the art of boring™

EP 224 | Memory, Part 2: The Risks from China, Technological Change, and Overcapacity

[00:00] [Rob Campbell] [RC] Welcome back for this second installment of our 3-part series on memory with my colleague Shan Yeo. We're going to dive right back into our conversation and this time through the lens of risk. If you missed part 1, go back and listen to episode 223, where we cover the basics of memory and how an industry that has spent decades as a graveyard for capital turned into a genuine wealth creator.

[00:25] [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.

[00:41] [RC] Shan Rui, I want to flip to the other side. I think you've painted a picture that really sets the backdrop for why we've seen the returns that we have from the industry. Just really big changes in the demand picture, the supply picture.

Can we talk about some of the risks? I'm sure we'll get to the overcapacity one eventually. But what I wanted to start with is you mentioned the three players, the Samsungs, the Microns, and the SK Hynix's. There are other players out there, to my understanding, that might be a little bit behind, and in China. Can you talk about what's happening there?

[01:16] [Shan Rui Yeo] [SY] I think China has always been the biggest fear of competition for, I believe, almost every industry in the world today, whether it's electric vehicles, not just industrial, not just memory. But I think the memory threat is real. There are competitors in China.

For DRAM, the first company is CXMT, ChangXin Memory. It's supported by state-owned capital, and I think it's about to IPO in the Chinese market. It's not easy to enter DRAM because of the amount of capital and IP required, and how CXMT tried to go about it was that they acquired the licences and DRAM patents from this company called Qimonda. Qimonda was a memory subsidiary of Infineon that went bankrupt in 2009, and this was how they managed to become the fourth largest DRAM company today.

I think everyone should be concerned about the potential oversupply rates from CXMT. They are going to turn to the domestic capital market to fund capacity expansion. And unlike the big three DRAM producers, who are driven by economics and the market, I think CXMT is unlikely to exhibit the same capital discipline that the three DRAM players have, because the top priority in China today is self-sufficiency, so that China is not reliant on foreign semiconductor control.

What happened over the last seven years was that the U.S. administration controlled the export of semiconductors and equipment services to the Chinese. So now their top priority is to make sure that they are not reliant on foreign imports.

In terms of DRAM technology, I think the memory shortage has allowed them to gain access to a new customer base and even under supernormal profit**.** There was news that Apple is interested in buying from CXMT, although this is a grey area because of the control by the U.S. administration.

A year ago, CXMT was in losses, and for the first half of 2026, the forecast is revenue growth of 200% and a net profit margin of close to 70%. IPO and high profitability, I think these are catalysts for CXMT to aggressively expand capacity. By capacity, I think they are about 15% of global DRAM capacity, and they have planned to ramp up multiple new fabs across China.

Should we be worried? I think there are reasons why this is less of a concern in the short to mid term. The U.S. government has implemented equipment export controls such that key equipment suppliers like Applied Materials, KLA, and Lam Research are not allowed to ship equipment capable of producing DRAM at 18 nano or better to Chinese customers.

For reference, the big three DRAM companies are around 12 nano today. The three memory companies have started to use EUV machines from ASML, and ASML is not allowed to ship any EUV equipment to China. So though CXMT is expanding capacity, the effective supply share is much lower than the capacity share. Compared to their 15% wafer capacity share, I think their effective supply share this year is about 10%.

And secondly, I think they are certainly qualified for smartphone and PC, but to achieve true self-sufficiency, they have to invest in HBM. And as we recall earlier, HBM has a higher trade ratio and the lower U rate. So with a legacy DRAM node, I think CXMT has even higher difficulty when they try to invest in HBM. A lot of this incremental capacity will actually be consumed by their investment in HBM.

On the NAND side, the counterpart is Yangtze Memory, YMTC. Founded in 2016, I think they were put on the U.S. entity list in 2022. But its global NAND share actually jumped from 8% a year ago to 13% last quarter, with revenue growth of 450% year on year. And unlike DRAM, I think for NAND, China is a lot closer with their global peers, just because it's easier to achieve higher density.

The way you achieve higher density for NAND is you stack multiple NAND layers together, and this is easier than when you try to shrink for DRAM and logic. And they do have their own proprietary Xtacking architecture, where they do wafer-to-wafer hybrid bonding. Just for reference, I think some licensed hybrid bonding patents from YMTC a year or two ago shows you that the technological threat from YMTC is a lot more real than from CXMT.

[07:09] [RC] So there's potentially a supply response that might come from China, certain categories in memory, maybe more than others. But I think, if I heard you right, reasons why that may not come to fruition given some of the export controls that have been put in place in terms of the tools that these companies need to actually produce at the highest technological levels.

Reminds me of another question, though, that I think is in the category of risks to the thesis that we've seen play out, and that is just technological change. Throughout history, there has been a wonderful propensity in capitalism, when there is a bottleneck or a shortage, for technology to come around and just create a solution based on pressure that wasn't there before.

Recognizing that we are in the realm of nanometres and this is still something very difficult to do, what are some of the technological changes that you think are more or less likely that could really upend, either in terms of something brand new that might displace, or do you think it's possible that customers might sort of be satisfied with something that's not at the very cutting edge?

The analogy I'm thinking of here is a little bit like DeepSeek, which, compared to ChatGPT and some of these other models, my understanding is it does 90% as good a job with a lot less cost. Just curious whether you think customers will settle for that given some of the dynamics that we've talked about. But a long intro to say, what are some technological changes that you're thinking about?

[08:37] [SY] This is always a risk that I think the industry faces. The argument is that they also benefit from this technological change. I think if AI demand continues to grow, the industry players are aware that there's insufficient supply to meet the required demand. And at the same time, memory will get more expensive, which is not what they want, because they want the AI token price to come down over time for AI inference to proliferate. So I think both NVIDIA and the AI funded labs, everyone is working hard on solutions to reduce the consumption of memory over time.

I would say there are three main broad areas of solution. The first is to improve AI model efficiency through software. These are methods like quantization. The counterargument here is the Jevons paradox: historically, efficiency gains in computing have grown consumption rather than reduced it.

One example is GPUs. If you look at Blackwell versus Hopper, I think Blackwell can generate 35 times more tokens than Hopper, but that has not actually driven consumption down. In the case of memory, when we get cheaper tokens, we don't pocket the savings, we spend them. In the past, when I did deep research, maybe I only fed it five to 10 documents, because that was the context window the memory allowed. Today, I can feed it 30 to 35 documents, so my inference is much better than before. So I'm not too concerned about AI model efficiency.

Second, I think AI accelerator designers are working on alternative memory solutions that exploit the entire memory hierarchy, as we talked about: SRAM, HBM, DRAM, NAND. This year NVIDIA introduced something called the LPU, which was based on SRAM. And going forward, I think the industry is working on a few approaches, such as DRAM memory, where you share your memory resources across multiple AI accelerators. There are also solutions such as high bandwidth flash, where you use NAND to actually offload some of the KV cache from HBM to DRAM, and from DRAM to SSD. And for high bandwidth flash, it will be to offload some of the function of HBM.

So there are definitely a lot of solutions being explored today. That's a niche market within the DRAM and HBM demand, but I think NAND is sufficient to change the overall supply demand balance. And in fact, they may even drive higher DRAM and SSD demand as you try to offload it down the memory hierarchy.

I'll say the biggest technological risk here is that memory is a constraint only because of the large language models, specifically the attention mechanism, which forces the system to hold huge amount of content. So the biggest risk is that large language models are not the path towards artificial general intelligence. There will be a new fundamental architecture shift that could arrive, and this new paradigm is not memory hungry.

Given how fast AI has evolved, I think this is a real possibility that we cannot discount. It is low probability right now, but you can't say it is zero if we just put the timeline a bit longer…which is why I think the terminal value is a distribution rather than a phase number.

[12:56] [RC] And so, low probability, at least for the moment, but high impact if that were to get there. Can we talk about another risk, which I think is front and centre for a lot of listeners, that might be in the category of there actually might be a greater probability of this, which is just the notion of overcapacity. Clearly there have been a lot of IPOs in China and elsewhere.

I think SK Hynix, for example, just tapped, not the IPO market, but capital markets in the U.S. And the reason for this is to be able to expand capacity as quickly as possible. How are you, and we, thinking about this notion of overcapacity, or just how much demand there is and for how long?

[13:40] [SY] I think it's inevitable that there will be oversupply in the future, because the industry is still a cyclical commodity. So the question is always, what's the duration of this undersupply today? From the customer's perspective, prices have gone up so much, so I think there is also a lot of pressure from the customers to invest in capacity expansion, because I think they are still very optimistic about where demand will be in the future. And if demand goes to where they think it is, they clearly need the memory suppliers to expand capacity more aggressively.

Last month, the CEOs of Samsung and SK Hynix appeared in front of the Korean president, they made an announcement that they will accelerate investment in memory capacity. SK Hynix announced that they will commit to invest 700 billion U.S. dollars until 2035. Samsung committed to invest 1.5 trillion dollars until 2040. These numbers are huge, and it represents a huge amount of capacity that can come online over the next decade.

A huge amount of this investment will be back-loaded in the 2030s. It is about four years just to get all this capacity online. So what we see here is that capacity for the next five years can double by 2030, and if AI demand growth continues, supply can continue to be tight until then.

There's also the question of whether there will be sufficient EUV capacity from ASML if everyone decides to expand, because it's not just the memory suppliers that have to expand capacity. TSMC has to expand capacity, Intel has to expand capacity, and everyone needs EUV capacity. And it's not easy to expand capacity for ASML.

For reference, I think ASML just guided that their capacity for EUV machines in 2027 is 85 EUV machines. Everyone is going to fight for the allocation of this equipment. So you can have trillions and trillions in investment plans, but you could still be a bottleneck by the equipment suppliers. So without the EUV, they'll be a bottleneck to the DRAM capacity expansion.

But of course, eventually, I think oversupply is very likely, whether it's five years, three years, 10 years, or 20 years. As far as demand is more than supply, prices go up, and there'll always be an incentive to invest in more capacity. But I do think that this is also within the market expectations because if we look at the memory companies, they're trading at three to five times forward earnings. So nobody is saying that these companies are going to enjoy this supernormal profit forever.

And for NAND, I think the demand supply although it is a bit better in the shorter term, because the DRAM companies are all focused on DRAM capacity expansion rather than NAND, because it's more profitable to produce DRAM over NAND, and they know that throughout the cycles, DRAM margin will be higher.

Some of them have even converted some of the NAND capacity to DRAM. And Kioxia has committed to building capacity in line with market demand growth. And YMTC is focused on NAND as well as decided to invest for their capacity in DRAM, although they are supposed to be the national champion for NAND. I think the supply outlook is much better for NAND. And so I think oversupply is definitely a risk that we have to take into account.

[17:40] [RC] That's it for part 2. Next week on the final installment of our series, Shan turns to valuations and making sense of big market movements.

Hi, 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, please leave a review on iTunes, which will help more people discover the Be Boring, Make Money philosophy.

Thanks for listening.

Companies Mentioned:
Samsung
SK Hynix
Micron
CXMT (ChangXin Memory)
YMTC (Yangtze Memory)
Apple
NVIDIA
Google
ASML
Applied Materials
KLA
Lam Research
TSMC
Intel
Kioxia