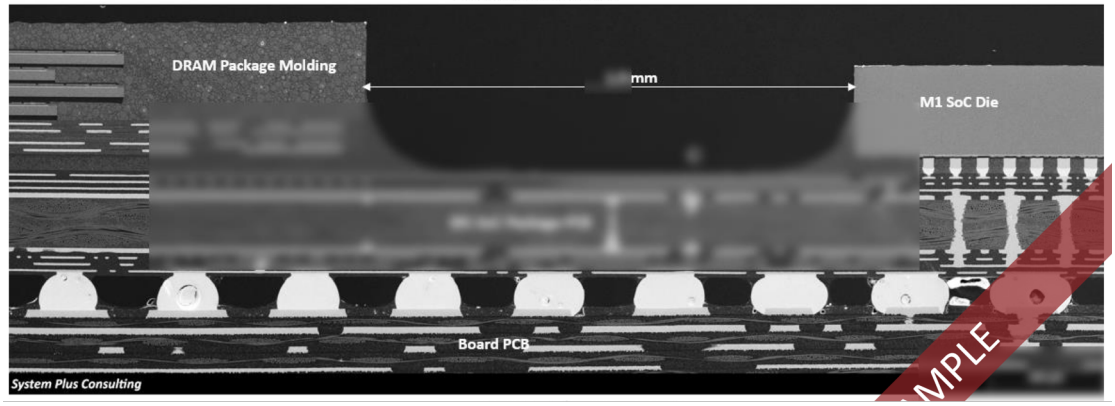
By now we can recognize the top block is the entire A14. We see two side-by-side layers of DRAM (two-stacked) on the DRAM substrate, with the A14 SoC and its ridiculously thin RDL below it, two land-side capacitors hanging off the bottom of the RDL, and the entire A14 package sitting on a PCB with various items on the other side of the PCB. All familiar.

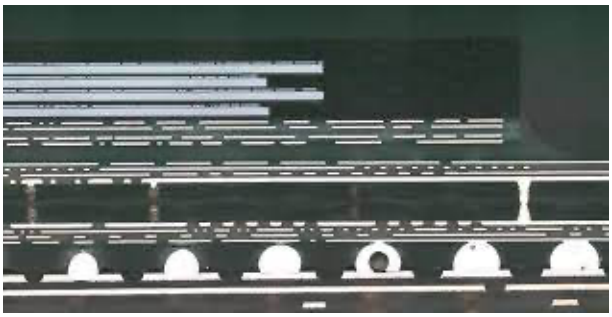
Of course in the A14 generation the big news is now the M1. Looks much A12X:

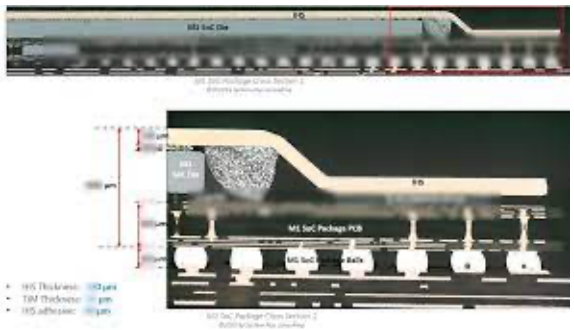




M1 SoC Package Cross Section







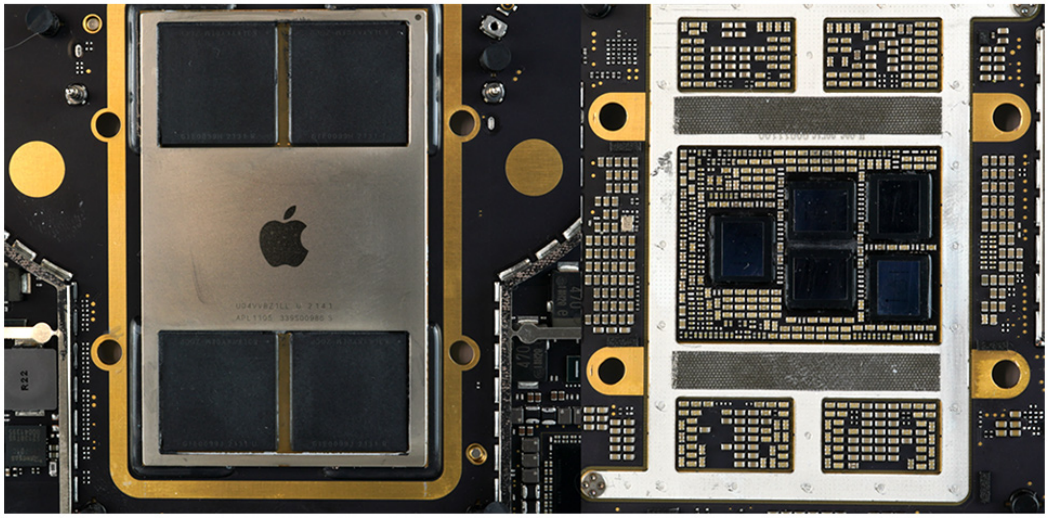
The original images are available at https://s3.i-micronews.com/uploads/2020/12/SP20608-Apple-M1-System-on-Chip-Sample-1.pdf , but they are not making it easy for us to see what's going on without paying for the report!

You can see in the third image above the metal case (probably both heat transport and some EM shielding) around the actual M1 SoC, the common substrate for both the SoC and the memory; and the thing that looks like a stone, which is, I think, a stiffener ring around the SoC to prevent warping the package warping. (Remember that along with all the other problems of these packages, an on-going problem is differential thermal expansion. The metals, molding compound, and silicon, all want to expand in slightly different ways as they get hotter, and you want something to try to limit the amount of warp that this generates.)

On to M1 Pro/Max:

Apple M1 Max SoC – Full Package Views - Frontside & Backside - Optical view

(Source: Apple M1 Max System-on-Chip, 2022)



SYSTEMPlus CONSULTING ©2022 | www.systemplus.fr - reverse-costing.com

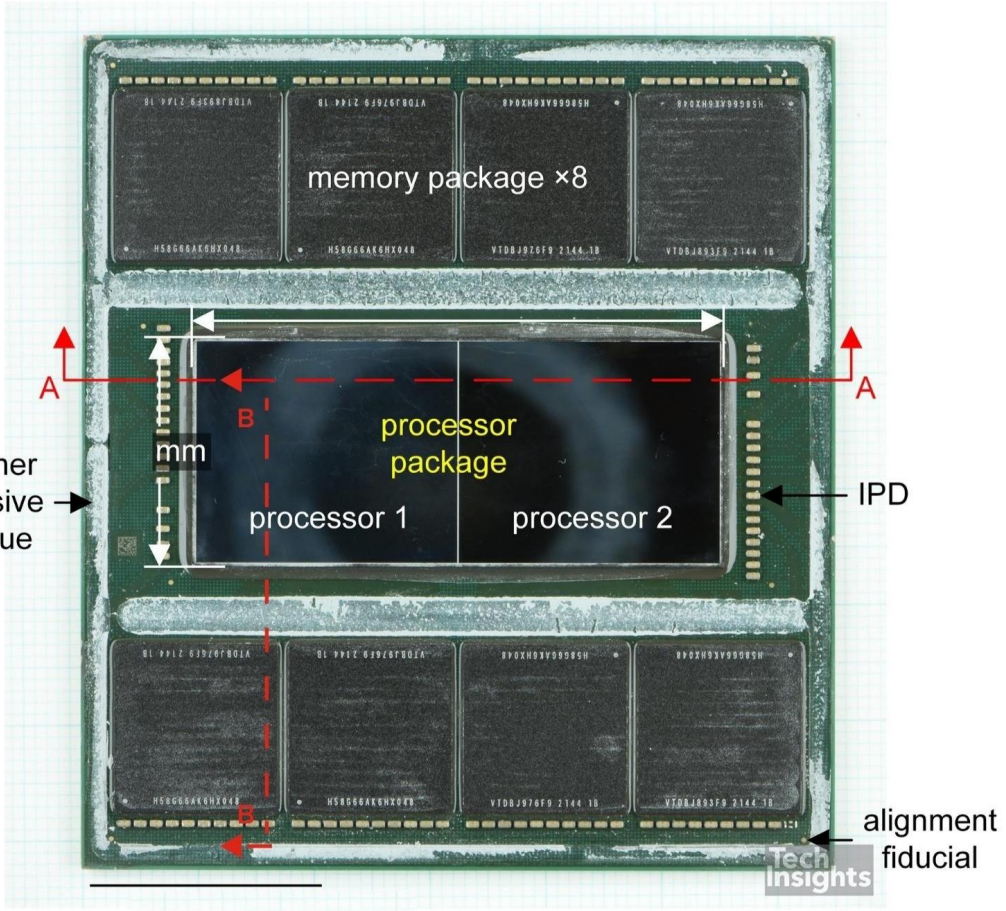
The cross section is is for the Pro, but Max is similar. We see that they are in fact a lot more different from the M1 than you might have expected. The differences are, I assume, primarily to handle substantially more power.
We can see that the packages are larger and a lot more robust. The metal surrounding the package looks twice as thick (as thick as the chip, rather than half as thick for the M1, so a better heat spreader. Meanwhile below the basic RDL connecting the DRAMs to the SoC, we have a more robust substrate on which are mounted all these capacitors and CIVRs (Coupled Inductor Voltage Regulators), five for the M1 Max, three for the Pro. Having so many of these allows for extremely precise voltage management, having them so close to the CPU (and controlled by HW) allows for very rapid DVFS changes. These are somewhat equivalent to Intel's FIVR, which appeared Haswell/Broadwell, then disappeared with Skylake.

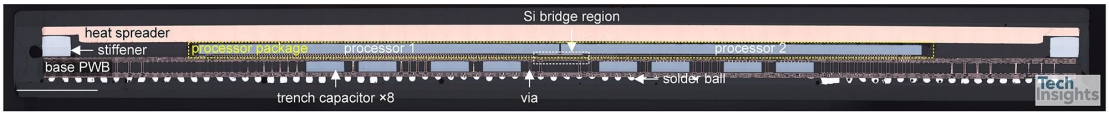
If you can't get enough of this sort of packaging/power detail, you may want to look at the various comments in this Twitter thread, https://twitter.com/marcan42/status/1557233825838936064 which explain why this stuff is unusual (in the context of the M1 Pro/Max) and how it saves energy (both by allowing for lower voltage margin, and allowing for very rapid DVFS). The comments also give some background to much of the SRAM details I gave in volume 2, with incredulity that Apple uses different SRAM voltage levels relative to logic levels because "that's such an additional hassle": https://twitter.-com/ignaloidas/status/1557255185898037249

If you keep going, the thread branches off into an interesting discussion of just where Apple uses various small ARM cores (mostly Chinook),
https://twitter.com/azonenberg/status/1557237018328870912
(Short answer is M1 appears to have at least 30 such cores, including in non-obvious places like on each NAND-chip, to provide a nice abstraction layer to the rest of the software.)

Almost done! Only one interesting device left! On to M1 Ultra. Pictures from https://www.techinsights.-com/blog/apple-m1-ultra-advanced-packaging

It’s unclear how to parse this but let’s pop the lid off:

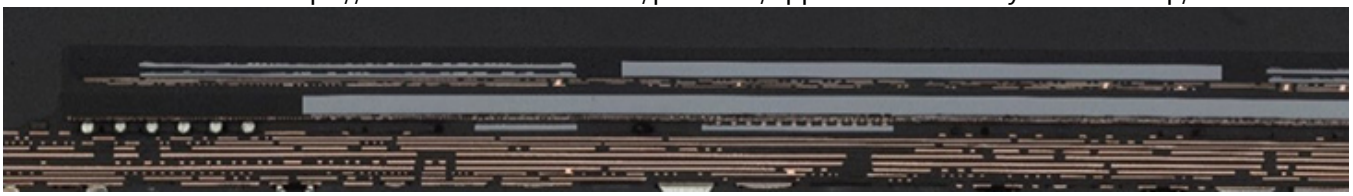




The primary new and unclear thing we learn is that the communication from one of the two SoCs to the other occurs via a silicon bridge (like EMIB, though apparently with substantially more dense wiring). TSMC's version of this tech is called InFO-L.

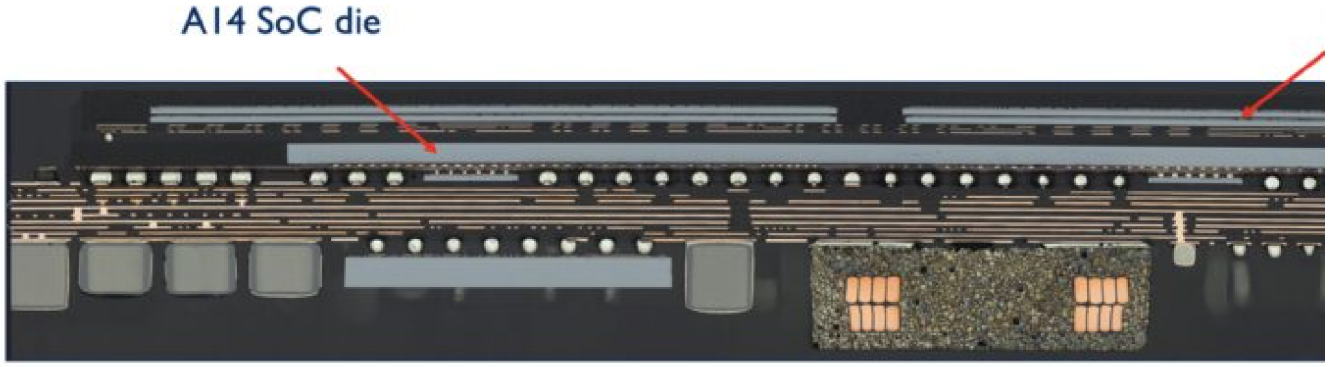
It appears that M2 has nothing interesting enough for the various companies that have provided the photos above to even bother releasing teaser web pages and photos. I guess the next big step is considered to be what will Apple do for the Mac Pro replacement.

There's barely anything about A15, but that doesn't mean there isn't a huge mystery there! Check out the cross section from https://www.i-micronews.com/products/apples-a15-bionic-system-on-chip/



OK, not great quality, sure, but enough that we can identify (compare with the A14 below), two stacks

of two DRAM chips, and the SoC (with associated capacitors hanging below). But what is that rectangle *between* the two DRAM stacks???



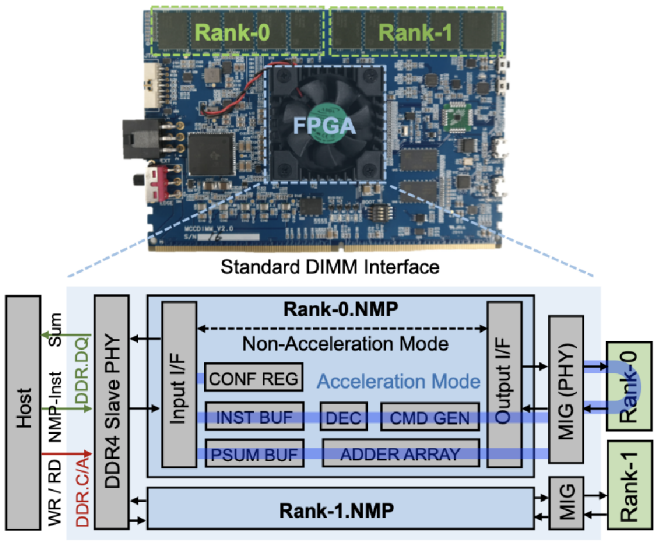
One can imagine various things (put more memory cache there?) but none of them really make sense in terms of economics or engineering. Except one option: PiM...

PiM stands for Processing in Memory, and it is based on the obvious point that moving data from RAM to a CPU is low bandwidth and energy-expensive. Suppose we could perform (simple...) computations in memory. The very simplest version of these are things like either move a block of memory from one address to another, or flood fill a block of memory; but more sophisticated (but still feasible) might be adding two large vectors, or performing large reductions.

If this is a PiM solution, then presumably Apple contracted with a DRAM vendor to have them place a custom Apple chip on the DRAM substrate. That chip looks large (about a quarter of the area of the A15, assuming it's squarish), but it could be made in a much cheaper process than N5, since it only has to run fast enough to keep pace with DRAM).

We've mentioned PiM earlier along with the summary paper (2022) https://arxiv.org/pdf/2012.03112.pdf *A Modern Primer on Processing in Memory*. Most of the examples discussed either utilize stacked DRAM (like HBM) or modify the DRAM and so look nothing like the A15. But there is one example, Samsung's AxDIMM, that looks very similar... Imagine this same idea but using Standard PoP DRAM interface and form factor, and replacing the massive hot FPGA with a dedicated low-power ASIC.

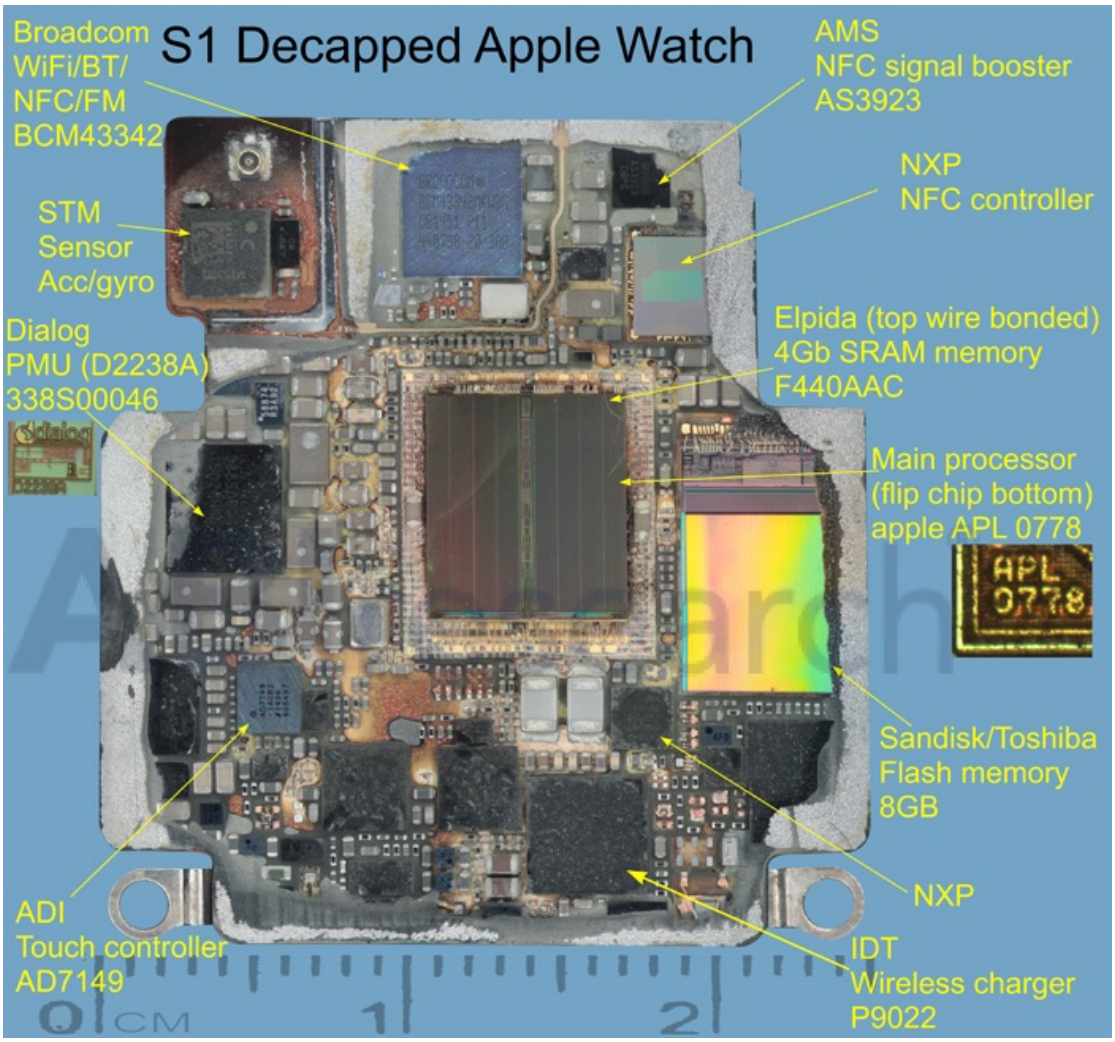
AxDIMM Design: Hardware Architecture



Ke et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM", IEEE Micro (2021)

Apple may make this visible via API (like the Vectorize APIs that route to AMX) but they may not, and may use it purely for internal purposes (ie AI, the current catch-phrase for all high-bandwidth computation!) Or it may be Apple-internal for a few years as they figure out where it does and doesn't work well, and they redesign the chip and API a few times, till it eventually becomes public?
Note patent (2020) https://patents.google.com/patent/US20220156045A1 *Performing Multiple Bit Computation and Convolution in Memory*! This is written vaguely enough that it could correspond to an SRAM, but could also be relevant to a DRAM...

Beginning around 2013 there are multiple Apple patents discussing various ways to connect memory to logic. Strange ways to stack DRAM chips, connecting DRAM to logic via a silicon bridge, DRAM mounted on both sides of a substrate (in all the examples we have seen DRAM only on one side), use of wideIO or HBM, triple-stacked PoP packages (DRAM on top of HBM on top of SoC) etc. Many of these patents are, in fact, abandoned (ie after starting the patent process Apple didn't bother to pay the fees for the next step) and none seem to be in products. So a waste of time?
To me it looks like a few different things going on.
One is that Apple was preparing various Plan B's for if DRAM bandwidth and/or capacity did not grow as expected, but bandwidth and capacity did remain on track.
A second is that the story does not end with iPhone/iPad/Mac; there is also, for example, the watch.
There's extremely little available on how watches are packaged, but we do have this image:

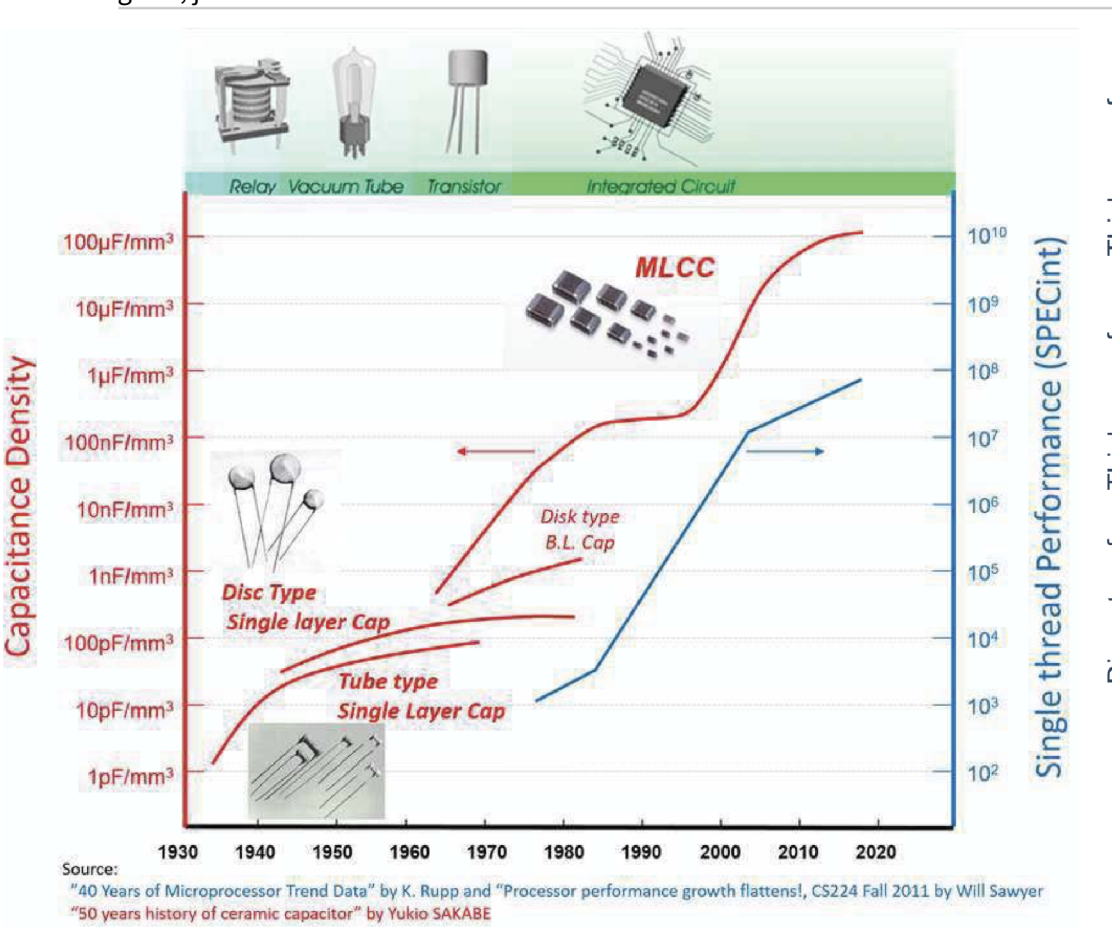


It looks, as best one can tell, like at this stage (and remember this is the first watch, when Apple was, like the first few iPhones, using a lot of third party chips rather than aggregating as much as possible on a dedicated SoC) we're mounting the DRAM on the processor much like with iPhones, only, to save space, with the DRAM as a "bare die" without the surrounding molding compound.
But once we appreciate that the watch is in the mix, we can also appreciate that there could be a desire to somehow mount a DRAM chip as "directly" as possible on the SoC without requiring either wire bonding or a substrate between the DRAM and the SoC. In this light, some of the memory ideas make more sense. Unfortunately, however, it's impossible (from what's publicly available, anyway) to track the progression of DRAM and the SoC in watches.

Likewise we've seen that, so far, both the iPad and Mac have been OK with "standard" DRAM mounted in a fairly standard fashion; close to the SoC, on the same package, but not using unusual mounting techniques. It's possible that once the Mac transition is over, Apple will consider the energy (and volume) costs of continuing to use standard DRAMs to be unacceptable, and some of these wilder ideas (bare DRAMs connecting to the logic via EMIB-style bridges, or RDLs) will come into play, perhaps once the volumes look large enough to justify the alternate tooling required?

Packaging passive components

It's an interesting fact of which few people are aware that capacitors have followed their own exponential scaling law, just like Moore:



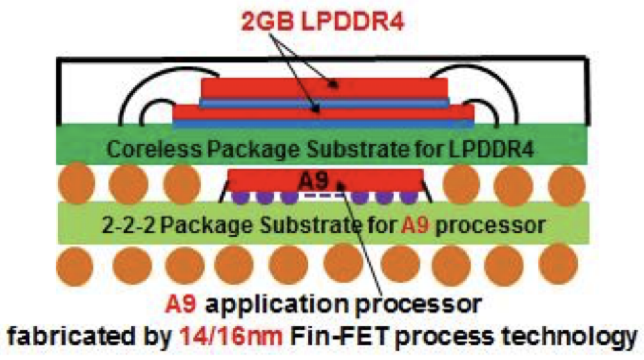
This history has made it worth providing capacitors at every level of packaging (since a useful amount of capcitance can be provided at every level).
Some aspects of this are discussed in http://pwrsocevents.com/wp-content/uploads/2020-presentations/S3_2_Shunsuke-Abe_36894.pdf .

capacitors integrated into the packaging

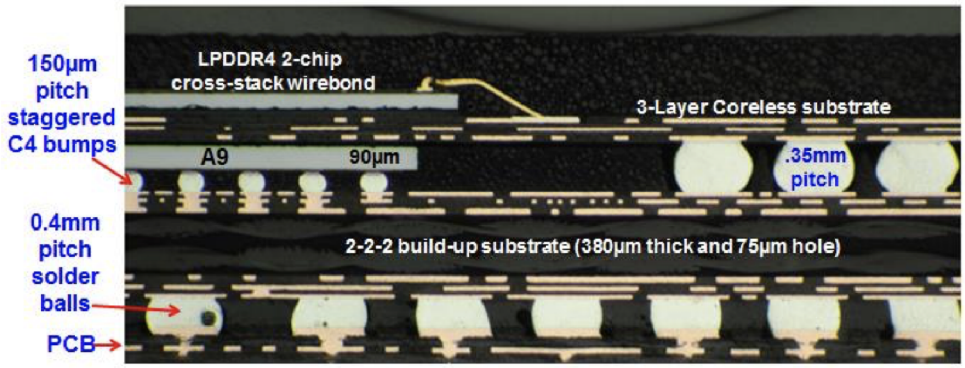
For example consider (2015) https://patents.google.com/patent/US9935076B1 *Structure and method for fabricating a computing system with an integrated voltage regulator module*. I think the most significant aspect of this is Fig 6 (and a bunch of related figures).

The approximate progression of interest is we start off with DRAM mounted on a substrate (which distributes signals), processor mounted on a substrate (which distributes signals),
and the whole lot mounted on a PCB. The name of the game, then, is trying to make these two sub-

strate layers as thin as possible (to shrink overall height). We see, for example, that by the A9 we have managed to make the DRAM substrate coreless, so it's an RDL (routing distribution layer, but very thin.

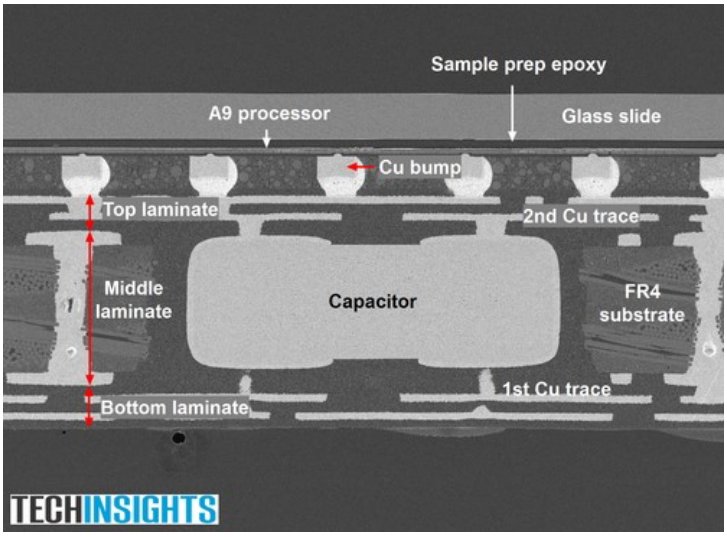


Compare that layer to the thicker layer on which the processor is sitting, in the photo below.

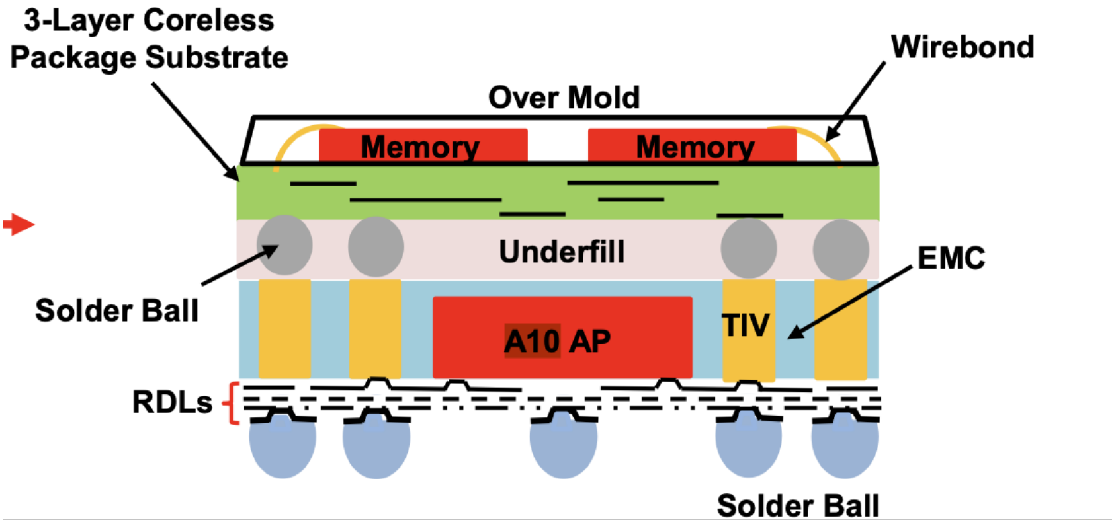


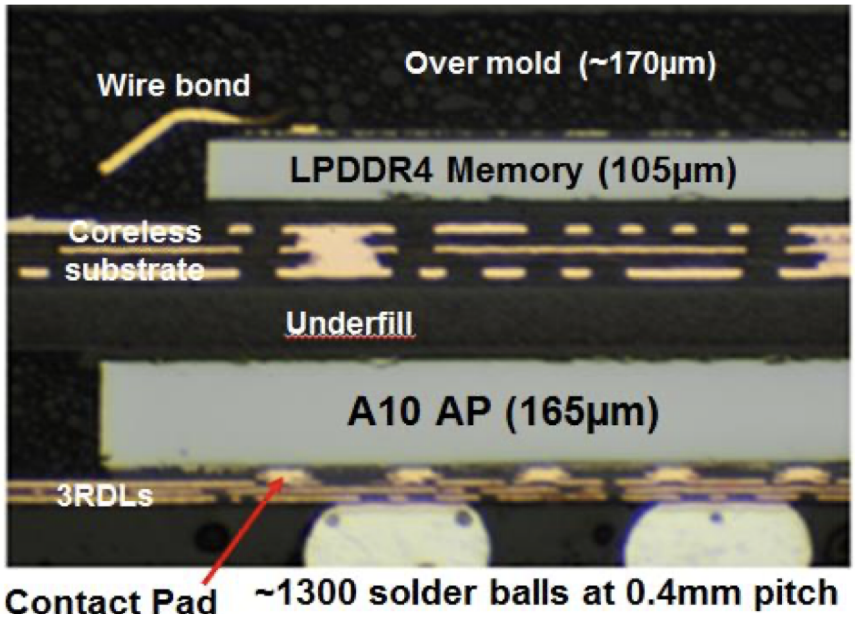
The idea of the 2016 patent is, if we have to have the substrate layer anyway (eg to make the overall package stiff enough) can we get some additional use out of it? And of course the answer is yes. The idea is to embed large passive components (like inductors or capacitors) in the substrate itself, rather than mounting them on top.

This was in fact done for the A9 as we see below, maybe also for earlier chips?



But by the A10, using TSMC's InFO, we have also removed the lower substrate (the strengthening is now provided by molding compound around the A10 itself) as in below:



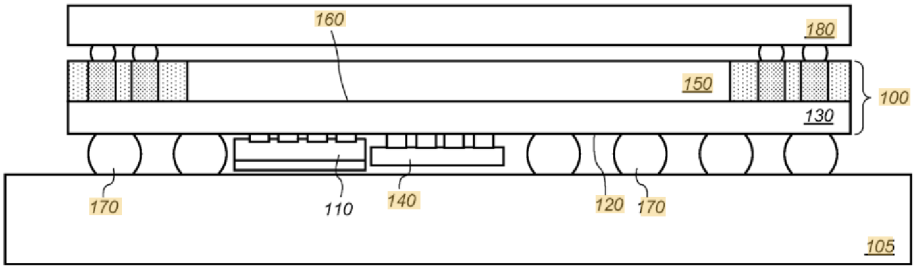


Two solutions (one background knowledge, one the new feature of the patent) are presented in (2015) https://patents.google.com/patent/US9691701B2 *SOC with integrated voltage regulator using preformed MIM capacitor wafer*, which seems to address the obvious question of how do we solve the A9's problem when we no longer have a substrate to hide the capacitor?

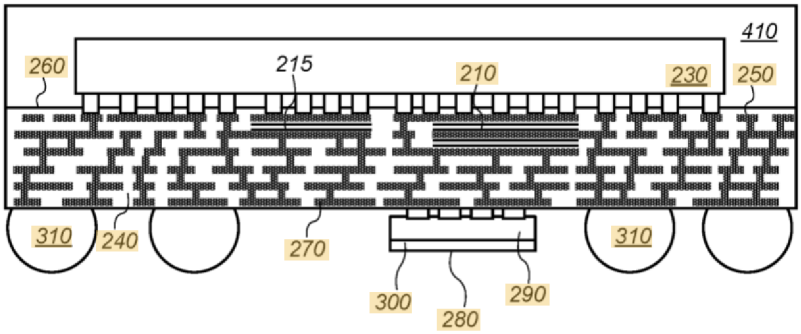
The answer is we now create the capacitor in the RDL layer as an MIM capacitor (a large block of metal, on top of which is a thin layer of insulator, then another large block of metal)! That's presumably what's being used by the A10 and later.

We do know that around this time TSMC introduced a similar sort of technology, called HD-MiM, in the context of silicon interposers.

You can see the idea here. The image below shows the background state of the art as of the A10: we try to pack an Integrated Voltage Regulator (110) and a Capacitor (140) directly under the SoC package (as seen, eg, as the black rectangles on the back of the A11).



But we can also use any unused area of the fanout wiring layer directly below below the actual SoC chip to build MIM (metal insulator metal) capacitors, as in region 210 below:



Another possible solution is to move some capacitors up into the EMC (Epoxy Molding Compound) surrounding the A10. The idea here is that since we're already punching holes into that EMC for the TIVs (Through InFO Vias) that communicate power and signals from the RDL up to the DRAM layer, we can make more such holes, coat them with the appropriate chemicals, and voila, instant cylindrically shaped capacitors! These are called DTCs (Deep Trench Capacitors) and are another TSMC technology introduced at this time and which we know were definitely used in the A10 package.

Both solutions can be used together, to provide larger capacitance in a smaller area.
Remember that the point of capacitors in this context is to provide a source/sink of electrons during rapid changes in current, faster than the battery can track. One advantage of the MIM capacitors, even if they do not store as much as the Deep Trench Capacitors, is that they can be placed in the RDL directly under (and thus very close) particular problematic spots like the GPU and CPU, which may modify their current draw extremely rapidly as they constantly power on for a few cycles then power-off.

Interestingly, Apple has a patent (2014) https://patents.google.com/patent/US9607680B2
EDRAM/DRAM fabricated capacitors for use in on-chip PMUS and as decoupling capacitors in an integrated EDRAM/DRAM and PMU system on a variant of the DTC idea where you create the Deep Trench Capacitor on the SoC itself, so that it can be maximally close to whatever block/sub-block it is helping out.
This patent assumes a process that can create eDRAM on the SoC, since eDRAM uses similar deep trench capacitors. TSMC has provided eDRAM for some process versions, for example some game consoles, and IBM has made aggressive use of eDRAM for

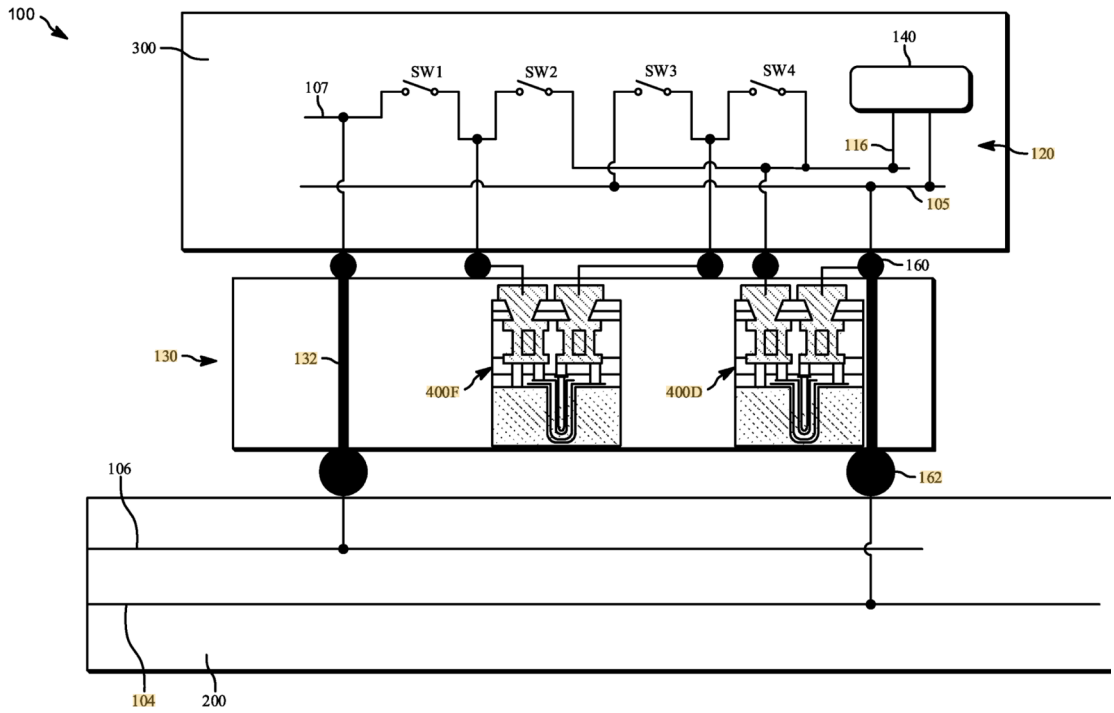
years.

So perhaps one day, when all the immediate excitement of the x86 to ARM transition is over and there's time to explore new ideas, eDRAM may be used for a future chip's SLC or something similar? (eDRAM is both more dense and has lower leakage current than SRAM, so it has advantages for such a cache; you can read about the tradeoffs here in (2013) https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.298.8449&rep=rep1&type=pdf *Technology Comparison for Large Last-Level Caches (L3Cs): Low-Leakage SRAM, Low Write-Energy STT-RAM, and Refresh-Optimized eDRAM.*

Another question you may have is why these RDLs keep getting thicker and more complicated. What are they doing apart from (mainly) connecting power, ground, and signal pins from the SoC to the external pins? We see one part of the answer in (2018) https://patents.google.com/patent/US10756622B2 *Power management system switched capacitor voltage regulator with integrated passive device*. This describes how we embed not just a few large capacitors, but a large number of small capacitors throughout the RDL, connected to logic on the SoC. This logic then dynamically ties together different combinations of these capacitors

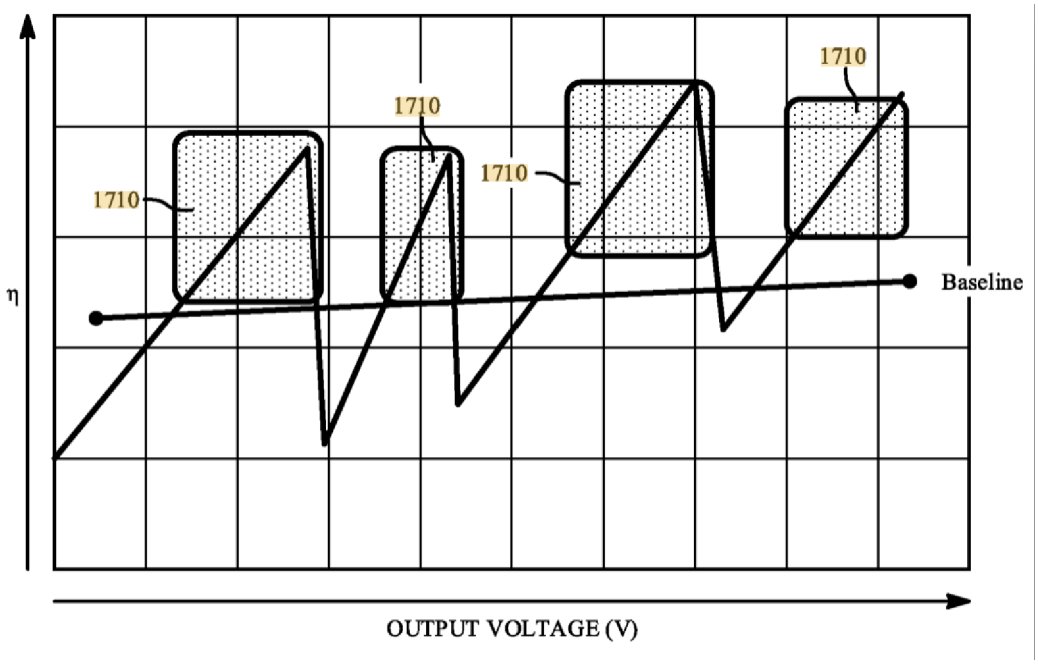
- to provide independent power supplies for different islands. (At an obvious granularity these would be things like CPU clusters, or the GPU; but one might prefer a finer granularity, so individual power supplies for each CPU [even though they are at the same voltage] or even individual power supplies for things like the L1 caches vs the CPU core.)
- to provide different, optimal, reactance at different DVFS points.

This viewpoints gets us to things like



where we now want the RDL/substrate to provide us with both a dense web of capacitors (and some inductors) along with a network (controlled by logic in the SoC) that allows us to modify exactly which of these passive devices is, at any given time, connected into the power delivery network.

The end result of all this is:



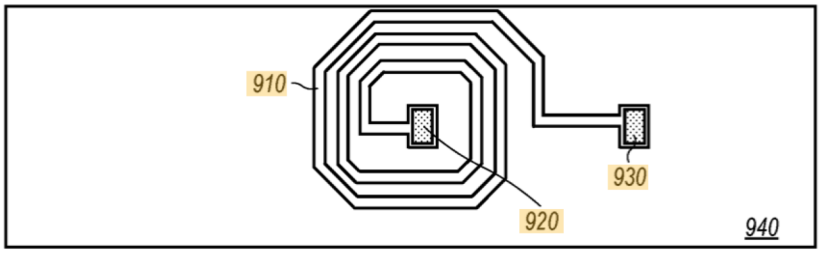
The Baseline curve gives energy efficiency for each output voltage for a traditional voltage regulation setup. We see that it is smooth and essentially flat.

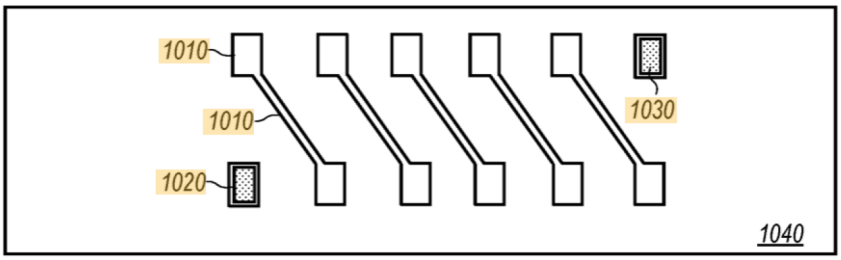
The Apple setup corresponds to the zigzag curve, with different efficiency depending on which collections of capacitors we connect. The grey zones correspond to voltages of particular interest (ie the voltages for which we design our DVFS). We optimize the collections of capacitors so that when we are operating in the grey zones we are in the above-baseline efficiency portion of the curve, and we don't really care what happens to efficiency outside those voltage ranges of interest.

The patent in fact claims (like other patents we have seen) that, after design has optimized the capacitors in general, each chip is tested to figure out exactly what the bounds are for each grey box (they will depend on the precise reactance and load characteristics of that particular chip), and that is recorded then used to determine the optimal DVFS settings for each subpart of that particular chip.

inductors integrated into the RDL

Of course a full solution needs not just capacitance but also inductance. (2015) https://patents.google.com/patent/US9748227B2 *Dual-sided silicon integrated passive devices* contains figures 11 and 13 which show two ways of creating an inductor using essentially the sort of thin metal film technology that would work in the RDL.





A scalable future

We' ve seen how the M1 Ultra is probably manufactured. But how do we go beyond the M1 Ultra? On the one hand, we want to include more logic; on the other hand we also want to include more DRAM. The existing Ultra design is neat, but doesn't really scale beyond two chips.
- with say four chips created on the initial wafer, there's too high a risk that at least one of the four is defective in some way
- the geometry of the M1 has communication on the north/south sides, and memory on the east and west sides. You could stretch this to four chips in a row, but you'd probably prefer a more compact arrangement, at which point there's a question of how we layout the chip-to-chip communication vs the memory.

(2018) https://patents.google.com/patent/US20190319626A1 *Systems and methods for implementing a scalable system* answers some of these questions and reveals an astonishing vision for Apple, vastly more ambitious than even the wildest have assumed!
The new items that are added are "communication" and "memory" bars which are small pieces of communication "hardware" added after the primary SoC is manufactured. In simplest form, these could be RDLs added via BEOL as described above; or they could be fancier being active silicon added either like EMIB bridges or like CoWoS wafer-on-wafer.

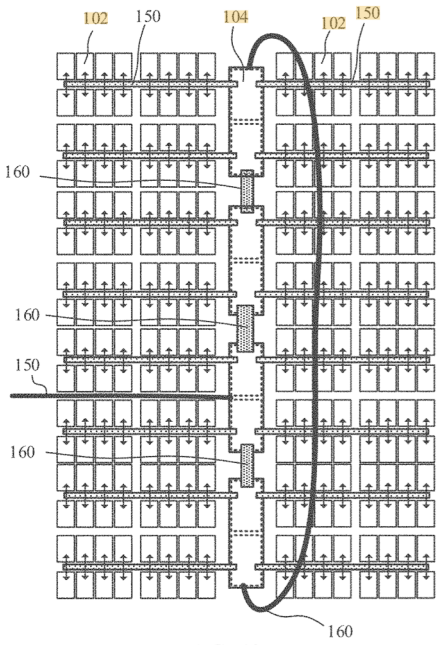


FIG. 13

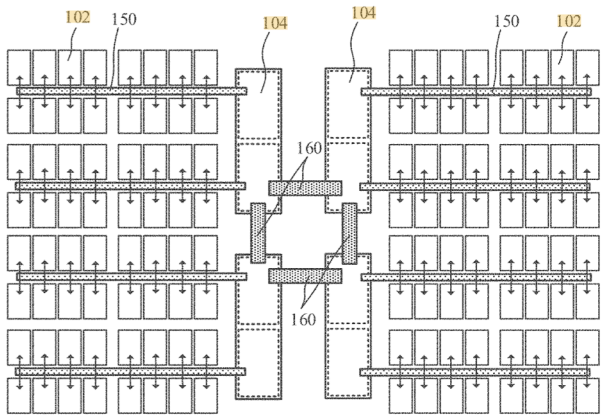


FIG. 14

Consider first Fig 13 above. We see pairs of M1-Ultra like SoCs, but these pairs are joined by the grey communication bars labelled as 160. These bars allow Ultra-like (very wide, vary fast, low power) chip to chip communication. Alternatively we can create two docking sites for bars on the base SoC, allowing for geometry like Fig 14.

Along with the communication bars, we also have memory bars, thinner, and labelled 150. These again attach to appropriate docking sites on the base SoC, but allow for much more memory to be connected.

Note that, physically, this is very different from traditional multi-socket DRAM! Multi-socket DRAM has multiple DRAM chips connected to the same shared traces, with problems in terms of power, frequency, and those shared traces can only communicate data from one on the DIMM chips at a time. The Apple scheme has an extremely fine RDL manufactured on the communications bar so that every DRAM chip can have its own dedicated path to the memory controller on the SoC.

You can now improve this scheme in multiple ways. One is that, as you may know, analog PHYs (ie the analog circuitry that talks to off chip hardware, whether DRAM or PCIe) do not shrink in size nearly as rapidly as digital transistors, and so it's a shame to have to waste much of your leading edge silicon area on these less demanding analog circuits. The patent suggests that if the memory bars are active silicon, not just an RDL, they can be manufactured in an older process and can have PHYs built into them, removing the PHYs from the base SoC.

This same idea can be used to create PCIe bars that attach to an IO docking point on the base SoC.

They also suggest moving as much of the chip-to-chip routing logic as possible to an active bar, which obviously means the base SoCs (which will sell to many more customers) can be a little smaller and cheaper.

Alternatively you can move memory logic (for example ECC logic, or even memory compression and

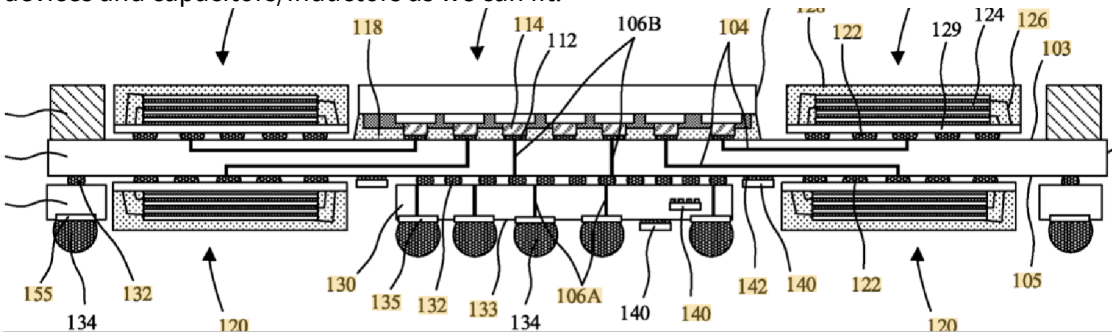
some caching) to the memory bars, in a way that again nicely scales to the number of memory chips rather than requiring this logic to be present at one fixed size in the base SoC. Or you can have memory bars that support CXL or some other alternative memory standard.

Alternatively you can even have the communication bar as optical! Apple suggests this for the curved communication “bar” in Fig 13 linking the top to the bottom via an electro-optical base stub and optical fiber.

The overall pattern here, as you can now see, is some docking sites on the base SoC, and the uses of (possibly active) communications/memory/IO bars to allow scalable (and independent...) growth of either the number of base SoCs, or the amount of memory, or the amount of IO.

a simpler (intermediate term) solution

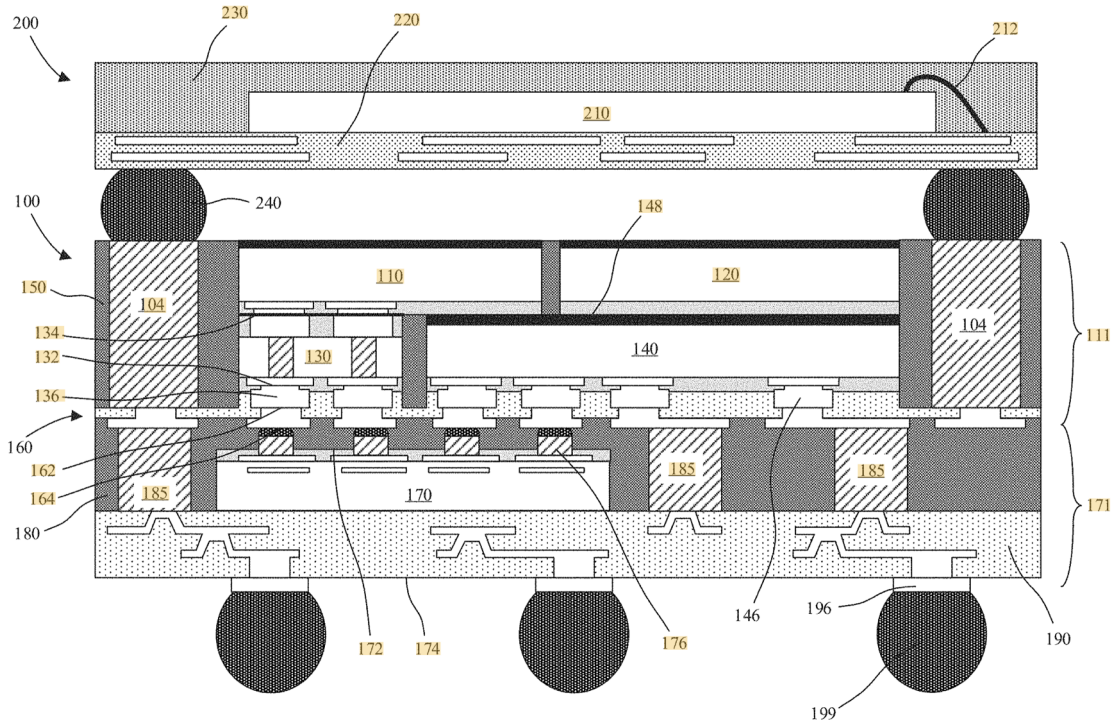
What about for the cheaper, M1 class, devices? Apple also has that covered in the event that they want to double their RAM: (2018) https://patents.google.com/patent/US10685948B1 *Double side mounted large MCM package with memory channel length reduction*.
This is the basic M1/A12X design, but we also put RAM on the reverse side of the package, as in the image below.
As you can see, we also use the extra height below the package to include as many voltage regulation devices and capacitors/inductors as we can fit.



This same idea, of course, could be scaled to Pro/Max/Ultra class devices as a way to double DRAM before we go the route of memory spines.

vertical substrates

Check out this image from (2019) https://patents.google.com/patent/US10770433B1 *High bandwidth die to die interconnect with package area reduction.*



I know you have seen so many of these your eyes start to glaze over, but each example is different, and the trick is to note what's familiar then look for what's new.

Obviously we have a PoP structure with DRAM as the upper package and logic as the lower package. What's interesting is that lower package.

We have seen two ways to compose chips into a single unity.

One possibility is side-by-side (2.5D).

- Ideally we can form an RDL which provides communication between the two.
- Second choice is a bridge chiplet to communicate between the two.
- If we have to, we can utilize some sort of substrate (thicker and less electrically ideal) to communicate between the two, a silicon interposer, or even an organic interposer which starts to look like a mini-PCB.

Alternatively we can stack dies on top of each other. But now how do they communicate?

- We can use wire-bonding like DRAM and flash. But that's a real pain if your connections are extremely regular, and who wants those tiny wires complicating subsequent processing?
- We can use fan out layers then solder bumps, like how the DRAM package communicates with the logic package. But those actually use a lot of area, use a lot of energy, and can't be switched as fast as we'd like.
- We can drill via's through the silicon chips, coat them with metal, and have those form vertical communication channels. This is how HBM works, but TSVs are still an expensive technology and Apple seems to want to do everything to avoid them! So many of their packaging patents start with a TSV solution as current art then go on to "here's how we can do it avoiding the TSVs".

So with than in mind, look again at the patent. We could have placed chip 110 side by side with chip 140, so that they share RDL 160. But that would increase the package area.
If we stack 110 fully above 140, we have the problem of communication.
But if we stack 100 with some overhang to the side of 140, then we can build up RDL 160 to have the vertical element 130 (not very high! just a few tens of microns) which can communicate with pillars on the underside of chip 110. This gives us most of the advantages of vertical stacking, but replaces the (rare and expensive) TSVs with more common RDL technology. The rest of the items are just there to make the patent more general! 120 is “mechanical” silicon, blank, non-connected to provide mechanical stability if chip 110 is too small relative to chip 140. Then chip 170 is just there to show that, if you want to be extra fancy, you can make RDL 160 two-sided, and stack in an additional layer (at not too much additional expense) via using the second side of the RDL.
So once again we have a somewhat scalable solution for the next few years. Perhaps the easiest immediate growth direction is to use the second side of RDL 160 to communicate with small chips like 170 (but there are limits to how much of this you can do because there needs to be enough space to push through some connectors all the way to the outside world. Once that growth path has been taken as far as it can go, we can then stack a third layer slightly offset relative to the second layer (again with some assumptions, this time that we can route all the connections required along part of the lower surface of chip 110 rather than having to use all of the lower surface).
As before, I remind you that we may not actually see these ideas in desktop chips! They might appear in iPhone chips, but it’s with something like an Apple Watch where space is so important that you are willing to pay a little extra in packaging to reduce area by 20 or 30%.