SDS PODCAST EPISODE 1003: BUILDING AN AI DATA CENTER END TO END, WITH LIGHTNING AI'S FRANK BASSO

Jon Krohn:	00:00:00	We've done over a thousand episodes of this show on every layer of the AI stack, except the one that physically runs all of it, the data center. Today we finally fix that. Welcome to episode number 1003 of the SuperDataScience Podcast. I'm your host, Jon Krohn. You are in for a treat with an exceptionally interesting episode today with Frank Basso, who is vice president of infrastructure at Lightning AI, the US-based startup that has over 35,000 modern GPUs, over $500 million in ARR, and that makes it easy to go from AI idea to product lightning fast. In this episode, Frank explains how he builds the physical AI data centers that allow us to do all the mind-blowing things that we do with AI. He digs into GPUs, compute nodes, liquid cooling systems, and much more. Enjoy this special episode. This episode of Super Data Science is made possible by Anthropic, Cisco, Acceldata, and Gurobi.
	00:00:57	Frank, welcome to the SuperDataScience Podcast. How you doing today?
Frank Basso:	00:01:01	I'm doing quite well. Thanks for having me.
Jon Krohn:	00:01:03	It is my pleasure to have you on. We are actually colleagues. We should probably get that out of the way for our listeners. We both work at Lightning AI. I'm a fellow there, which is quite a loose role, but you have a very specific role at Lightning AI. What are you up to there?
Frank Basso:	00:01:20	Oh, wow. Yeah, fellow.That's anything we throw your way to get it done in any way we need, right?
Jon Krohn:	00:01:25	Exactly.
Frank Basso:	00:01:26	I really love those ambiguity within that title. No, I'm Frank Basso. I'm VP of infrastructure. I'm responsible for basically everything that plugs into the wall from the

physical data centers and the DCO teams through network engineering and operations and infrastructure and platform engineering. And

Jon Krohn: 00:01:46 Where are you based?

Frank Basso: 00:01:47 I'm based out of Los Angeles, California.

Jon Krohn: 00:01:49 Los Angeles, the data center capital of the world.

Frank Basso: 00:01:54 If only electricity costs half as much here.

Jon Krohn: 00:01:56 Yes, it would be. And so I'm so excited to do this episode with you, Frank, because while we've done over a thousand episodes of the show, they've all been on a different part of the AI stack. We've had tons of episodes about open source Python libraries. We've had lots of episodes about tools, software tools that make it easier to build and deploy AI models, just like Lightning AI does. And we even have done episodes on GPUs and on chips, but we've never talked about the physical centers that actually run all of those chips. So I think potentially an interesting place to start because this is Lightning AI specialty. What makes an AI data center different from a generic data center for computing?

Frank Basso: 00:02:47 That's a great place to start. I think the differences are density. So traditional data center, even hyperscale data center upwards of five years ago, the maximum power density you'd see Microsoft had the highest at 50 KW racks, which were kind of insanity, but the industry standard was 18s and 18 kilowatts. That's one server now. And so that's a huge differentiator. Also, the physical data centers themselves weren't built to handle the densities of not just power and cooling, but in the space itself, raised floor environments are no longer the case because these cabinets weigh upwards of two tons now versus they were eight or 900 pounds before, now they weigh 4,200 pounds. And so this has presented a lot

of design challenges and constraints when it comes what we knew five years ago versus what we're building for today and into the future. In the future, it becomes even more exciting.

Jon Krohn: 00:03:52 Yeah. Even more power consumption in the future. I understand from the research that we did for this episode, which is actually kind of an interesting thing. I was going to ask you this later on, but just kind of seems to fit in nicely here with you talking about the power consumption. How do you build a data center that today say needs to run H100s when you know that more B200s are coming and the next generation after that is going to be even more power dumps?

Frank Basso: 00:04:17 Well, you run into some physics constraints with these designs and that's what many are solving for. Different companies are solving for them in different ways. NVIDIA's way is to basically co-locate more equipment together, meaning the higher density. Right now we went from eight GPUs in a stack to liquid cold model where you have 72 or 144, 288 and on and on that kind of doubling within the same cabinet or what is a new factor cabinet footprint and getting all the way up to half a megawatt of power. And currently we've recently deployed GB300s, a 10,000 chip superpod in our Chicago data center that's a liquid to chip cooled solution. It's a 10,000 square foot room with 20 megawatts of power running in it, 20 megawatts of power. That used to be more than an entire 100 or 200,000 square foot data center. Now it's in 10,000 square feet in a room.

00:05:20 Of course, there's 50,000 square feet of supporting infrastructure like chillers and cooling pumps and UPSs and generator lineups and all of those things that you need to support the heart of that building, which is that room.

Jon Krohn: 00:05:33 Frank, that sounds like a ton of different variables to think about when you're getting a center together. My understanding is that today in 2026, Lightning AI does something called co-location when it's provisioning the infrastructure for all of its GPUs. And I don't think I've mentioned on air yet that this is a lot of GPUs. So Lightning has over 35,000 GPUs. And so when Lightning's building a new AI data center, they use something called co-location. Tell us about that process and how it works.

Frank Basso: 00:06:04 Yeah, that's a great question. And just so you know, we're always expanding. So by the end of this year, it'll be more like 50,000, which is kind of the growth curve that we're going through at Lightning now. But co-location is same as you, slightly different than you'd seen in the past. It used to be, say, "Hey, I need a cabinet of gear." In a data center, you'd go and lease it out with a small amount of power and you'd install your kit and you'd get your providers lined up and you'd be on the air. That was great if you were to co-locate. Co-location at this scale is more hyperscale. When you start talking about a minimum of 10 megawatts of power and upwards, it's a slightly different conversation. We have a number of data center partners that we work with and that list is getting longer by every day almost.

00:06:59 And basically we work with them and say, "Hey, what's the available grid utility power that they have available to the site?" And whether that's a new build or a brownfield retrofit of an existing facility and how much IT load can they support in the building? And if it's an existing center, there may be some fixed constraints that this is what we have. We have 10 megawatts and that's what we have and there's no room for expandability. Or they might say, "Hey, there's another 20 megawatts in the parking lot at the substation." And so then we talk about a build. And so we then take the amount of available power and work with the data center provider and in this case, everything moving forward is liquid to chip because it's

more efficient overall and consumes less overhead and has better what they call PUE or efficiency rating of actual workloads in the building because the cooling is directed to chip.

00:07:55 It's not going to air to the chip and then back to air and then you have to cool the air. You lose a lot of efficiency with that if you're using it in a mechanical way like that. So we work with them on a design and the design of available power will tell us how many chips can we put in there of a certain type, say GP300. We say, "Okay, we want to build in GB300, a serviceable unit is 1,152 GPUs, that's 16 cabinets of gear and we want nine serviceable units in a building that's 10,300 and change GPUs. And then we need on top of the GPUs, well, we need all the networking and networking cabinets and we need the storage and the storage subsystems and we need all the internet routing and all the compute, all the CPU to support because GPUs don't run on their own.

00:08:49 You need lots of VMs to do all kinds of things I'm sure you talked about on this bad podcast at length. And so there's a ratio and that ratio is increasing for CPU. So there's a lot of other things you have to put into your design constraints. Then you have to do load calculations based on what is our usage profile going to look like for this gear because you have what's called a plate rating. Say the plate rating is 22 megawatts based on what it says on the side of the box, you're going to draw 22 megawatts of power, but the reality is that plate rating of the box means you have the maximum amount of CPU or the maximum amount of memory installed and the maximum amount of drives for our configuration with a lower amount of drives or a lower amount of memory per box or what the basically design for the customer, there's no way that you could actually draw that much power.

00:09:42 So we apply a D-rating. We do a safety rating at 81%. It's no industry secret. Most companies use 70 to 75%, which

I think is a little bare, but we do a conservative number at 81%. So we come out at instead of being whatever it is, we come out at under 20 and we know that power envelop is what's available in the building is 20 megawatts even though the plate rating or the delivery and the infrastructure's designed for more, it actually will never exceed 20 megawatts and that's what we're actually going to be paying for and that's what we contract for. And so after that point, we have a design and the data center provider says, okay, great, this is how much it'll cost per megawatt to build. And that varies a lot depending on where you are in the world and how much things cost or whether you have union labor or things of that sort, hugely variable from one to four million per megawatt right now.

00:10:45 And that's just cost to us. The cost to them is probably six times that, but they're charging us per megawatt per month as a base fee plus our usage from the utility. So the base fee per megawatt is how they make back all their money over the term of a contract and a contract these days is usually 10 years in length.

Jon Krohn: 00:11:03 Wow. It is wild to be planning over those kinds of timeframes. When you think about things like you were talking about how CPU usage is going up and you can correct me with these kinds of like rough finger in the air numbers, but my understanding is that in the pure GenAI era, when we were talking about just the chatbot experience when you're in something like ChatGPT or Claude, that was something like a 12 to one GPU to CPU ratio. But now as we're increasingly in this agentic era, agents are doing more processing and kind of coordinating of tasks that then get sent out to GPUs. And that means we're getting closer to a one-to-one ratio of CPUs to GPUs and that's just over a timeframe of a few years. So it's pretty wild to think how complicated it is to be planning 10 years ahead on what kind of infrastructure, what kinds of power demands you're going

to have in a given center when the industry, when the AI industry is changing so quickly.

Frank Basso: 00:12:02 No, that's true. It is. The good news is we have a little help. NVIDIA produces what they call our RA or reference architectures. Now, NVIDIA reference architectures are a starting point. They're the effectively the minimum viable product or bare minimum you need to do to make this system work in a performant manner. And it includes what your layout should look like, what serviceable units are, what your network should look like, whether you're Infinite Band or Rocky on your east-west GPU interconnection network or on your north-south network and then the internet access you may need for that. They have recommendations. Well, we use this as starting point. A lot of companies build that or build somewhere not near that. It's all over the place, which predictability is becoming a very interesting thing in the consumer side of the market. But we take and we build on top of reference architecture, for instance, we do no oversubscription within our network in any way, shape or form.

00:13:07 And that's common on the east-west network. You don't oversubscribe.

Jon Krohn: 00:13:11 What does that mean, no over-subscription?

Frank Basso: 00:13:13 Oh, over-subscription. This is a very old networking term that means say you have two things connected at 800 gigs of speed. I'm using a modern number and you only have eight from the switch, you only had 800 gigs of Uplink. Well, you're oversubscribed two to one. If it goes into a switch and you have two ports going 800 gigs each, there's two ports that are going up to the next layer of the network to match that. And so we never oversubscribe. This is a bad thing because of contention within the network and also the RAs are mostly designed for single tenant systems where we're a multi-tenant cloud and so

you never know what the customers are going to do. I've heard people in the industry say NeoClouds have it the worst because you never know what your customers are going to be up to versus people who build it for themselves like XAI and OpenAI.

00:14:11 Those guys are like, "We build to support our engineers and we don't have to do anything like NeoClouds." So they do one specific tasking for their designs and it makes it a lot easier, just like hyperscalers did predictable workloads, predictable outcomes. For us, it's the next customer comes in and says, "Well, what do you think about X?" And we go, "Ooh, okay, we can do that. " And sometimes they're like, "Wow, no one said you could do that before." I'm like, "Well, we designed for future headroom and future performance levels that people aren't thinking about when they build. We have to because otherwise we don't future-proof our systems."

Jon Krohn: 00:14:48 Wow, that is a fascinating insight into how you're doing this. For people who aren't aware or who haven't listened to the episode that I did about three months ago with the CEO of Lightning AI, Will Falcon, tell us a little bit about what a NeoCloud is, like what Lightning AI is.

Frank Basso: 00:15:05 NeoCloud meaning unique is the ultimate definition non-hyperscale, even though we build things at hyperscale. So the NeoClouds are being unique or bespoke industry versus the normal Amazon, the Amazon, Google, those type of systems that we provide, we're focused on the GPU, the GPU and the GPU performance and then secondarily the memory bandwidth and performance and then the storage bandwidth and performance because it's insane compared to what a normal cloud provider would have to put up with. We have customers that they may be just doing training, but they're doing reinforcement learning and they're doing a distributed basis and they're pushing almost a terabit of traffic in and out the front door on the internet access

side through our private network interconnects across our backbone. And we've built a substantial network of not just, hey, we're connected to the internet, but we build regional networks within the metro that interconnect our data centers to all the internet exchange points within the region, plus we have a nationwide backbone that connects all of our data centers together so we can shuffle and move data on our client's behalf and also run them as one cohesive system.

00:16:26 I mean, we're effectively running a telecom carrier style network in the North America to tie together all of our sites. Nobody really does that, but maybe since as a recovering network engineer, I knew better. And so when I joined, I went, "Oh, this is going to be a problem." So we started on that a year ago and it's just come to fruition and customers are very happy about the upgrades, but that same customer that's pushing all that traffic is also pummeling the storage arrays at like per GPU a gigabit or gigabyte per GPU per second is not unheard of and any normal storage arrays would just have lots of wait times and hourglasses waiting on those apps in the background and not be performant. But we have to design for those kind of performance constraints. We have lots of demanding clients and with Inference, it becomes even more challenging than with training.

00:17:26 Training is fairly predictive for a long run over a period of time where Inference is very spotty and it's, what do you call it, peaky or bursty

Jon Krohn: 00:17:35 Traffic. Yeah, bursty.

Frank Basso: 00:17:38 And that's a tough one.

Jon Krohn: 00:17:39 When you talk about storage arrays, is that something like when you're training a model, you are updating weights on a whole bunch of different GPUs and you need to store those weight updates somewhere so the

information is getting sent back and forth between the GPUs and the storage array?

Frank Basso: 00:17:56 Correct. So the nodes, it depends really on how you're running it. Every customer does this differently. Some fetch, they've pre-trained somewhere else and they're fetching it in real time chunk by chunk. And when they process it, they just use the local storage for scratch space and when they have a result, they push it back out of the network in real time. Some download all their pre-training and have it locally and then all the nodes pummel the storage array and then make all their updates and when they're done, they batch load it. We never know how they're going to use it. Some customers are more demanding than others, but storage arrays and storage performance along with the CPUs that they're using to interact and load things in and out. The workloads are so random on all the profiles. It's kind of fun to watch to see what customers are up to next, to see if we've built and designed a data center that can work for them and if we need to make any adjustments on the next generation that we build or do a tweak or a refit to an existing site to support those things.

Jon Krohn: 00:18:59 Something that I didn't appreciate before this episode, I guess I had kind of heard conversations around the office that should have tipped me off to this being important, but it sounds absolutely critical that you are listening to each client's needs and actually in a lot of cases it sounds like developing bespoke infrastructure systems for them and their needs. That seems to be essential to being able to deliver the kinds of services the clients are looking for. And it sounds like before something that you said earlier is that before we have these clients come on as prospective clients, you have to be designing the whole system with extra leeway with extra headroom to be able to meet whatever client demands come up.

Frank Basso: 00:19:43 Yes, yes. And that's the challenge that the XAI and OpenAI, my peers over there kind of look at me and say, "Wow, good luck, man." Because it's a much harder thing. I mean, in a lot of ways we're taking a swag at it based on what we know, based on trends coming forward, working with the industry, listening to the industry, talking with NVIDIA about future items and then finding a middle ground that is the best balance between performance and economic costs. These things are not inexpensive to build. GPU data centers are very expensive because of their high density requirements. Liquid cooling systems are extremely expensive to install and operate. They just don't sit there and run on their own. They take a lot of active tuning and to keep everything running at peak efficiency. And same thing with the GPU nodes themselves. GPU nodes are fickle.

00:20:45 They're very specific and prescribed. And so if a customer tries to do something different, it may break an entire cluster by doing that. And then of course we have to have all the guardrails in place with our VPC networking to isolate the customers so that in a multi-tenant environment where they're not sharing GPU nodes, but they're sharing common north-south network fabrics. The east-west fabrics are large cluster fabrics that are segmented up to based on how many nodes that whether a customer has eight nodes or they have 512 nodes isolating them and keeping them segmented for the highest level of security, highest level of performance and other key measurements and KPIs that we measure against. Those design constraints have to be done upfront and they have to be discussed openly upfront. And then there's knobs you turn for the performance of how much you really want to spend because the sky's the limit on how much you can spend here.

Jon Krohn: 00:21:44 For our listeners who aren't aware, Frank, what is a node?

Frank Basso: 00:21:47 A node or a GPU node is effectively the actual box itself, the server that is a GPU server. It's not unlike a traditional X86 or ARM or other compute server, except for it's built and added onto, it's much in an air cooled scenario. It's upwards of 12U tall. So 12 rack units are 1.75 inches each. So it's a very tall box. You can put four in a full size cabinet and it moves a lot of air through. It has eight GPUs in a typical configuration. And besides that, it's a normal server, but it has eight GPUs in there, plus it has eight network cards for one for each GPU, plus it has two other network cards in it for what we call the north-south network, which is your internet and storage layer access networks. The same server that would be 12U with the same configuration comes in a 3U as well, but now it's liquid.

00:22:56 So instead of four per cabinet, I can put eight per cabinet easily and so my density is increased.

Jon Krohn: 00:23:04 Because you need less space for air to move around because you're using the liquid cooling. Correct. And when you talk about liquid cooling, so I've built tiny little GPU boxes for myself and they have been liquid cooled and something that might surprise people, you hear a lot about, and we're going to talk about energy later on, but when you hear liquid cooling, I think people think that there's like you're hooked up like a water hose to the municipal water system and you have this like cool air, sorry cool water running through the surface. But I've had liquid cooled GPUs and tiny little servers that I built myself and the water just stays in the system. It's a fixed amount of liquid. I guess it's not even really, I don't even know if it's water, but the liquid is fixed in there and it just loops around.

00:23:52 It gets heated by the chips and then it gets cooled down in another part. You're not needing to add more liquid on a continuous basis

Frank Basso: 00:24:01 That's correct. And this is one of those large pieces of misinformation based on designs from 20 years ago in data centers that aren't really used anymore. So your home system, you've got the little water reserve and little radiator and you're taking heat away from the chip through that interface. When we say liquid cool, this is not immersion. The chips aren't dunked into a bat of water. It's not like a Bitcoin mining where you submerge the entire system in a water bath or a dielectric fluid bath. Is it non-conducting binary fluid, but this is instead of a heat sink on your chip, the heat sink has a water loop interface to it. So cool water's coming in or what they call PM25 fluid, which is very special fluid that isn't water. It's like a very specialty glycol fluid that you would put in your, like something you put in a race car versus regular car, very lightweight, very highly performant and it flows on what's called an SFN or a secondary fluid network.

00:25:08 So the secondary fluid network is the same as the one that's in the case at your house. This circulates this fluid from the chips itself, comes into the cabinet with these inch and a half to two inch hoses, runs down through manifolds in the cabinet, manifolds have connections or points the servers plug into the manifolds and now they're on the secondary fluid network. The secondary fluid network goes to a heat exchange point. So that's not where it stops. This is the beginning. So instead of saying, "Hey, it's going to vent to your den or your room or your office at home," it actually has another radiator, a heat exchanger that exchanges with the primary water system of the building, the hydronic system for the building and that exchanges the heat at that point. So the secondary fluid network is super clean. It's sub 2.5 micron filtered water so it never plugs up the litle capillaries on the chips or on the cooling plates and things like that inside the box.

00:26:12 There's a number of different technologies in the boxes depending on who makes them. And then it goes to the building primary loop. Okay, now you have cold water coming in at as low as seven degrees C to 20 degrees C, so 45 to 68 degree water coming in from the building. That then goes outside the building and goes into giant heat rejection systems. So this is where the misnomer about water happens. Back in the day, the cooling towers used water. They used water sources to basically spray over these radiators and then blow air across them and you'd get this great Bernuli cooling effect across the surface with the water or other methods of using direct air cooling could be a waterfall of the water coming through and the air is cooling the water. Those techniques are pretty much banned everywhere now. So what it has outside because of water waste and water use, everyone's been very concerned, but the industry started going about 10 years ago to what's called heat rejection.

00:27:22 What it is, is a giant radiator that's in your car, but it's huge or there's 30 of them lined out outside with fans and the fluid in the building, the glycol-based fluid in the building, just like in your car, never gets changed once you fill it up. Once you fill it up, it's full. So you may need to, if you have to fix a broken pipe or change something out or do maintenance, then you may have to top it off, but you're not using millions of gallons per month. You're using no more than a household does in a month and in a giant data center. So you reject, basically it's called heat rejection. You're taking the heat and you're exchanging the heat from the secondary to the primary. And then in the primary, you go through a giant radiator like in your car, but there's bigger ones and you blow air across it and you vent all that heat and you blow that heat upward.

00:28:14 So it blows up and away from the building and blends with the air above the building. And so sometimes even when you drive by data centers, you'll see like on a cold

day, if you're in Chicago or something, you'll be like, "Wow, what's that plume coming off a building?" Same thing off a data center. It's the differential of heat and it condenses and it makes a pretty cloud. But when you do see that, you know you're not using water, you're doing heat rejection. And that's one of the biggest myths today on why people are turning towards not wanting data centers. The NIMBYism, the knot in my backyard, "Hey, you're taking all our water." And we'll talk about power I'm sure in a minute, but the water part's not true anymore. And I think that's just old information that is no longer being done, but it keeps coming up in the news.

Jon Krohn: 00:29:01 Yeah, we'll get to that power situation shortly, but almost like 95% of what you said there was new information to me. I did not know any of that and it is really interesting. So I hope a lot of listeners have enjoyed that as well and maybe now you can also debunk the concerns of people who saw TikTok about all the water usage that the new data center is going to use in their area. So yeah, that is interesting when you were talking about nodes, I don't know, five minutes or so ago and I asked you about nodes, no, no, not at all. Nothing to apologize for, but one of the things that you were talking about there, and you'd mentioned it earlier in the episode as well, also something that I know nothing about, you were talking about east-west connections and north-south connections, what are those?

00:29:44 And they also, they sound different.

Frank Basso: 00:29:46 That is a good question. Networks within the data centers are broken up into two ... Well, there's actually four networks within the data center itself. The first one being the one that a lot of us focus on is called East West. I So if you have all the GPU boxes lined up in a row, they need to talk to each other. They travel east and west to talk to their adjacent GPU nodes. And this network is very fast. Every single GPU has a dedicated physical interface to it,

but then logically can connect to all the other GPUs. And this is done-

Jon Krohn: 00:30:24 All the other GPUs on the east-west connection?

Frank Basso: 00:30:26 Correct.

00:30:28 So if you have 255 nodes, say a 2K cluster, 2,000 chips in a cluster, then all of them can talk to each other and they can talk to each other at full line rate. So whether it's a 400 gig interface or an 800 gig interface in the newer deployments and soon to be 1.6 terabits per GPU, leaving the box, you have eight of those connections in an eight configured server or if it's a water cooled four GPU server, you have that many connections connecting to all the rest. That way the GPUs can work together, share memory, share bandwidth, and you can cluster the GPUs to load larger models and do other things with them and distribute the work across multiple GPUs together. That networking east-west is very powerful. That makes all the difference. And different manufacturers have different versions of that, but NVIDIA is by far the furthest out and the most performant.

00:31:31 The EastWest traffic is outstanding. And there's two ways to do East-West. You can do it with Infineband or IB. That is what NVIDIA purchased from Mellanox years ago. They're the only manufacturer of this. Others had license to it, but they stopped making it, but NVIDIA has built on it and I think all those old other manufacturers are regretting their decision to stop making it. And then there's Rocky.That's RDMA over ethernet and Converged ethernet, R-O-C-E. And that is using traditional ethernet framing and a traditional ethernet technology to pass GPU traffic across it. This works very well. It scales incredibly well, but we're going to leave that religious argument alone for this discussion because people are seated in one house or the other and it's literally one of

those religious arguments within technology that people are passionate about.

Jon Krohn: 00:32:30 Specifically Infineband versus Rocky.

Frank Basso: 00:32:33 Yes, that is correct.

Jon Krohn: 00:32:34 So

Frank Basso: 00:32:35 That's a big one. Then you have the North South network. The North South network is, it goes up and out and back down. So from the GPUs up to the different layers out through firewalls and border networks and to the internet, that's your north south network. On the inside, it'll pass through a services layer so it can go to get to the CPUs and it can get to the storage networks and those are tied in as well. So in the north-south network, there are complexities about storage on the east-west network, but we're not going to go there.

00:33:10 But traditionally north-south is for that. And then you have another network on top of that, which is your management network. And then you have a true out-of-band network, which is the brake class network and how you manage these things from the outside in. So if there's a problem on the inside, you can work on it and run all your observability and see the node health and the network health and the traffic and the flows and kind of what the customers are getting at based on from looking at it from the outside in. If they're having issues getting to some endpoint, you can look at all of that through your out of band network. So it's truly out of band and then you can kind of break glass and barge in if you need to from the outside to ensure everything stays running. Now, a lot of people don't build that network very totally segmented.

00:33:57 They just have the inband management and then if something breaks, you send someone with a laptop to the

data center, which doesn't work. So the larger you build, the more kind of guardrails you need. And that true out of band management plane is one of those things and expense that a lot of people don't necessarily buy in on, but it's critical for lights out operations of these site locations.

Jon Krohn: 00:34:21 Right. Yeah. Lights out operation, meaning this was a new term that I learned just in the past few days as I was doing research for this episode. Lights out means that you don't need to have any lights on in the center because there are no humans in there. It's a fully autonomous system.

Frank Basso: 00:34:35 That's the idea at least. That's the notion. I've built lights out systems for years with modular data centers and other things. Think about GPUs are they get run so hard that they do break. There's a certain percentage of failure. Some models have more failures than others and the newer chips are just so much better. The liquid to chip cooling has really pushed that number down. You don't see a lot of failures within the LTC chip sets, which are great because you can adjust your staffing models in the data centers. We're staff twenty four seven around the clock at our data centers with our own teams to ensure that we can provide that Four Seasons white glove experience for our customers, whether they're on Bare Metal or MKS or anything else that their nodes are always available and always running.

Jon Krohn: 00:35:28 What does MKS mean?

Frank Basso: 00:35:30 Manage Kubernetes. So more than just, "Hey, we run Kubernetes." No, we run a full managed Kubernetes suite. So full platform as a service, control plane management provisioning, slurm on top if you want it, other things. But it's more than just, "Hey, here you go. Kick it over the fence and you can run Kubernetes on

your own. Good luck." No, we have a full managed Kubernetes suite and a very robust platform as a service.

Jon Krohn: 00:36:00 So basically that means that a client of Lightning can come with their Kubernetes, whatever they want to be scaling up with Kubernetes and they can just have that configured already packaged up and then they can bring it to this MKS managed Kubernetes service and have it scale up easily.

Frank Basso: 00:36:18 Yeah. They can effectively plug into our system directly and it just makes it easier for them. We have customers that range and skillsets all the way from the 10 scale, the smartest, brightest, craziest idea AI native companies that are just doing the things you read about all the way to down to the enterprises which aren't doing those foundational and frontier things that want to be users of the system and they want an easy button, they want assurances, they want more security, they want more of the enterprise features and we maintain all of those things as well for them. And we have an easy button through our provisioning and management systems to do that. They don't know anything about bare metal. They don't know anything about north, south, east, west storage networks, internet access. They don't and they don't care and they shouldn't have to. They want to consume it like they would AWS or GCP.

00:37:21 They want to click and deploy and we have those options as well. Then we want some folks who, "Hey, I want to change the bias settings on the boxes for this esoteric performance thing that I'm trying to obtain. Can you test that with me? " Sure, of course we can. It's a different customer and it's a different understanding and the amount of customers that know those things in the hardware space are becoming less and less and they're relying on us more and more to make those things happen.

Jon Krohn:	00:37:50	That's really cool. It's nice to know the lightning does that. Another question that came to me as we've been talking about this that is pro, I think I know what the answer is, but it'd be nice to hear you say it. When you talk about all this east, west, north, south, obviously that is completely independent of like magnetic north, right? It's not like you need to set up your data center so that all of the roads are like east, west.
Frank Basso:	00:38:13	No, it's Feng Shui. No, I'm just kidding. No, you could line the cabinets up at a 45 degree angle in the room if that was your groove. And that's actually a joke I use with the providers. I said, "If the pipes are running at a 45 degree angle under the floor, I'm not going to make you move them. I'm going to line my cabinets up to them so we can value engineer the deployments as much as possible." But no, it's a logical assignment of east, west, north, south. That's a good question.
Jon Krohn:	00:38:47	Yeah. So at some point people just decided to use that convention because obviously it could have ended up being the other way, presumably the nomenclature, like everything that you're talking about is east-west, that could have been the north-south. I don't know if that makes sense.
Frank Basso:	00:39:00	Well, east-west, meaning you're looking left to right. So you're looking at things that
Jon Krohn:	00:39:05	You
Frank Basso:	00:39:05	Had a row of people or row GPUs, you're looking left, you're looking right, you're looking east
Jon Krohn:	00:39:08	And west
Frank Basso:	00:39:09	Versus north and south,
Jon Krohn:	00:39:11	You're looking up,

Frank Basso: 00:39:13 Looking down. I

Jon Krohn: 00:39:14 See. Yeah, that makes a lot of sense. And then so digging into this a little bit more, when people have seen pictures online of any kinds of data centers, like we have been able to see online for decades, probably most of us haven't actually been to a data center. So when you see those photos, you see these long hallways that the servers make up. And so when you're talking about east-west, it's like you're standing looking at one of those server racks and you're examining it from the left to the right. That's kind of like looking west to east regardless of where your compass would actually be pointing.

Frank Basso: 00:39:48 Yeah, 100%. And then the east-west is also limited into what we call serviceable units. So there's a logical distance on the cables that you have to maintain. They don't go very far. They go between 50 and 500 meters and that's it. So they can't go down the street, can't go across to another data hall. So data halls have to be contained and the serviceable units are there. Now how you connect all those serviceable units together, that is some magic, but there's limits and scaling factors that each manufacturer has for the design and constraints of their east-west connectivity within the actual structure. I

Jon Krohn: 00:40:30 See. So basically based on that kind of 500 meter cable limit, if you build a big enough data center, you could potentially have a bunch of different ... What was the term you used there for kind of like a-

Frank Basso: 00:40:43 Serviceable units.

Jon Krohn: 00:40:45 Well, yeah, there was another term though where if I could potentially have that east-west, I'd reach my east-west maximum for one data hall. That was the word that I was looking for, data hall. And then so you could have then another data hall to my east and another data hall to my west. You could have as many as you wanted

basically. And then those would be connected not by east-west connections, but by the north-south connections if they needed to be

Frank Basso: 00:41:11 Connected. No, actually. It's more complicated than that.

00:41:14 So within the east-west fabric that connects it together, there are distance limits. The limits are actually between the GPU nodes for timing and the pico second level. So from GPU and one to GPU, say 255, they can't be more than 50, 100, 500 meters depending on the type of equipment you use apart. So if your data center's the size of a Costco or a large factory store like that, if you had one at each end of the other, it would be too far. So they have to be grouped together physically and the data halls are now designed to group for those groupings when you design and lay them out. Now, you can have one data hall connected to another by going up a tier within east west. So you have your leafs that connect your local connections, just like traditional networking, then you have your spines that connect all of those leafs together and then you can have a third tier or superspines, which could allow you to go between data halls.

00:42:18 Now when you're doing training, that kind of literally adding X picoseconds to go to the next room, it might break your training. The timing may be off, it may mess you up. So while they're all connected so you could share data, certain types of workloads just will function badly unless you tune them very specifically for that data center and that topology and that design. So working with the clients to let them understand what the underlying design is and sharing it with them openly is super important. Otherwise, they'll make an assumption that they were using at another provider and they're like, "Hey, why does this work?" Or, "Why does this work so much better?" It's like, "Oh, well, we'll share with you and show you why."

Jon Krohn: 00:43:01 Wow. So much to think about there. And even in my simple minded view of how these data halls could connect, really appreciate you elucidating for me how these spines work and superspines that is great to know. So I think probably for now I'm going to take a pause on talking about these physical centers, unless you think there's something interesting that our audience needs to hear about just what it's like to be in one of these AI data centers. A lot of us can probably picture a photo that we've seen online of these long hallways of server racks, but is there anything else that's kind of interesting, maybe particularly interesting about an AI data center when you're physically standing there?

Frank Basso: 00:43:47 When you're physically in there, one of the differentiators from a traditional data center is the noise level. These systems are very noisy. We call them screaming banshes.

Jon Krohn: 00:43:58 Oh my God, I had no idea.

Frank Basso: 00:44:01 Yeah. So especially within air cooled and inside the data hall, the levels range from way beyond what you'd hear at a rock concert if you're in the front row. And so hearing protection for our staff, require two types of hearing protection at all times. You have both the buds that go in your ear, you can use molded ones or not listen to your music or whatever. So something occlusional in your inner ear and then cans, right? And cans being all passive, you cannot wear or use active noise canceling systems within a data hall. This is a big thing that people are like, "Yeah, put my noise canceling on. It's great." Well, if the data hall is 105 to 110 decibels, to cancel the noise, noise canceling generates 105 to 110 decibels. So that does just as much damage. You're not hearing it, but it's damaging your eardrum.

Jon Krohn: 00:45:00 Wow. I had no idea about that. It makes so much sense now that you say it, but I had no idea that with my noise canceling headphones, I'm here thinking, I'm wearing

noise canceling headphones right now. Obviously, it's not canceling a hundred decibels of noise. I don't have screaming banshees in my recording studio, believe it or not, but actually that's an interesting take-home tip. For anybody going to a rock concert or whatever, you need to have passive noise, not canceling, but just suppression.

Frank Basso: 00:45:28 Yeah, occlusional or suppressive. And so really it blocks the two different kinds of hearing protection, the inner ear hearing protection and then the outer ear blocks different frequencies of noise as well. So the frequencies of noise that cause your hearing damage, the higher frequency noise from the fans and the motors and the power supplies that are humming that you can't really hear to your native ear, they're present. And so you need to block all those out for safety. We take that very seriously at all of our locations and we actually have OSHA sound studies done and we maintain OSHA compliance and everyone has to get trained to be in the data center. Even visitors, we warn visitors when they're coming like, "Hey, this is a very loud environment." And what's funny is that the occlusional blocks out so many things, but if I talk loud enough, like that loud grandma talking at you because she can't hear anymore to someone in the data center with a hearing protection on, the frequency of my voice comes clearly through, but you don't hear any of the high frequency noises or things that can damage your hearing.

00:46:34 So you might think that, "How does anyone work with somebody else?" Well, there's a couple ways. One, we have some systems that are intercom based like racing radio style that you can talk to each other with and we also have hand signals that we use in the data center and things like that. There's a bunch of different combinations depending on which location we're at, how loud it is. And you think, "Oh, the liquid cooled ones don't have as many fans. They're just as loud." They have rear door heat exchangers. They have other cooling systems

and pumps and things running in the room. It is a very loud industrial environment. The liquid to chip

00:47:09 Data centers are more industrial if that makes sense. The traditional data centers that are pretty and they're really nice, they raise floors and they're super orderly and things like that. You look at some of the pictures you see online on liquidity chip centers and there's hoses and pipes and cables and everywhere all of ours that currently are fed from above. So there's 20 inch water mains running through the room that are insulated so they don't make water or sweat because the temperature differential with the room. There's hoses to every machine. It's just if you look up, you'd be like, "Oh my gosh, what is all this stuff in here?" Well, that's how the sausage is made. It's very industrial. It's like you're in the reactor room on a submarine or something.

Jon Krohn: 00:47:56 It's pretty cool. Do you think that there is a higher rate of sign language fluency among data center workers relative to the general population?

Frank Basso: 00:48:09 That'd be an interesting question. I don't know. I don't know, but you think there should be at some point.That's actually a really good idea even though I'm sure they use their home version of sign language, if you know what I

Jon Krohn: 00:48:21 Mean. Right, right, right. Well, yeah, that's interesting. I guess it's potentially a good career choice for people with a hearing problem.

Frank Basso: 00:48:28 Oh yeah. There you go.

Jon Krohn: 00:48:32 And so for people, Frank used the word OSHA, which people in the US probably will everyone will know what that means, but if you're outside the US, it means occupational safety and health administration as a federal body that is keeping workers safe in all kinds of industries. Quick question for you with these data centers

being so large, do you just always get around on foot or do some people use pedals or motorized vehicles ever?

Frank Basso: 00:48:58 It depends on how good your insurance is, I think. A lot of data centers, some of these now have scooters and bicycles just sitting all over the place. I'd say don't wear Heli's because you need to wear actually proper work boots in these locations because things are heavy and if something were a drop in your foot, that would be bad. So you need to wear a proper work attire.

Jon Krohn: 00:49:24 Do you wear hard hats?

Frank Basso: 00:49:25 During construction phase, they wear personal protective equipment. So hard hats and vests and ceramic towed boots and non-flammable things during constructability and provisioning. And then once it's online, the data center technicians aren't required to wear that, but they're for hearing protection or if they're working in a cabinet safety glasses and then of course they need proper ESD projection. We issue ESD shirts for our teams. So they're wearing a shirt that's not polyester that won't spark every time they touch a cabinet kind of standard issued uniform stuff that we've been working through.

Jon Krohn: 00:50:08 So ESD is like electrosensitivity something?

Frank Basso: 00:50:12 Yes, electrostatic discharge.

Jon Krohn: 00:50:14 Electrostatic discharge.

Frank Basso: 00:50:17 Yeah. And you have to wear a wristband if you ever touch or open a box. So you put

Jon Krohn: 00:50:21 It on

Frank Basso: 00:50:21 And then the cabinets literally have like these little lightning bolt and plugs on either side, every single cabinet, front and rear, and you plug yourself into those.

So you're now connected to the cabinet because with all the air flowing through the systems, they can generate static electricity. So it's for safety of the worker and for safety of the gear.

Jon Krohn: 00:50:41 Wow, that is so cool. Thank you. I didn't anticipate talking about these kinds of physical things like sound and scooters and stuff, but it just kind of occurred to me as we were talking about this more and more. Let's go back before we wrap up this episode. There is something that you mentioned earlier that you said we would discuss it later on. So let's make sure we get to that, which is the electricity issue. You mentioned NIMBYs earlier, you hear a lot of that there's a lot of NIMBYism all over the world. You and I are both in the US. We hear it particularly in the US press around concerns of electricity usage in a given region where lots of data centers are coming online. Now I read The Economist every week and the economist has a number of times in the past year done articles on how at least up to this point, if you are in a region that has an increased electricity bill, it is almost certainly not due to data centers or AI specific data centers being built there.

00:51:43 Yeah. I'd love to hear your thoughts on this issue. People probably, I don't know if you're at cocktail parties and are like, "Frank, I can't believe what you're doing."

Frank Basso: 00:51:52 Well, yeah, I get that a lot like, "Oh, you built those things?" I'm like, "What do you mean by that? " And everyone thinks that- You're one of those people. Those people, you AI people, I've been building data centers a long time and this isn't the first time this has come up. Everyone thinks that this is why, "Oh, my power bill went up." Well, it has nothing to do with anything. It has to do with so many different public utility commissions, decisions and taxations and other things that drive the cost of electricity to a residence. I live in California. The

power is very expensive here. It's pretty much the highest in the state, but-

Jon Krohn: 00:52:30 In the country.

Frank Basso: 00:52:32 Well, yeah, sorry.

Jon Krohn: 00:52:33 LA is the highest in the state, and then California is the highest state.

Frank Basso: 00:52:36 Well, no, no, actually it's not. I was going to say, this is kind of the highest in the country, but we have multiple power companies here. So within the same state generating off the same generation, meaning the same natural gas fired plants, we don't have any coal out here. We do have nuclear plants still. We have one remaining, but the power varies from 10 cents per kilowatt hour all the way to 58 cents a kilowatt hour for residential delivery. Why is that? Well, it's because the state has messed it all up.This is state regulations, state changes here in Los Angeles is the second cheapest power in the state. Silicon Valley power in Santa Clara, California is the cheapest power in the state and they have their own gas fired plants within the city, just like LADWP, LA Department of Water and Power does here in Los Angeles.

00:53:28 Then you have two other power companies. You have Southern California, Edison, our wired wildfire specialists, then we have Pacific Gas and Electric. Pacific Gas and Electric and Edison, there's some of the most expensive power in the entire country, whereas LEDWP is the same as it is in Chicago in the Midwest and Silicon Valley power is as cheap as it is in Texas, which is super cost effective or in Florida where there's plenty of inexpensive power. There's plenty of power out there. It's all about regulation. So when a data center goes in, a couple of things happen and let's say you don't have a data center and they come in and your power rates won't go up because the data center itself is going to pay

millions to tens of millions of dollars to the utility to improve the power grid to connect to that site.

00:54:20 They're going to pay for upgrades. They're going to pay for new transmission lines. They're going to pay for upgrades to the generation near you. They're going to pay for more generation or they're going to put gas fired, what they call behind the meter power stations or fuel cells, which are silence at the site. And oh, by the way, those fuel cells, the byproducts, they make water. So they make distilled water coming out of the back of the fuel cells, whether they're natural gas-based or hydrogen-based fuel cells, you're getting water output as that. They capture that water and they use it to keep the building full and topped off or they use it to put it through a filtration system and they water the grass out in front of the data center. They're making water at these locations with new fuel cell technology and that's the way forward.

00:55:05 These gas fire plants are called peaker plants, which you see going in that like, "Hey, Goliath is using these peaker plants." Well, they're burning what's called dirty gas. So the stuff that's coming right from the oil fields and the gas fields and they're burning that, but then they burn it and they filter it and they keep it clean versus like if anyone's seen an offshore oil rig or a gas field, you see these big burning torches in the middle of the night, they're burning all that gas, all that excess gas. Now they're capturing that gas, pipelining it over to the data centers and they're using it to generate power for the data centers. What's burned and ended up being CO2 that was wasted in the atmosphere now can at least be used for something. And so there's a lot of misinformation about how those plants work.

00:55:52 Sure. Those plants need some water, but if you use fuel cells, you make water. And you notice like Project Jupiter down in New Mexico, fuel cells, couple companies out there that are building fuel cells. All of them are using

fuel cells now. Fuel cells are great technology, just like heat rejection, fuel cells in the future are data centers. So they're not only upgrading the grid, but they're adding additional capacity. And then here's something people don't know. Data centers have interconnection agreements with the grids or your utilities. When your utility has a problem or a big storm comes or you have a heat wave event or whatever it is, data centers are required to start their onboard generation, whether it's fuel cells or diesel generators or whatever it is, interconnect and push power back to the grid to support your local grid in time of crisis.

00:56:45 And so the more data centers you have, they're like many power stations all over the country. They can provide stability of the grid when the grid operator can't handle that hundred year ice storm that comes through town. So it's a symbiotic relationship that is just not well understood. I have to admit to all the listeners, this is a complex topic and I covered it at a very high level, but at the same time, data centers aren't bad for the grid. Data centers don't cause your power build to go up. They just don't. That's a misnomer just like they use water. No new data center uses water.

Jon Krohn: 00:57:21 That was obviously a well practiced set of information and although it was high level, you did have a lot of detail there and I learned a lot of things from what you were saying. I think there will still be some people out there who will say, "Well, it's bad to be putting gas fired plants online. We should be using all sustainable power." And I think that that's an ideal and I think a lot of data centers do get built with commitments to be using nuclear power, maybe worst case, but also to be using as much solar and wind as possible. So that does happen. You see lots of data center contracts being agreed to today where it is entirely renewable power sources and that's a great way to be moving. I think there may be some data center scaling happening globally where it's happening so

quickly that a gas fire plant is needed in the short term to power that.

00:58:12 An argument that I make that kind of, well, at least allows me to sleep well at night as someone who is so deep in the AI world is that AI is helping us increasingly push the frontier of what humans are doing, including things like making nuclear fusion commercially viable, which is hopefully the energy of the future where we can be splitting water to be ... There's small amounts of water. It's not like we're getting around out of water because of the nuclear fusion plants, but creating huge amounts of energy, like having suns on our planet and then that allows us to pump carbon dioxide back into the earth. And so anyway, it is a very complex topic. I understand the concerns about energy use in general and I'm not asking you to necessarily have any comment on what I just said, Frank.

Frank Basso: 00:59:03 Yeah, no data center industry is pioneering this tech. Data centers are why we're looking at small reactors now. It's why reactors are back on the scene. Nuclear power is not evil. Nuclear power is great. It powers our entire Navy and has without incident for many years. I remember when the tsunami hit Indonesia, well, the USS Ronald Reagan, a nuclear aircraft carrier, pulled up and connected itself to the shore and provided power for the country for like a month.

Jon Krohn: 00:59:33 What?

Frank Basso: 00:59:33 Yeah, it was in the harbor. Oh yeah. They provided shore power for emergency services and stuff like that. Nuclear power is not the enemy. It's more uncertainty and unknown, fear of the unknown, but data centers and data center technology companies are, you might have heard like we're increasing the voltage, we're going away from AC power because we can make the entire place 5% more efficient by going back to DC power and Going higher

voltages, 800 volts DC and then you're talking about coming out of the fuel cell is DC power. Then you put a grid tied battery energy system in there like a giant Tesla grid battery or some other brand.

01:00:15 Those are battery tied. Now you have no diesel generation. You don't need generators, you don't need UPS systems with big batteries in them because you already have the battery bank outside and then it's straight DC power all the way in and you have efficiencies of five to 10% gain on the power lossiness within the building. So your overall efficiencies go up and the data center business and the supporting industry is pouring billions of dollars into this technology for cleaner running data centers and always looking for better. I wish they'd spent this much money on cars to make gas fired cars this efficient. And the battery systems are being driven by data centers, not just the car manufacturers. The data centers are looking for cleaner power, longer term power, salt water based, water-based electrolytes and non-chemical electrolytes. There's no fire hazards and salt batteries, salt chemistry batteries.

01:01:13 These things are all being driven by the data center business. They really are because they're grid level storage and four data centers. So you can use renewables, fill those up, run on batteries overnight and then let the sun come back up and power the data center. That's all happening.

Jon Krohn: 01:01:29 You talking about those kind of 5% gains and efficiency, that reminds me how before we even had the GenAI era, I believe the first commercial value provided by the DeepMind acquisition that Google made now a decade ago was creating efficiencies in the way that power was being routed within Google data centers, allowing them to get those kinds of single digit or low double digit efficiency improvements. And so yeah AI has for a while been helping reduce consumption even as we build more of

these AI data centers. So yeah, complex topic and I'm certainly not an expert, but I'm hopeful for sure in the long term with what we're doing with AI. And I strongly believe as any regular listener will know that I'm techno optimist and I think things have never been better, things are going to continue to get better. Hopefully that's kind of a nice note to start to end the podcast episode on.

01:02:25 Frank, you and I were discussing before we started recording about what book you might choose as your book recommendation for our listeners. And if you are going to go with the one that you recommended, it's kind of perfect because we were just talking about alternating current, direct current and yeah, is that your pick? Are you going to go with the same one?

Frank Basso: 01:02:42 No, that's my pick. It's a good book. It's been out for a while, but it's a good listen. It's based in truth with a little twist because it's about Edison, Tesla and Westinghouse in the early days and it's called The Last Days of Night. So talking about electrical power and electricity to homes and buildings and streetlighting and things like that. And it's written from the point of the view of an attorney who's involved with all these things. And so it's very interesting that it's a total outsider watching Westinghouse Edison and Tesla argue and discuss these things and things that happened and very interesting book, The Last Days of Night.

Jon Krohn: 01:03:27 Cool. Yeah. And a good reminder actually, now you mentioned that kind of the last days of night kind of also ties into my techno optimism and how there's probably not many listeners that would like to go pre-electricity. And in the future, I think people couldn't imagine going back to a pre-AI society where we have unmetered intelligence, just making everything easier than ever before for us and hopefully allowing lots of positive human outcomes as well. Frank, I actually didn't warn you about this, but my final question that I ask every

guest is just how people can follow your thoughts after this episode. I don't know if that's, should people be following you on LinkedIn or what? Did you post anywhere publicly?

Frank Basso: 01:04:12 No, not really. Occasionally on LinkedIn I'll share things that I think are salient and what I get exposed to day in and day out, like some of our customers who are working towards cures for human illness and things using AI tech and those things that are not like, "Oh, I can play this game better or I can do this better." No, real things that move the needle, things that move society and the world forward, that's why we have this AI tech. Those are the things that really drive us. And so I'll post and discuss those. Very interesting. But yeah, I am of course always on LinkedIn. Besides that, I'm kind of a reckless online, typical security background guy that doesn't post a lot online.

Jon Krohn: 01:04:56 Well, so almost an exclusive here into Frank's brain. Unless you are also, if you're a Lightning AI employee and you have access to the Slack, Frank, you've recently been a very heavy user of the random thread on Slack and posting things in there. I think you've been the number one poster.

Frank Basso: 01:05:14 Well, there's lots of fun things to share and random they are.

Jon Krohn: 01:05:20 Maybe you're just doing that because you can't shout it loud enough in the data centers. So you're like, "Here you go. It's in the Slack."

Frank Basso: 01:05:26 Exactly.

Jon Krohn: 01:05:27 Frank, it's been really interesting having you on the show. I have learned so much. I'm sure a lot of our listeners have as well. This was an important episode to help us understand how data centers are built that are allowing

us to have all the AI capabilities that we're talking about on the show all the time. So thank you, Frank, for taking the time out of your super busy schedule. Really appreciate it.

Frank Basso: 01:05:46 Thanks for having me, John, and thanks for everyone who actually tuned in to listen.

Jon Krohn: 01:05:51 Love that episode today in it. Frank detailed how Lightning AI uses co-location to provision its GPUs, sizing each build around available grid power, applying a conservative 81% derating, and then contracting per megawatt over typically 10 year terms. He talked about how liquid cooling doesn't waste water at all. Modern data centers run on a sealed glycol loop and reject heat through giant radiators so a huge facility uses no more power per month than a single household. He talked about how GPUs talk to each other over ultra fast east-west networks limited to a few hundred meters and how thanks to screaming Vanshees, AI data halls run at 105 to 110 decibels louder than the front row of a rock concert. As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Frank's, social media profiles as well as my own at superdatascience.com/1003.

01:06:49 Thanks to everyone on the Super Data Science podcast team, our podcast manager,Sonja Brajovic, media editor, Mario Pombo, partnerships team Natalie Ziajski, our researcher, Serg Masís writer, and our founder Kirill Eremenko. Thanks to all of them for producing another outstanding episode for us today for enabling that great team to create this free podcast for you. We are deeply grateful to our sponsors. You can support this show by checking out our sponsor's links in the show notes. And if you ever want to sponsor the show yourself, you can see how to do that at jonkrohn.com/podcast. Otherwise, please help us out by sharing this episode with someone

who would love to learn about AI data centers. Review this episode on whatever podcasting platform you listen to podcasts on or on YouTube. Subscribe if you're not already a subscriber, but most importantly, I hope you'll just keep on tuning in.

01:07:40 I'm so grateful to have you listening and I hope I can continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.