The heat, the liquid, and the people powering every AI answer (Ep. 60)

Weston Morris [00:07]
Welcome to possibly the final episode of the Digital Workplace Deep Dive podcast, at least with me, Weston Morris, as your host. Yes, I am retiring, leaving the technology area to focus on volunteer work with my wife. That doesn't mean my curiosity and my desire to learn about technology is ending. Absolutely not. I will continue to do that, and I really look forward to seeing how this podcast, or one like it, evolves in the future.

Weston Morris [00:36]
For the final episode, I thought, what is it I really am curious about? We are inundated with AI. We're on the user side of it, right? We're seeing how we consume it. But what goes on behind the scenes? When I make a query or a prompt and it goes off into the cloud, what is going on, and what challenges is the industry facing in that space? Well, I'm thrilled to be able to have as my guest today Vivek Swaminathan, director of solution management, focusing on this exact topic: what's going on in the data centers behind the scenes with AI. Welcome, Vivek.

Vivek Swaminathan [01:14]
Thanks, Weston. Happy to be here. I'm absolutely excited for this podcast.

Weston Morris [01:20]
I noticed the background. You are not in your home office, like normally when you and I chat. Where are you?

Vivek Swaminathan [01:26]
I'm in our collaboration hub at Bangalore today, and I'm doing this podcast from there. I was also last week at our Hyderabad center with a couple of technicians doing training on liquid cooling.

Weston Morris [01:44]
How appropriate. That's why I've had such a hard time getting a hold of you. You've been heads down on this. Well, that's exactly the topic I'd like to dig into today.

Weston Morris [01:52]
In preparation for this podcast, we've all of course been using AI, but I said, hey, for the podcast, I asked AI a simple question. What are the top three concerns that enterprises are having with rolling out AI, and what is going on behind the scenes? It gave me a very nice answer, but it got me thinking: what actually happened with that prompt? It felt almost magical.

Vivek Swaminathan [02:19]
Exactly. And that's the illusion. The moment you hit enter, there's a whole chain of events that starts that most people never see. Electricity moves, chips start lighting up, cooling systems engage, data travels from point A to point B. And it's the people who maintain that infrastructure who make this answer possible. Today, I want us to follow that answer from the moment you ask that question to the moment it appears on the screen, and uncover the industry that's hidden behind every AI response.

Weston Morris [03:00]
That's exactly what I want to do, Vivek. Let's follow it. So I've asked the question. I built the prompt. What happens next?

Vivek Swaminathan [03:07]
Your question gets broken down into small units called tokens. They are little pieces of language that computers can process. These tokens travel across networks to a data center somewhere in the world. It might be a few miles away. It might be on another continent. Inside that building are thousands of GPUs working together. The entire job is to predict what token should come next, over and over again, until the right answer emerges.

Weston Morris [03:42]
So this is something that's happening all day.

Vivek Swaminathan [03:45]
Absolutely. Millions of times every minute. Every AI-generated email, every image, every meeting transcript that is generated. What feels instantaneous to us is actually one of the largest distributed computing systems that humanity has ever built.

Weston Morris [04:06]
You mentioned the tokens, and that explains probably why everyone's so eagerly focusing on token costs, but we won't go down that path for the moment. So what's the catch? What's the challenge we're facing as this enters into a data center?

Vivek Swaminathan [04:20]
We go back to school, right? It's physics. Every calculation consumes power, and nearly all of that power becomes heat. The computer doesn't just generate answers. It creates heat while generating those answers. We've spent centuries using machines to do physical work for us. Now we're asking machines to do cognitive work. We are asking machines to think for us. We've outsourced the thinking to the machines. The catch, to your point, is that every answer arrives with a small amount of unwanted heat and a surprisingly large electricity bill, if I may say so.

Weston Morris [05:07]
So AI isn't just a computing story.

Vivek Swaminathan [05:11]
Exactly. It's a heat story. Everyone is focused on making AI smarter. The industry is now discovering that keeping AI cool may be just as important.

Weston Morris [05:23]
I know we're recording this podcast with video as well, but a lot of people listen to it with audio. Maybe give us a way to picture what's going on here.

Vivek Swaminathan [05:36]
Absolutely. Imagine a dollar's worth of electricity arriving at a data center. The goal is simple: to turn as much of every dollar of electricity as possible into an answer. But some of that money never reaches the GPUs. That is spent in moving the heat, in running the pumps, running fans, running chillers. The industry measures this through something called PUE, power usage effectiveness. A perfect facility in an ideal scenario would have a PUE of 1, meaning every watt of electricity goes to computing. As you can imagine, there's no data center that's perfect. Some of that energy always goes into cooling pumps, fans and supporting the infrastructure.

Weston Morris [06:33]
A lot of times when we think about data centers, regardless of whether it's AI or anything, we're focusing on the chips, the CPUs, GPUs, making faster chips. All the focus is on that technology, but I'm seeing that there's a whole other focus here. It's more than just building faster chips.

Vivek Swaminathan [06:52]
Exactly. It's how to figure out how much of every energy dollar becomes intelligence and how much gets spent dissipating the heat. That's what it's all about.

Weston Morris [07:04]
I'm looking at a news article here that came up just a week ago about a data center failing, and it seems to be relating to cooling problems. It wasn't that the chips weren't doing their jobs in the CPU and the GPU. It was the cooling. I go back 40 years in this industry. I remember coming into raised floors and air conditioning and cooling. It seems like we've been solving this problem for decades. What's new here?

Vivek Swaminathan [07:34]
The example that you're talking about was an organization that suffered a major outage. There were severe storms in that area, and they shut down the affected systems. The key lesson here is that redundancy is not just about server infrastructure. Cooling infrastructure can also become a single point of failure, because once the temperatures exceed safe operating limits, even a server needs to be shut down to avoid hardware damage. There is so much heat generated by the GPUs in these servers. That's exactly what happened in that example.

Weston Morris [08:20]
So is there a difference in, I'll say, the density of heat that we're talking about today? Is that playing a factor?

Vivek Swaminathan [08:29]
Absolutely. If you look at the underlying physics, it hasn't changed. But AI has really outpaced all of this innovation. Every watt going into these systems eventually becomes heat. The difference is that we're packaging more and more computing power into a smaller space. The heat is not new. The heat is always there. But the concentration, the density of heat, is what is new.

Weston Morris [09:00]
Thinking about us as humans, right? We sweat, and water is incredibly effective at carrying away heat as it evaporates. What is it, 240 to 1 calories or something like that? I keep hearing about direct-to-chip cooling. I'm thinking that's using the same principle. Maybe I'm oversimplifying it. Is that what's going on?

Vivek Swaminathan [09:22]
Absolutely, 100%. Humans use the same principle. But if you compare this to a Formula One example, it's like trying to cool a Formula One engine with a household desk fan. That's when liquid cooling becomes necessary, not because liquid creates less heat, but because it's more efficient at carrying that heat away from the point of generation. That is why we sweat, that is why humans use the same principle, and water is very effective at carrying the heat away. That is what direct-to-chip cooling uses from a principle perspective.

Weston Morris [10:00]
That makes sense. If the heat density is increased, we've maxed out how much heat air can pull away effectively. So now we're moving towards liquid. So we remove the heat from the chips using liquid. Problem solved.

Vivek Swaminathan [10:17]
Not really, because you need somewhere for the heat to go. That's where the location of the data center matters so much. In colder regions, like the Nordics and parts of Canada, you can reject that heat into the atmosphere through dry coolers. The most efficient facilities achieve a power usage effectiveness value of approximately 1.1, somewhere around that, because nature does most of the cooling work.

Weston Morris [10:50]
Now, that's probably the easy version, where you have the ability to deploy this in a colocation. You were just talking about your temperature there in Bangalore, and of course here in Michigan, very similar weather today. I don't think we can take advantage of that year-round cold. What other options are available for removing heat and trying to be efficient with it?

Vivek Swaminathan [11:13]
That's a great question, because the interesting version is when this heat becomes useful, when the heat is reused. Some operators in Europe are feeding that waste into the grid

heating systems that warm offices, homes and public buildings. So instead of treating heat as a problem, they're treating it as a resource. I think that's probably one of the greatest outcomes of what we're discussing today.

Weston Morris [11:44]
I love that. Taking that heat and using it for heating homes where they need it in a cold climate. It reminds me, I recall seeing, this was a few years back, a similar sort of thought process in the Netherlands, where a fitness gym was using the mats that are pressure sensitive to produce electricity. The people running and jumping and everything like that produced electricity that was keeping the lights going and the fans and the air circulation in the building, like for free. I love that sort of thinking.

Weston Morris [12:18]
Now, that's not going to be possible with the example you were giving about building a data center in a cold climate. What's going on in hotter parts of the world?

Vivek Swaminathan [12:29]
In hotter parts of the world, components like chillers, cooling plants and methods of active heat removal become very critical. Compute may be identical, but the economics of removing heat are not identical. That's where all these other methodologies come into the picture.

Weston Morris [12:52]
I guess I expected we might be talking about cooling, but maybe not to this extent and how important it is. I think back to the air-cooled data centers. When the cooling went down, stuff would stop, but I don't remember it being like a meltdown, like a nuclear center or something like that. How is that different? How serious is the need to have proper cooling in an AI data center? That's really what I'm asking.

Vivek Swaminathan [13:21]
Great question, Weston. In a liquid-cooled AI cluster, the system is designed assuming that the coolant is carrying away heat from the CPUs and the GPUs. That's exactly what I was doing last week in Hyderabad with the technicians, working on systems that evaluate the efficacy of the liquid cooling system. If you lose that coolant flow, temperatures can rise really fast. The performance throttles, systems protect themselves, and workloads may shut down. It's not like losing air conditioning in your office. It's more like losing a

radiator in a Formula One car that's traveling 200 miles an hour. Everything's okay until it isn't.

Weston Morris [14:21]
I would like to dig into some of the human side of things. You were busy looking at the training there in Hyderabad last week. Before we do, if I could just put you on the spot a little bit and talk about the technology of cooling. I think we all get what air cooling is like. We have air conditioners. It's just like taking what we have in the office and putting it in a data center, maybe directing the air and measuring the temperature and making sure it stays cool. What are some of the technologies involved with liquid cooling? It's not going to flood the data center with water. What's going on in there to actually provide liquid cooling?

Vivek Swaminathan [14:59]
Liquid cooling generally is of a few types. You have direct-to-chip, which is quite popular these days. Then there is immersion cooling, in which you immerse the server rack inside a dielectric fluid. And there is the rear-door heat exchanger. Generally there are these three types of cooling.

Vivek Swaminathan [15:23]
The rear-door heat exchanger you can compare to a traditional hot aisle or cold aisle, which was traditionally there. But immersion is obviously new, and direct-to-chip is also new, wherein the coolant flows on top of the GPU, which generates the heat, carries that heat away and dissipates it out into the atmosphere, like we spoke about. So there is liquid that is introduced in the data center, which traditionally wasn't there, and we were using air. Hence the skill and the ability to be able to troubleshoot and work on these systems becomes more critical than ever.

Weston Morris [16:11]
You mentioned three there. I'm trying to picture them, so I'll go from the most traditional. One is there's air that is then transferring heat to cool, so it's a hybrid sort of thing. Then you mentioned direct-to-chip. I'm not quite sure I understand what that means. And then it sounds like, I was joking saying you flood a data center with water, but you've got this dielectric that the servers are actually in, and they're in this pool of liquid. Can you explain a little bit more about those two? What does direct-to-chip mean? Where's the liquid, how's it being moved and how's the heat transferred?

Vivek Swaminathan [16:46]
Without getting too technical, there is a GPU, that is the processor. On top of the GPU, there is generally something called a cold plate, and the coolant, which is propylene glycol, PG25, flows over the cold plate and carries the heat away. The ability to carry that heat away depends on the flow rate of that coolant, which is controlled by the heart of the liquid cooling system, which is called the CDU, the coolant distribution unit. Think of it as a smart pump.

Vivek Swaminathan [17:31]
The coolant distribution unit is responsible for pumping all of the fluid across the server racks, which enters the servers through manifolds behind the server. They enter through one loop and come out from another loop. That's what they're typically called, the primary and the secondary loop. Once that is done, it is routed out to dry coolers or chillers, like we spoke about previously.

Vivek Swaminathan [17:58]
That is called direct-to-chip. It's gaining a lot of traction these days, because there are organizations coming up with innovative technologies to be able to retrofit data centers with such technology, which makes it extremely effective, especially when you talk about workloads more than 40 or 50 kilowatts. Up to a certain threshold, you can use air cooling. But once you cross a certain threshold, liquid becomes absolutely necessary.

Weston Morris [18:36]
So in that direct-to-chip, I'm seeing the chip is a solid, a metal of a kind, and there's a metal transfer, and inside of that is liquid. The liquid's very close, and being pumped and managed and monitored, and you're checking the temperature all throughout the environment with that CDU distributing it.

Weston Morris [18:56]
Now, immersion. The chips are surrounded, everything is sitting in this. Talk to me about that a little bit. That just sounds crazy. It seems like it's all going to short out.

Vivek Swaminathan [19:08]
Exactly, and it's so funny, because a few years ago people used to dread water in a data center. But now we're talking about liquid cooling. Immersion is a very specific type of liquid cooling wherein, like I said, the server is immersed in the dielectric. But there is also a refactoring of a server that happens before it is inserted in the dielectric liquid. There is a

change in the BIOS that needs to be done so that the system doesn't shut off when water touches the components. There is a lot of work that needs to be done before that server is immersed in the immersion tank, as they call it.

Weston Morris [19:59]
And it's not just regular water. It sounds like you said there's a dielectric liquid that's being used there.

Vivek Swaminathan [20:06]
One of the most common coolants used in the industry is PG25, which is propylene glycol, and a mixture of deionized water as well. That's what we generally use.

Weston Morris [20:23]
Not the same, but kind of similar to what we put into our car radiators. Yeah. So that's it. Thanks for explaining that. That's helping me a little better. Now I'm getting a picture of something quite different in terms of the skills that are needed.

Weston Morris [20:38]
If we want to get to the human side of things: you were just in Hyderabad training several people. In the past, when we're talking about setting up data centers, it's very much about rack and stack, connecting, testing out the network connectivity and things like that. I'm hearing electricians, plumbers, I don't know, carpenters, a whole different set of skills. Talk to me about that.

Vivek Swaminathan [20:58]
This is the favorite part of my podcast, Weston, because every keynote talks about chips, models, frontier models, and very few talk about the people, the technician who installs it, the engineer who brings it to life. You need a field services expert to answer a call when something breaks at 3 a.m. All of these environments require a very new and complex skill set: part electrical, part mechanical, networking and part IT. That's what we saw last week when we were there. Also, during our Unisys Innovation Program here, Patrycja did a fireside chat with the CEO of that organization, and that was extremely well received.

Vivek Swaminathan [21:53]
The global demand for those skills is growing more than the workforce that is available to cater to it.

Weston Morris [22:04]
That's kind of an aha moment I'm having here. What's in the news, of course, is the GPUs. You hear about Nvidia and other manufacturers. It's about building a faster chip to be able to do more AI in a more dense thing. But it sounds like that's maybe not the real bottleneck in advancing AI.

Vivek Swaminathan [22:24]
It may be people. What the AI industry is discovering is that technology only scales when skilled people can deploy and maintain it consistently. That is something we have always known as a part of field services traditionally as well. And that is what people are appreciating and realizing more and more as we go.

Weston Morris [22:54]
That's explaining why you've been making such a big investment in training the people to be able to support this cooling infrastructure. The chips are great, but they're not going to work at all without that in place. I get it.

Weston Morris [23:06]
Well, let's go back to the very beginning. We talked about that prompt that I submitted. We said it all seems magic. A lot of times I'm paying for a few tokens, whatever, kind of seems free. Walk me through that again, end to end, please.

Vivek Swaminathan [23:23]
Absolutely. You remember the answer that appeared on your screen in a second. It traveled through networks. It was processed by thousands of high-performance chips, and those chips generated heat. There were cooling systems that carried that heat away, infrastructure teams that managed and deployed in that facility, field services engineers who install and maintain that equipment. All of that so that you could receive your paragraph of text in maybe under a few seconds. That's the hidden industry behind that answer.

Weston Morris [24:07]
Gotcha. So it's certainly not free, is it?

Vivek Swaminathan [24:11]
Absolutely. It's not free, it's invisible. That's the real story of AI, because the future of AI won't be determined by who builds the smartest model. It will also depend on who can power it, cool it and maintain it at scale.

Weston Morris [24:33]
Wow. That's something. I'm coming to the conclusion here that whereas humans have built the machine, it's also humans that are going to keep it alive. That's not disappearing.

Weston Morris [24:44]
Well, Vivek Swaminathan, thank you so much for carving out a bit of time for us to better understand what's going on behind the scenes with AI. I know your passions are coming through here. If people want to reach out to you, what do you suggest?

Vivek Swaminathan [25:00]
Absolutely, Weston. I'm available on LinkedIn, and they can reach out to me there. No issues.

Weston Morris [25:06]
Very good. And I know you've been prolific with blogs and articles in the media on this topic, so I think if people just search, they can find information on that. Thanks again for joining the podcast today.

Vivek Swaminathan [25:18]
Thanks, Weston.

Weston Morris [25:20]
It was seven years ago that I started this podcast, the Digital Workplace Deep Dive. The topics ranged back then. We were in the middle of the pandemic focusing on employee experience. We talked a lot about employee experience, OCM. More recently, the focus has been on this topic, as we've been discussing here today, Vivek: AI, what's going on behind the scenes with AI and how to take advantage of it in our digital workplace.

Weston Morris [25:45]
I've certainly enjoyed being your host. And the question of what may happen with this podcast in the future? I'll invite you to stay tuned. In the meantime, thank you very much to all of my listeners and subscribers. I've enjoyed serving with you and interacting with you.

This is the Digital Workplace Deep Dive. I'm your host, Weston Morris. All right. There we go, stop recording.