



[00:00] [Rob Campbell] [RC] Today on The Art of Boring, my colleague Irena Petkovic joins me to talk about where all the capital expenditures being spent on AI are actually going. Irena walks me through the different ways in which hyperscalers are monetizing their data centers, how the token economy actually works, and what needs to be true for those hundreds of billions of dollars to pay off.

[00:23] [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.

[00:41] [RC] Irena, welcome to the podcast.

[00:43] [Irena Petkovic] [IP] Thank you. Long time listener, first time caller.

[00:46] [RC] First time caller, that's right. Well, very excited to have you. And we have you on here because I know you've done a bunch of research into a question that I think we've danced around, is maybe the wrong term, but that we've certainly spent time on in other podcasts that haven't really tackled directly.

And you've spent the last couple of weeks really focused on this particular question with a few of our holdings, and that is there's so much capital spending that's going into AI. What does the return on that spending look like? Obviously, a huge question for specific businesses, but I think for the market writ large.

So can you maybe start just by setting the scale for us? What are the scale of the investments that are taking place right now?

[01:29] [IP] Yeah, for sure. And I think there are two reasons why there's so much debate around this topic in the market. The first is the point that you've just made, the scale is something we have not seen in our lifetime.

So the four big hyperscalers are expected to spend \$700 billion in CapEx this year. And to give you a sense of just how large that is, the Apollo moon landing program, which lasted 13 years, cost \$300 billion. So we've got two Apollo moon landing programs, more than two, in one year.

In terms of a percentage of GDP, this has far eclipsed the telecom or dot-com bubble CapEx for telecom players. It peaked around \$100 to \$120 billion a year. In terms of historical analogs that we can find, the only historical analog that would compare is the railroads. So it's a huge investment.

And the second point I'll make is that I think the cash flow complexion of these businesses has changed significantly over the past decade. And that's something that investors have had to reckon with.



If you think about the companies we're talking about, Microsoft, Meta, Google, Amazon, these were not always very capital intensive businesses. If you look back to the early 2010s, they were spending around 5% to 10% of revenue on CapEx. They were known as being very capital light businesses that didn't need to invest a lot to grow.

But as they built out their cloud businesses, that capital intensity got a bit higher, 10% to 15% of revenue, but still pretty stable. And now those businesses are spending upwards of 45% of revenue on CapEx. So there's been a big shift in how those businesses are creating value.

And it's understandable that investors are reckoning with what the return is going to be on these investments. I think, importantly, we as investors have to remember that if you have an investment opportunity before you that generates a very strong return on invested capital, then your free cash flow, your most efficient free cash flow rather, should be zero. You should invest as much money as you can behind that.

But I think because we invested in that capital light era, we were conditioned to see high free cash flow generation and being capital light as an intrinsically good thing.

[03:46] [RC] That's such an interesting point. And of course, we've come out of this era where so many buybacks, there wasn't this giant deployment of capital into actual CapEx. It feels different today, but it still begs that question, okay, if you're going to spend all this capital, well, gosh, there better be a return there.

So clearly, these players are seeing the potential for return. Can you just maybe go a bit deeper on what does it mean to actually generate a return on this AI CapEx?

[04:11] [IP] Yeah. When you think about that \$700 billion figure I just talked about, it's easy to ask, what are they actually doing with all that money? So, what they're doing is building data centers, and about half of that \$700 billion goes into buying chips and other networking equipment.

And then the other half goes into buying the land, buying the cooling equipment, things like that. I think there's a big shift that investors have had to make in how they think about data centers.

I was very fortunate. I got to have lunch with the founder of NVIDIA earlier this year, Jensen Huang. And one thing that he really tried to impress upon us was that the nature of the data center has changed. These aren't your mom and dad's data centers.

Previously, a data center was a cost center. You wanted to minimize the cost of it. You wanted to run it most efficiently. And it was sort of tertiary to the business. It wasn't core. But now the data center is something that can actually generate revenue. And the way that it generates revenue is that it produces tokens.

Jensen has this term that he talks about called the AI factory. And what he means by that is that now a data center takes electricity and it uses chips to actually generate tokens that can be sold.



So now investors must think about these as revenue producing assets. Things really matter like time to market, efficiency, and lowering that total cost of ownership. And that's a big lens that has had to shift.

[05:32] [RC] When I think about return on investment for the hyperscalers, it's the revenues that they generate. But if these data centers are becoming more revenue generators themselves, doesn't that cascade forward? Like the actual clients of these hyperscalers must see a return on investment that they're making into what they're spending on the hyperscalers.

I think you've done a little bit more work on that too.

[05:53] [IP] Totally. So again, it's like, how do I take this thing that was a cost center and how do I actually generate revenue from it? What is actually the breakdown of the volumes and the price of that line?

There are four main ways that the hyperscalers can generate revenue using these data centers.

The first is on raw compute rental. So, this is the traditional cloud model just extended to GPUs. This is generally the most commoditized way that they can generate a return because you're just renting out GPU hours. And that's kind of the same from one hyperscaler to the next.

The customer base here is generally more sophisticated because they have to set up more infrastructure on their end. The typical client for GPU rentals would be the AI labs, OpenAI and Anthropic, smaller AI native startups, or enterprises that want to run their own open models.

[06:44] [RC] Do you have a sense for just how much of that is dominated by the labs, by the Anthropics and the OpenAIs? It seems like that must be just a huge amount of it.

[06:52] [IP] Totally. They're the anchor tenant there. They haven't quantified it, but you're usually renting out GPU hours if you want to train a big model, and that favors scaled players who are well capitalized and they need a lot of compute to do that. So, it really is dominated by the labs.

[07:07] [RC] Okay. So that's number one. How about number two?

[07:10] [IP] Yeah. Number two is AI-enabled productivity products. So things like Microsoft Copilot or Gemini Enterprise that you can sell to people to make them more productive. So that's billed out either by usage or by user. And that is more akin to a software like margin. It's not as big as the GPU rental, but it's still a way that they can monetize it.

The third way is harder to quantify, but it's using data centers and the resulting compute to enhance their other businesses. Meta is a really good example of this. Meta is actually not renting out any GPU hours externally. They're building those data centers and using that compute to make models that help them place ads better. And as a result, if they can place better ads, then they can charge a higher price to those advertisers because presumably those ads convert to sales at a higher rate. So that's an example. You're basically just using that AI to make your other businesses better. Amazon with Amazon.com, the list goes on and on. It's just harder to quantify.



The fourth way is the newest business model for the hyperscalers. And this is where they're selling tokens. It's hard to understand what that actually means. I'll break it down. How do you generate revenue selling tokens? And then what are the costs associated with that?

A token is the volume in that equation. What actually is a token? A token is a unit of work for AI. And whenever you are typing something into an LLM and then you get a response out, you're consuming tokens.

And to give you a sense of magnitude, a million tokens is about 750,000 words. So that's roughly sort of the quantity that we're looking at. On the price side of that revenue equation, the price of those tokens is determined by how proficient that model that you're using is to generate those tokens.

So if you have a really complicated task where you have a very low tolerance for failure, you might pay more to use tokens from a more proficient model, from the frontier models. If you have a lower level task where you're more tolerant of errors and maybe there's less sort of thinking required from the AI, then you pay less to use tokens from a less proficient model. To give you a sense of just how much more expensive those frontier tokens are, it's a factor of 5x.

So the nice thing about the hyperscalers is that they can offer all those different models and a full range of possibilities for how you want to route those specific tasks. The costs associated with generating an incremental token are very low. If you think about it, it's basically just electricity.

You have depreciation, but that's a fixed cost. And the real investment is this initial data center build-out that you're seeing right now. It's a very upfront cost heavy investment.

So that's why it's a very high margin channel for the hyperscalers now. And people are paying a lot of attention to it because it's a new business model and the primary way by which they'll monetize this compute.

[10:10] [RC] Just based on your description, that's where the huge amount of investment is taking place. Well, it's across all four, but just taking that last one. That's where the big upfront investment is taking place today.

The assumption is that there'll be a return in terms of these tokens they could consume. And because there are different types of tokens, from understanding you right, the pool of potential users is actually quite wide. So those that need and are willing to pay up for the most sophisticated models will get them. But that's not the only offering. You know, there isn't just a business class on the airplane. You can buy a cheaper seat that's a little bit in the back that, to go back to the tokens, does the job that's adequate for the solution.

And so I imagine this is where the hyperscalers are really trying to envision a large swath of potential use coming out of this.

[10:55] [IP] Absolutely. And they are positioning themselves as sort of that routing layer. They're basically saying if you want to use a really cheap open weight model, you can run that on AWS [Amazon Web Services].

Or if you have a super technical task, you can pay to use Fable from Claude or something like that. So they're positioning themselves to address as much of the market as they can, which puts them at an advantage relative



to some of the labs who are more constricted to the models that they offer. Even though there is a range, they can't offer open weight models, for example.

[11:26] [RC] What evidence are you seeing so far with respect to the early signs of potential returns on this investment?

[11:32] [IP] This quarter was very informative in terms of the types of returns we're getting. There was a continuation of a lot of things that we've been seeing for a while now. So, there's continued acceleration in revenue growth in these companies' cloud businesses.

And just to give you a sense of how much these businesses are seeing accelerated growth. AWS, they have the same growth rate this quarter, 37%, as when they were half the size. So that revenue growth has really accelerated.

They're expected to add more revenue this year than all of the revenue they generated in 2019. And there's been this huge acceleration across the hyperscalers, including Microsoft and Google. You continue to see margin expansion, which suggests that this growth is very profitable.

But I think most incrementally, we got some data around the diversification and durability of that growth. Microsoft talked about how over the last 12 months, 90% of the revenues that they generated in that cloud business were from outside the frontier labs.

Earlier we were talking about how OpenAI and Anthropic were sort of dominating that GPU rental market, they're trying to make the point that we're starting to see AI proliferate through the enterprise, and token consumption is not just coming from those two big AI labs.

The other piece of evidence we got was that the executives of these businesses have started to be more explicit around the returns that they're getting on these investments. So before it was sort of like a trust me kind of thing. But now, most notably, the CEO of Amazon shared that they pay back their investments on servers and networking in less than three years. And that's a very high implied return on invested capital, well in excess of their cost of capital. That's very encouraging for investors.

And then more broadly, across those four hyperscalers' results, they talked about how they're all very short on compute, demand continues to exceed supply. And even through our own work, talking to experts and people who are starting their own AI businesses, they've talked about how it's very difficult to get supply of compute.

[13:35] [RC] So early signs, what about on the durability side? Is there any evidence or things that you're watching for as time goes on here? Because I imagine, like you said, the return doesn't come immediately. It's over time.

[13:46] [IP] Yeah, for sure. I think one thing that we're looking at when we think about durability, in addition to a lot of those metrics that I talked about, like accelerating revenue growth, backlog, et cetera, is we're talking to non-AI corporates about how they're using AI. I think that's one of the best primary sources of evidence that we can have.



When we talk to companies that aren't hyperscalers, what have you, we ask them questions like, how are you using AI in your enterprise? Are you seeing adoption outside of a coding function? Because the vast majority of enterprise usage today of AI is in coding and software engineering.

So when you hear them talk about things like it's helping in customer service, or that we built a model routing functionality, all those things are very encouraging signs because it suggests that not only is token spend growing and token usage is growing, but it's growing in a very value additive way.

[14:39] [RC] A lot of this, or at least my interpretation of a lot of the evidence that you've seen so far, seems quite supportive of the spending that's taken place so far, to your earlier point. If you have the ability to invest in a return, like you ought to do it. Can you play the other side for me though? Just put on a different hat and, what are the ways in which this really goes off course and what are some of the signs that you'd be looking for in that regard?

[15:00] [IP] One of the main ways that this could go off course is that, like I said, the majority of AI usage in the enterprise right now is in coding. So if you have all these pilot programs outside software engineering and they fail to materialize into actual programs, so it's not as helpful in customer support or procurement or what have you, then you don't have that same proliferation of AI in the enterprise. And as a result, you could have demand for tokens not keep up with falling prices of tokens.

So the price of tokens continues to fall and that has been well offset by increases in volume. You can get the demand right, but you can actually get the margin wrong, because then token prices may not be supportive of a strong return on invested capital, which is sort of one of the frustrations of this business, right? You can be right but still be wrong. That is something that we would be looking for.

The second way that this could go wrong is on the financing side. As I talked about earlier, if you have an opportunity like this in front of you that's presumably very value creative right now, you want to invest as much capital as you can behind that.

So the hyperscalers have depleted their cash flow from operations. And if they want to continue financing this build-out, they're going to have to turn to the debt markets, and you've already started to see that with Google issuing a hundred-year bonds.

And I think understandably that makes investors nervous, because the telecom build-out, which ultimately was value destructive, was mostly funded by debt. And up until this point, a source of great comfort has been that the hyperscalers have been using their own internal cash flows to finance this build-out.

A related point is on some of the circular financing that you're seeing from chip designers like NVIDIA. And that's also sort of a callback to prior bubbles.

You can sort of see both sides of this. If you're sitting in NVIDIA's chair and you see that token demand is continuing to grow and the hyperscalers may be capital constrained or unwilling to invest in continuing to build these data centers, there are certain things that you can do to encourage continued construction of the data



center.

Things like providing backstops to the neoclouds or investing in some of these companies to allow them to invest in your products. And obviously it's not a great headline, but you can sort of see both sides of this debate.

[17:17] [RC] What about just some of the popular backlash that we've seen? Maybe this is more media driven, against the actual construction of the data centers themselves.

[17:25] [IP] Yeah, it's a great point, Rob. And that's something that's a real concern for a lot of these hyperscalers and for some of the chip designers themselves. It's actually something that Jensen talked about at the lunch.

He said it's incumbent on us to explain to people that AI is a good thing and not something to be feared. But there are real environmental concerns that people have, right? When you think about a one gigawatt data center, that pulls the equivalent power of 750,000 homes.

When people talk about rising electricity prices and the grid having to expand, it's a very appealing argument to say, why should I be funding the hyperscalers' build-out? Some of the hyperscalers are paying their own rates for electricity to sort of get around this. They're basically creating small backup power and or even small power plants to fund these data centers.

But that's absolutely something that could get in the way of this build-out. And you're already seeing it in states like Texas that are sort of having a moratorium on data center build-out.

[18:21] [RC] I ask, what do you do with all this? I mean, this has been a fascinating discussion and it remains an unanswered question, right? What is the return that will be generated on this?

Is this a prudent allocation of capital? You've given us a great framework for how to think about it, what some of the early evidence might be, some KPIs that we might watch out for. But what do you do about all this?

I know you work most closely with our global equity strategy. How are you and the team thinking about some of these positions in the context of the portfolio at large?

[18:49] [IP] So I think there are three things that we can do with this information. The first is how we reflect this view in the portfolio in terms of portfolio composition and the weight of this point of view in the portfolio. So most of our AI weight is in our hyperscaler weight.

We also have some picks and shovels as well, but we positioned ourselves in a way where we're not super overweight these companies relative to the index. It's not an outsized bet that we're making for clients, and we've diversified our investments within that basket as well. So that's the first thing we can do from a portfolio composition point of view.

The second thing is that we have some natural contradictions in the portfolio. I've spent this time talking about how AI is so great for society and there's so much demand for tokens, et cetera. But we also have some names in the portfolio that are thought of as AI losers.



They're going to be replaced by AI or what have you, but we still own them and actually have been more opportunistic to add to these companies as they've come down. So examples of that would be things like some software businesses, or even some businesses in more like the professional services or travel spaces. And we have those in the portfolio, A, because we believe that some of those could actually be beneficiaries of AI, but also to counterweight some of that AI weight that we have.

The third and final thing that we can do is things that we've added to the portfolio, especially more recently, have been very immune to AI disruption. We're trying to find things that are aggressively excluded from this AI trade, if you will. So things like cement companies, things like building inspection companies, things that are just not disruptable and mostly in the physical world.

That is also sort of a ballast that helps balance that AI weight.

[20:35] [RC] Fantastic. Well, Irena, this has been most informative. Thank you so much for your first appearance on The Art of Boring.

[20:41] [IP] Thank you for having me, Rob. This has been fun.

[20:43] [RC] Hi, everyone, Rob here again. To subscribe to The Art of Boring podcast, go to [Mawer.com](https://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:

Microsoft
Meta
Google
Amazon
NVIDIA
OpenAI
Anthropic