the art of boring™

EP 223 | Memory, Part 1: From Boom-Bust Commodity to AI Bottleneck

[00:00] [Rob Campbell] [RC] Right now, as you listen to this, your brain is doing two things at once: holding the sentence in short-term memory and drawing on years of long-term memory to make sense of it. Computers work the same way, and both sides of that equation have been at the centre of one of the biggest stories in markets lately—with memory companies generating returns over the past year of hundreds, if not thousands, of percent. And it's also worth mentioning, on any given day, week, or even month, these stocks have had big swings in both directions.

Today, Part 1 of a three-part series on memory with my colleague Shan Yeo, who is based in Singapore. In this first installment, a bit of memory 101: what DRAM, NAND, and high bandwidth memory (HBM) actually are, as well as how the industry has evolved over the past 50 years leading to the supply and demand bottlenecks—and the explosion in profitability—that we're experiencing today.

In Part 2, we cover some of the bigger risks: overcapacity, technological change, and greater competition from China. And then finally in Part 3, Shan leans into his role as an equity analyst and talks through how he's thinking of valuations as well as putting big share price movements in context.

Throughout the series, we get a bit technical in places, but when we do, Shan is so great at making it real. If you want to better understand why an iPhone's memory content went from $50 to $400 in the space of a year, this three-part series is for you.

[01:23] [Disclaimer] This podcast is for informational purposes only. Information relating to investment approaches or individual investments should not be construed as advice or endorsement. Any views expressed in this podcast are based upon the information available at the time and are subject to change.

[01:44] [RC] Shan, welcome to the podcast.

[01:46] [Shan Rui Yeo] [SY] Thanks, Rob. I think it's very exciting. It's my first time on the podcast.

[01:50] [RC] Well it's your first time and it's your much-hyped appearance, I think. On my last two podcasts, I spoke with Wen and I spoke with Paul, and in each conversation, we referenced this upcoming conversation on memory, as listeners will see.

And as most people probably know, one of the defining topics of the last 12 months in markets has been not just AI, but this particular niche within AI that has had eye-popping returns. Now, at Mawer, in our portfolios, we've had the benefit of owning a few of these securities.

I know you've spent a lot of time, not just in the past 12 months, but in your career, on this particular industry. And

we thought we'd take the occasion to do a bit of a deep dive into the history, what's changed, how we're looking at things right now, and how we're managing positions in what's been a pretty crazy sector in terms of the returns that have been generated.

But as we've seen a little bit more recently, it's certainly been volatile, with big single day moves both for the positive and for the negative. Let's dig in. And there are aspects of this conversation that I imagine will get quite technical. But if we can start, just to set the stage on this topic of memory in general, can you walk us through the various types of memory and how they're used?

[03:08] [SY] Okay. So broadly, there are two types of memory. One is DRAM and one is NAND. For reference, it's like for humans we have short-term memory and long-term memory.

So DRAM is your short-term memory—what you actively hold in your mind right now, like when we're doing this podcast. It's fast to access, limited in capacity. And now when I go to sleep tonight, I forget about everything. So same for DRAM: when you power it off, it loses all data.

NAND, on the other hand, is what we call the long-term memory. It is durable. It holds a huge amount of data. But it is slower to retrieve. I will try to recall what happened five years ago. It's going to be some time to retrieve it. So there is a latency involved, but it can store your data for years.

So every major electronic device requires both types of memory: smartphones, PCs, data centers, AI, robotics. So every consumer electronic requires memory. So these HBMs are 20 times faster than your usual DRAM, and they are a lot more expensive. But this is the brief overview of memory.

[04:47] [RC] So high bandwidth memory is just on steroids in the way that they're stacked on top of each other to achieve the type of performance. Okay, that's helpful in understanding the landscape. And by the way, I can imagine that the reason that you have the short-term memory is because it would be really expensive trying to store everything in long-term memory.

There are things that you don't need to carry over when you turn the machine off or when you go to sleep at night. There are aspects, from an efficiency perspective, and I'm sure we'll come back to this later in terms of where we might be going in terms of efficiency, there's a reason to have both types of memory.

[05:18] [SY] There is what we call the memory hierarchy. So the memory hierarchy is: the higher you are, the more expensive, the better the performance, but the lower the capacity. So at the top of the memory hierarchy, we have the SRAM, followed by the HBM, the DRAM, and the NAND and the hard disk drive.

DRAM is more expensive, and you can't use too much of it because it's expensive. If you don't need the performance because you only use it once every five to ten years, you store it in your NAND.

[05:33] [RC] Thinking back over the past year, I think some of these stock prices have just exploded because a real bottleneck has emerged. And Paul referenced this just in terms of Apple raising their prices by 20%, I think, just a couple of weeks ago, largely based on supply and demand for memory. Can you give us a bit of a historical perspective on the industry? Some listeners will have heard and know that memory historically has been brutally

cyclical. Can you trace a bit of that history for us, up through where we are today?

[06:25] [SY] I think historically, memory has always been highly cyclical, both DRAM and NAND. They are the commodities in semiconductors. So essentially, your demand and supply drive the price, and one supplier's gigabit is interchangeable with another. There is no difference of a DRAM from, let's say, a Micron and an SK Hynix. There is an industry standard body called JEDEC which sets the standard for everybody to follow.

They remain cyclical, and they are always cyclical, because there is a mismatch in the elasticity of demand and supply. On the supply side, it takes about two years to get a new fab, one year for equipment to move in and qualification, and another half a year for wafer to go in and for the wafer to come out. So a total of about three to four years to get incremental greenfield wafer capacity. So if I'm a supplier today, I must focus what demand is in 2030 for me to actually build my capacity today, and demand changes pretty frequently.

So there's always this mismatch in demand and supply, which has historically driven the cycles every three to four years. But throughout all these cycles, I think the industry revenue has continued to compound at about high single digits over the last 30 years.

[07:56] [RC] Given the growth and demand for memory at large But I would assume, given your description of this notion that when a customer doesn't really care where they're getting the memory from, it's commoditized—it's the same thing from each supplier —I imagine that despite growing revenue, there might not have been a lot of wealth creation historically, just with a lot of competitors in this space.

[08:18] [SY] Yeah, absolutely right, Rob. I think the memory industry is what we call a wealth destruction industry, for more than 40 years, ever since DRAM was commercialized in 1970. I think nobody earned a return above the cost of capital through the cycles over the last 30 years before 2010.

In the 1980s, there were more than 30 companies involved in memory, including companies we know of today such as Intel, Texas Instruments, and Infinium. So these are big companies today that we know are associated with memory, but back then, almost every major semiconductor company was involved in DRAM.

So, in the 1980s, the Japanese firms competed with better cost structure, because it is a commodity. So one, issue was when Japan was subjected to the Plaza Accord, where the United States wanted to adjust the exchange rate, just to make sure that the American companies were more competitive. And I think another big move back then was in 1985: Intel made a strategy pivot under the previous CEO, Gordon Mooreto the withdrawal from the memory market, because they recognized it is a commodity that they do not have an advantage, to focus on microprocessors.

So because of this wealth creation I think what happened subsequently were numerous spin-offs, mergers, and bankruptcies. So today, the likes of Intel and Texas Instruments, they are not involved in memory anymore.

I think what really changed for the industry was in 2013, when Micron bought over Elpida and the competition became rational. So overnight, the industry consolidated to three main players: Samsung Electronics, SK Hynix from Korea, and Micron in the U.S. So these three companies control more than 90% of the DRAM market today, and they became a wealth-creating industry ever since 2013. So these three companies, from 2013 to 2024, if we

exclude this supernormal return of AI in 2025, have delivered at least 15% return on capital. So this was across three up-cycles and three down-cycles, with only a single year where they went into losses. So clearly they have earned above their cost of capital throughout the cycles.

What changed, I believe, was a mentality change. They understood that it is difficult to consolidate the industry further. If you go from three players to two players, it is actually pretty difficult there be antitrust concern**.** So perhaps what is better to do is to start to monetize, rather than to compete with each other for market share.

And secondly, I think Moore's Law slowed down. So previously, you could get a huge competitive advantage by being one year ahead of your competitor. So by being one year ahead, you could be 50% more cost-efficient. But because of how Moore's Law has slowed down, you could only get a single digit more bits per wafer from tap migration these days. So nobody has enduring cost advantage. So I think that's when the industry started to mature, after 30 years of this competition.

And on the behavior side, there is also evidence. So in the down cycles what you see is that instead of further capacity expansion, they start to announce production cuts on their earnings calls. Micron started with production cuts, SK Hynix will follow. And then finally, Samsung will say , okay, we are going to cut our DRAM and then production.

[12:26] [RC] So some degree of rational competition.

[12:29] [SY] Yes, rational competition definitely helped. And I think another aspect here is that both Samsung and SK Hynix have started to introduce formal dividend policies, where they commit to return 50% of their free cash flow to shareholders. What this means is the excess capital is returned to shareholders rather than invested in new capacity. So you definitely see signs of this rational competition that has changed the industry.

[13:00] [RC] Now, that might be a good description of the last 40, 50 years. So for most of that, as you mentioned, wealth destroying; maybe a change 10 to 15 years ago with consolidation, where it's become a little bit more wealth creating; but a much bigger change in the last year.

[13:16] [SY] I don’t think anyone who has looked at this industry for the last 50, 60 years, has ever seen anything like this before. And what really changed was that I think AI is memory hungry in a way that no previous computing wave was. So I think the technical understanding here is that memory is the key bottleneck in inference.

So, inference is when you type a problem into ChatGPT and you get results from it. There are two different stages of inference: one is pre-fill and one is decode. Pre-fill is where the model processes your problem to produce the first token. The speed of pre-fill is determined by the amount of compute you have. But from the first token to the last token, the speed of token generation after the first token is determined by what I call the decode stage, where memory bandwidth determines the speed of every subsequent token generated.

For the first time, I think memory is very important to the performance of the AI. And the fundamental reason is because we are in the large language model era where there is what we call the attention mechanism. Every new token generated is computed based on the relationship with all the previous tokens generated. Let’s say I do a

deep research and I have 50,000 words generated. The 50,000 words generated are based on the relationship it has with all the 49,999 words before that. That is a huge amount of computation power needed.

To avoid this computing every time you generate a new token, there is something called the KV cache, or the key value cache, where this is stored in the memory, which is why memory is the key bottleneck for inference. And when we have an agentic flow where it runs for hours, this definitely adds a lot of pressure to the memory content.

And we have also seen that with every new generation of GPU, there is significant content growth in memory. NVIDIA will release the Rubin GPU this year, which comes with 384 gigabytes of SOCAMM DRAM per GPU. This is an increase of more than 50% compared to Blackwell, which was released last year.

So just for reference, 384 gigabytes is the equivalent DRAM demand of 32 iPhone 17 Pros' DRAM content. If we assume 5 million GPU shipment, that is the equivalent DRAM demand of 160 million iPhones, or two of the total iPhone shipment this year. This is a huge pressure, suddenly there are 160 million more iPhones required. Any second if we look forward, NVIDIA will launch the Rubin Ultra GPU, which will have 33% higher HBM content compared to the Rubin GPU this year. So, we do have visibility into AI memory demand, even if the unit demand for GPUs stays the same.

[16:39] [RC] I think that's a key insight. The demand for GPUs and the number of GPUs might not grow. I mean, it has been growing, but it might plateau. But you're saying the amount of memory that's required for those is increasing 30, 50% year on year with each release of these new GPUs. So far, we've talked a little bit about DRAM. On the high bandwidth memory side, the HBM, presumably we're seeing even greater demand for those.

[17:07] [SY] HBM is very important and it's very much required by AI. HBM prices are three to four times higher than your traditional DRAM. And it has huge implications for DRAM supply and demand because of what we call the trade ratio. The current trade ratio for HBM is three to one, which means that for the same gigabyte of HBM demand, it consumes three times the wafer capacity of a traditional DRAM. This further disrupts the demand-supply balance for memory.

HBM dies are physically larger than your DRAM die because you need to drill what we call the TSV. So, imagine you have an apartment where you need a bigger space to accommodate for the elevator shaft. This is exactly what happened for HBM. It has to be a bigger and bigger die in round wafer means you get less dies per wafer and more likely to suffer from defects. And when you try to set 12 dies together, all you need is one bad die and you have to scrap the entire 12 dies.

[18:17] [RC] The technical difficulty goes up in the production of these, where in terms of profitability, that yield becomes that much more important.

[18:24] [SY] Yes, and lower yield means lower supply. All of these really drain a lot of DRAM supply. So just some numbers: I think in 2024, HBM accounted for 5% of bit demand, but they accounted for 15% of the total DRAM wafer demand. And if I look at 2027, the forecast is that HBM will account for 30% of total DRAM wafer demand. And the trade ratio, as we thought about, 3 to 1, will further go to 4 to 1 when we move to HBM4 next year.

[19:03] [RC] We've talked so far, Shan, mostly on the demand side of things, just how the number or the amount of memory that's required has increased. When I think of a bottleneck, it's a combination of both supply and demand. Can you talk a bit more on the supply side? I mean, you mentioned that the industry has gotten more rational, the number of players has gotten smaller, it takes years to sort of plan out your capacity expansion. Speak about the supply side of the bottleneck that we're experiencing in the moment.

[19:29] [SY] Supply is also a very important part of the equation for how we got to this shortage we have today. Based on the supply, unfortunately, the supply-demand balance will only get worse in 2027, because there's very little new supply that's coming online in 2027. There's three to four years to bring new wafer capacity online from scratch.

And back in 2024, I think the companies just came out of a down cycle. Everyone was very conservative, everyone was worried about overcapacity, and no one foresaw how much higher AI inferencing demand would have been today. No one planned for this capacity expansion.

So, despite all this huge jump in DRAM content and HBM trade ratio, I think DRAM wafer capacity will only grow by about 10 to 12% in 2027. And if we include migration bit growth of about a single digit, we could only get about 17 to 18% DRAM supply growth in 2027. And DRAM demand is expected to grow much further, just because of the amount of new memory content that the GPU suppliers have shown. So we expect DRAM shortage to widen from about single digit this year to low to mid-teens in 2027. Many new supply only arrives from 2028. So the shortage will ease in 2028, but our expectation is that there might still not be sufficient supply if the AI demand continues as it has today.

[21:13] [RC] So what's happened, then? You've got this massive mismatch between supply and demand for an industry that, even when the industry turned wealth creating, my understanding is that pricing was still deflationary, meaning that prices would go down over time. That's changed radically. Can we speak a little bit about that? Just help us understand the nature of how the business has changed on that front.

[21:37] [SY] I think you're right that price should go down over time. I think this is the nature of the entire semiconductor industry, because ultimately it is driven by Moore's Law. So just to recall, Moore's Law is a business model. It came from Gordon Moore, where it says that the number of transistors will double every two years. What will happen with double the amount of transistors, your cost goes down. And when your cost goes down, your price should go down. So that should have been what the industry is. Over the last 50, 60 years, the price has always comes down over time.

We have a mismatch right now where price is going up. And obviously there are demand responses to price increases, because there is, first of all, not enough capacity. And secondly, customers also cannot handle the price increase. So just an idea of how much price has gone up: since the start of the year, DRAM contract price has gone up by about 200%, and compared to one year ago, price has gone up by 400 to 500%.

So just now we mentioned the iPhone 17. It has 12 gigabytes of DRAM. That 12 gigabytes of DRAM may have cost about 50 US dollars last year, today, it costs about 400 US dollars.

If I'm Apple, that is a huge amount of cost that I have to bear. And for all of these smartphone and PC OEMs, they

do not have high gross margins. They have to either reduce the memory content or start to increase the price to pass it on to customers. So not only iPhones, but even Nintendo Switch, laptops have all announced price hikes.

Earlier this year in the Singapore we were like, let's go out and replace our iPhones and laptops, because memory price is going to continue to go up. And when prices go up, what we see is that demand will come down. So, you may never buy an AI server, but as a consumer, you are already paying for one. This is part of how the demand responds to prices, that has historically been a way for price to eventually come down, because demand destruction takes place.

[24:04] [RC] I can imagine another feature is, with so much CapEx taking place on these data centers, and this notion that speed to get there is that much more important, I imagine that customers might be a lot more willing to lock in longer-term contracts, whereas pricing might have been day to day or quarter to quarter historically. Has that changed in terms of the nature of contracts of customers too, in terms of locking that in and ensuring that they have that supply regardless of the cost?

[24:32] [SY] Historically, I think the industry has always been cyclical. So, everyone has tried to make it less cyclical. So historically, memory was sold on quarterly contracts. When we moved to HBM, I think the AI customers like NVIDIA and the other AI accelerator customers, other than price increase, I think they're worried that they do not have sufficient memory supply. So if you do not have sufficient memory or HBM, you might not be able to ship out your GPU. Without memory, you can't sell even your phone or your PC.

And as we all know, and the industry is aware, supply increase is very limited in 2027. So customers are starting to sign three-to-five-year agreements, what we call long-term agreements, or LTAs, with the memory suppliers, where they lock in the volume commitment with certain amount of floor and ceiling prices. These are not fixed prices, but they are a range where the price can fluctuate.

So, there is the question of whether the customers will walk away from the LTAs when this cycle is done, because the industry has tried LTAs many times in the past, and usually it does not work because when the cycle turns, customers will come back to renegotiate the contracts. But I think there is some change this time around, which is that the LTAs involve some form of prepayment. So, if the customers try to renegotiate, there is a penalty to it.

These LTAs are a lot more likely from that in the past. But I think that if the cycle rate turns and memory prices collapse, I think it's unlikely that these LTAs will hold. But at least the memory suppliers get to collect the prepayment. I think these LTAs will be important drivers of re-rating for the memory companies, which are trading at low single-digit forward price-to-earnings ratios, because of the lack of medium-term visibility into the memory cycles. If the LTAs turn out to be binding through the down cycle, I think we will definitely get much better cash flow visibility into the future.

[27:00] [RC] And certainly, so far, maybe there's been some multiple expansion. But in terms of the stock prices, it's really been based on the earnings that have come through and those that are expected, as you mentioned, just given the way that pricing and volume and all that has changed in the last little while.

That’s it for Part 1, and largely a pretty optimistic picture. Tune in next week as Shan puts on his black hat and outlines some of the key risks.

EP 223 | Memory, Part 1: From Boom-Bust Commodity to AI Bottleneck

Hey everyone, Rob here again. To subscribe to the Art of Boring podcast, go to mawer.com. That's M-A-W-E-R dot com, forward slash podcast or wherever you download your podcasts. If you enjoyed this episode, leave a review on iTunes, which will help more people discover the be boring, make money philosophy.
Thanks for listening.

Companies Mentioned:

Samsung Electronics
SK Hynix
Micron
NVIDIA
Intel
Texas Instruments
Apple
Nintendo