Patrick Moorhead: Sam it's great to see you again. And thank you so much for talking here at the Six Five Summit 2022. I know this is going to be a great adder to the semiconductor track.

Samuel Naffziger: We're excited to be here, Pat. I know I've got some big shoes to fill since my CEO was the last guest at your summit from AMD. But I've got some really exciting content here that I look forward to sharing around the energy efficiency challenge that I think is probably the biggest thing that we're going to have to deal with in our industry over the next decade.

Patrick Moorhead: Yeah. Sam, you and I both worked at AMD the same time when I was there. And even back in 2003, we were talking about power efficiency with Opteron. And in fact we put a, you save this much power and this much money up in times square during the launch of the product. I know you have been front and center, and it was one thing for me to talk about that power efficiency and even today talking about power efficiency, but you're actually engineering this capability. So what makes you think though that things are different now and that there's really a looming energy challenge on the horizon?

Samuel Naffziger: Yeah, no. That's great. Because that's the background I'd love to go through and I've spent, as you mentioned, the last 16 years at AMD and we overlap there. Power technology has been a primary area of ownership for me and now I work across all our product lines deploying and selecting the best technologies that make for the most efficient, highest performance, just all around best products. And so that forces me to study the underlying physics, the trends, what drives performance and see what's happening in the future across our industry and the technology landscape.

So going to why this is a bigger challenge than it has been in the past. And it's being fundamentally driven by a culture of more, I would say. And the culture of more is coming out of the benefits that people are garnering from all of the technology that we've been designing. They've really become dependent on it, we all have for navigation, for our video sharing and editing and all of the things that we don't see that go on out in the cloud, voice recognition with Siri. It's just a pervasive aspect of our life and all of that requires a lot of computation.

So the demand for these computes is just not abating. As this graph shows, the server performance is a perfect proxy because these are the things that set out there on in the cloud and do the work that we all depend, whether it's financial transactions or image recognition and stuff. And these have been seeing the historical exponential growth that's been going on for decades and it's continuing, it's up unto the right. And it's this sort of annual improvement that we become dependent on. If we don't deliver these kinds of gains, then the expectations for the next generation of features just aren't going to be met, people aren't going to upgrade and the capabilities we need aren't going to be there. So this exponential improvement, which is often correlated to Moore's law, has been a critical aspect and we've our industry depends on it.

Now, the problem that is much discussed is that the fundamental physics of Moore's law is running into serious headwinds. The technology nodes are slowing down in introduction, just because they're so hard to develop. We have billions of dollars in R&D and the best and the brightest on the planet working on this stuff. But all those PhDs are still coming up short of the

historical norms for process scaling. And as shown in the little chart on the right there, the density gains, generationally, the nodes are slower at introduction, but also the density gains we get are less and they're varying by component. And kind of the fundamental part of Moore's law that we depend on is that density improvement. And it doesn't just give us smaller chips or more gates in the chip, it enables lower energy cooperation because you get things closer together and smaller it's intuitive That there's less power required to get them to operate.

So, this represents a serious concern and an issue in meeting that exponential growth. The power demands, as shown here, are a corollary to that slow down colliding with the growth and performance. So if we aren't getting the energy improvements from our base technology, but we're still having to deliver the performance, it's just going to take more energy to get the work done. And unfortunately we are seeing a serious uptick in the thermal design power focusing in on the CPU server example. And that means it's a lot more watts and the system a lot more heat generated to do the computations.

So to wrap up the answer here on why do I see this as a problem that the biggest issue is, or the question is, are we still getting more work done per joule consumed? And the reality is that improvement curve is tailing off, is shown here. So we can just simply plot the performance per watt of our servers. And the performance is still going up, which the market and all of the world requires, but we're just taking more watts per unit of performance. It's kind of like if you drive an EV, EVs are great but when you start driving into a headwind, your range starts tailing off. So this is a lot like driving our compute performance into a headwind, we need bigger and bigger batteries just to get the miles to reach our destination.

Patrick Moorhead: Yeah. I loved your opening. And, and I think it's not only a culture of more, it's a culture of more and now. Particularly when I look at millennialism and Gen Z coming in and we want more and we want it faster. I love the charts that you show, and it really does paint a very sobering picture as it relates to servers. I mean, I can't believe we're at 400 watts right now for a TDP of a two piece server, that is astronomical. And the curve of the amount of work we're getting done is not necessarily getting any better. In fact, it looks like it's getting work. But CPU servers are one thing, but you know, the IT industry is a multi-trillion dollar industry. And I'm curious what maybe the other trends look like in some of these other segments, maybe like GPUs or something like that. It's funny as a follow-up to that. If you look at what Bill Gates talked about with 640K of memory is everything we always needed, reminds me a little bit of, we'll only need three computers in the entire world. And others said that the gigahertz are over. By the way, I said that in the year 2001, and it's more about performance, but it seems like aren't processors already faster than anyone needs? And most of these statements are inaccurate, but do you think we'll ever hit the point where statements like this will become a reality? Because it looks like we're about to hit a wall here, and then really the innovation moves to more of software and applications, and computation becomes essentially a commodity.

Samuel Naffziger: Yeah. Well, that is the big question. And then you and I have been in this industry long enough to have made some of those statements and heard some of those statements that have been proven wrong over time. And there is the culture of more, and the fact that the world is now addicted to the gains we get from IT improvements, but there are also a lot more things, a lot

more improvements, benefits to society driving the continued growth that I do not think will ever abate in the foreseeable future unless they hit an immovable object. And a good example is the scientific computing domain. Here is just the top 500 number of flops for these supercomputers. And it's been going on exponentially doubling much faster than Moore's Law for decades.

And it's not just to get the bragging rights for the top spot, although we won't turn that down. And AMD is super proud of breaking the XPlot barrier with our frontier supercomputer deployment, but these things are used for critical capabilities for the benefit of humanity, climate modeling, and genome sequencing, drug discovery, all these things that have a boundless appetite for computation. And they really impact people's lives. And protein folding calculations for disease prevention, those have just vast demands for compute and we're not even close to being able to do them accurately. So super-computing is one area that is just not abating.

And then a completely different one, but similar underpinning technology is in the gaming space. And we've all seen with lockdowns and such just how people have immersed themselves in gaming. Esports has become a multi-billion dollar spectator event.

Patrick Moorhead: Yeah.

Samuel Naffziger: And people, once again, they get addicted to the photorealistic capabilities, rate trace, lighting to faster frame rates, and just the immersive experience of gaming. And it's more than just an individual holed up in his basement. The thing that has transformed it is the interactive nature of all the online gaming, connected, meeting people around the world, and interacting with them. And it has to be a realistic immersive experience. And it requires a ton of compute and we are not even close to reaching the levels of realism that people would really want to see. This drives a flop growth, another exponential here in game requirements.

And probably the mother of all workloads, the killer app, if you will, the thing that disproves three computers is enough is the growth of machine learning. And what we're seeing here is just an astonishing increase in compute demand that's doubling in a matter of months rather than in years. And, of course, this is an exponential that certainly can't last forever, but it's been going on for several years now. And we're, in this industry, all working on meeting the demand several years out based on these trends. And the reason these things are getting so huge is once again, they're feeding the culture of more, they're meeting very real felt needs by all of our population. Things like language translation capabilities. You and I travel a lot, so the ability to translate language real-time when you're overseas, or photograph a foreign menu and then read the translated text, transformative capabilities, but they require a lot of compute.

They require these huge billion or trillion parameter networks to perform that recognition function off in the cloud somewhere. And you got to train those models. And training those models takes days and weeks on these gigantic compute systems. The demand isn't going away and it all takes energy. That brings me to the final slide in this section, which is the trends in energy use. And the semiconductor research council did a great bit of work here showing the

growth over time of energy demand. And the fact that these compute trends are going to have IT equipment consuming a very large chunk of the world's energy unless we figure out how to do things differently. It's something that's going to require a completely new set of solutions.

Patrick Moorhead: When I first saw this slide in the green room, I stared at it because I had never seen anything like this before, but actually, it makes perfect sense. We casually talk about trillions of devices on the edge. Well, guess what? Those trillion devices on the edge consume power, but they also have to be orchestrated and managed by a bunch of edge compute, a bunch of data center hyper-scalers, and all of that has to be trained in some sort of machine learning model. That's amazing. Could be 75% of all of the energy consumed is from electronics and semiconductors. So actually compute, that's pretty daunting. You've convinced me that yes, we have a challenge. And you alluded to new approaches must be found, so I'll hit you straight up. What are these new approaches? And do you think they'll be sufficient to meet this incredible energy demand?

Samuel Naffziger: Yeah. Well, do I think they'll be sufficient? I have a lot of confidence in the ability of our industry and our engineers to innovate. I think we will find a way. Let me walk through what I think some of those approaches are going to be, although certainly neither I nor anyone else knows exactly how we're going to put all those together to meet the demand. But starting at what we need to do differently. This plot here shows the trends in efficiency of a general-purpose CPU. And this is why that CPU server performance per watt curve is bending over because... This is a real simple chart, once again, from the semiconductor research council. And it just shows the energy on the X-axis and the work done on the Y. And sure, as you do more work, your Y axis goes up, it requires more energy. The problem with this chart is that the empirical data from way back in 1971 to 2020 follows a surprisingly consistent trend that shows that we need to switch a lot more bits, in other words, consume a lot more energy to get a unit of work done, the more performant our processors are.

So that's what this says, and it's not a one to one, it's an exponentially decaying return on investment in CPU performance. If we had scaled at one to one from 1971, we'd be a million times more energy efficient to date than we are now. So that 400 watt server it'd be less than a watt if it were perfectly energy efficient.

Patrick Moorhead: Right.

Samuel Naffziger: So yeah, this is why general purpose compute, as valuable as it is, it's just intrinsically inefficient. And so we're going to have to do things differently, and so that's where we can look to GPUs, are they more efficient? And to answer that question, we really have to get to the performance per watt curve. But you can look at the flop rates, it's been following an almost identical curve to the CPU side and doubling flop rates, floating point operational rates every couple years or so. So huge demand for these, for gaming, for compute, the power consumption looks like deja vu all over again here, a serious uptick in recent years. So meeting that demand, we're throwing more power into our systems. And I think we see this as even gaming Notebook cards are coming out at 450 Watts plus. Not Notebook cards, desktop. 450 watt Notebook, that would be-

Patrick Moorhead: That'd be big.

Samuel Naffziger: But then the million dollar question here is are GPUs is actually more efficient? And the encouraging answer here is that just looking at floating point operations, not delivered workloads, the CPU plot was more of the spec rate, which is representative real workloads. But if we look at floating point operations, we're still doing pretty darn good. We're continuing that improvement curve. So we can drive the fundamental atoms of compute to higher efficiencies over time. Now, I do expect this curve to tail off due to the physics that I mentioned earlier, but this gives us a hint that these kinds of special purpose devices can keep the efficiency gains going longer. And so that's where the special purpose accelerator architecture has become a real interesting direction.

Patrick Moorhead: Yeah, it is funny. I had my head going, "Hey, we're all going, ASICs, baby, here we go. We're going to push everything onto the software developer, and we're going to be really slow as an industry and we're going to be locked in." Because ASICs might be efficient, but they really lock you in more than not/ specifically on the GPU, I find that fascinating. It is good to see that at least on the floating point, it looks pretty good. But GPU's are a pretty mature architecture now, and therefore, do you think that it'll flatten on the efficiency curve, like you showed with the CPUs two or three questions ago? And if the answer is yes, is there any other technology or architecture that can keep these gains going?

Samuel Naffziger: Yeah. Well, the physics of process scaling are going to result in even GPU slowing down on that efficiency curve. Now the answer to that second part of the question is, are there ways ... So the move from a CPU to GPU did produce a nice efficiency improvement, what's the next step in that continuum? And really, we are going to have to not necessarily an ASIC, but there's a better answer that I'll get to in a bit, I think that-

Patrick Moorhead: Oh, wait a second. So maybe I'm partially right. Wow.

Samuel Naffziger: You've been doing this in a while, Pat. So I think you see the opportunities, but we do have to solve that programability problem, and we have to retain the general purpose capabilities or else people won't be able to access the performance.

Patrick Moorhead: Time to market will just shift out exponentially, and we might get the energy, but it's going to be years after we actually needed it.

Samuel Naffziger: Right. So what we're seeing though, is that we're just being forced into the mode as this conceptual chart shows, using more at domain specific architectures. The general purpose CPU, as convenient as it is, and ubiquitous, easy to program, it's just not sufficiently efficient to meet the world's compute demand in the future. GPUs are doing a lot better. We've figured out in large part how to program them, and we have some very sophisticated software stacks and making rapid progress there on enabling easy offloads to these more parallel devices. And it's a continuum, where for a narrower set of applications, we can go to more special purpose accelerators, and they are just intrinsically more efficient. The other big key, I think we're going to require is around package innovation.

And I get to this a bit, but we've got to be able to integrate these domain specific architectures efficiently together. And that means we're going to be splitting up our die into smaller components so that we can have application specific acceleration for given market segments and in a different mix for other segments. And if we're going to enable that, we can't have a lot of overhead communicating between these chiplets. And so the 3D stacking is going to be a huge enabler for these architectures of the future to have heterogeneous compute where we have low overhead, low energy connectivity. So I see that as a really big opportunity, and these things are going to transform the way we design these processors.

Patrick Moorhead: Yeah. So I'm learning a lot throughout this conversation. I definitely was dialed into the specialty of compute, but I didn't actually know that modularity through either tiles or chips could actually help in energy efficiency. I had no idea. So I've learned a lot there.

Samuel Naffziger: Well, maybe I can jump in there. And I'm enamored of what we call a decaf solution, which is a little diagram I showed. And what we did there is that little tile you see on the top is a special purpose cash chip, and all it does is store bits as densely and in as low power way as it can. And it sits on top of the CPU compute die below. Well, what we were able to do on that little tile is pack in twice as many bits per millimeter squared as anyone can in a more general purpose CPU cash down below, because that CPU cash down below has to have all the control, the interface circuitry, and all the wires to shuttle the bits around.

Patrick Moorhead: Yeah.

Samuel Naffziger: That cash expander die, all it does is store bits and send them down through the vertical, through Silicon vias. So, we ended up with a very low energy pathway from that denser S Ram down, and we get 20, 30% accelerations in performance for virtually no power by going to this 3D. Now, it is domain specific because not everything needs a big cache, but there's a lot of workloads out there that really do. And so yeah, this is a great example.

Patrick Moorhead: I appreciate you. Yeah, I should have known, yeah, you don't have to go to the memory bus. It's right there.

Samuel Naffziger: That's the uniquely 3D capable, right? If you tried to triple the cache and get that 30% performance on a 2D form factor, you'd have to go millimeters away, spend a bunch of energy shuttling the bits back and forth.

Patrick Moorhead: Yeah. So you talked about 3D V cache. You talked about cDNA and rDNA GP architectures. Do you have any other examples you can share of how specialization can improve efficiency?

Samuel Naffziger: Yeah. So let me jump into a few that I see really promising and show that ... There's a software implication and it's going to take a lot of innovation, but big opportunities. So, one of them is just around the precision we use to do calculations.

Patrick Moorhead: Yeah.

Sam: And we've been traditionally using these I tripe E formats, and they're great. They're like the gold standard. They're like the general purpose CPU. They can do everything. You don't have to think about it, but they're overkill for a lot of applications. They're great for climate modeling. They're great for collision detection in your car, but if you're just doing audio, you're balancing in the cabin, or you're asking Siri for a weather update, that doesn't need 64 bit math calculations, right? And so we can get by with many fewer bits for a lot of these computations that the world's using computers for.

And so there's been a lot of great research going on in these alternate precision arithmetic formats and how far they can be pushed and applied. And in the machine learning space in particular, there's been a lot of work, way down on the bottom end of this chart around 16 bit or even eight bit formats, that are good enough for the calculations we need.

And so how much of a benefit is that you might ask? Well, one simple example is here with the size of the multiplier required to do the calculations with these math formats. The gold standard 64 bit, it's got 52 bit, well, multiplier area goes as the square of the number of bits, right? So you go down to 32 bits, it's a lot better, about a fifth the size, but if you can get down to these much smaller, tighter form factors, you're more than a factor of 50 smaller, obviously much lower power, much lower area and for a ton of applications. Good enough. So there's going to be a lot of work here and the way we use the binary bits to do the calculations that people need.

Patrick Moorhead: Yeah. It's good to see people starting to get on the bandwagon in real implementations of things like B float 16, and those are huge proficiency gains. I had no idea that was the case, but you definitely have to do things differently because we got really comfortable with FP 32, FP 64, because we had been doing them for so long and that's why we built it into everything. But it sounds like, though, we're going to have to change the way that we do things with future architectures and I'd like to hear what you have to say about that. What will these future architectures look like, just to close out this session?

Samuel Naffziger: Yeah, well, it's going to be a lot different than the nice homogeneous architectures that we've gotten used to, which is exciting and daunting at the same time. It's going to be also a much more multidisciplinary future where we're going to have to be bringing together ... We touched on packaging technology there a little bit and 3D stacking. Well, that's a whole different set of skills and constraints than traditional Silicon design. And what we're seeing at AMD is that our package engineers, package experts, they have a front seat at the table as we define future products, because the way we integrate these different modular components and chiplets is very tied to the underlying package technology. Then we are working with the software guys because they they have legacy libraries and code they have to support, and we can't just invent hardware willy-nilly, as you've already hinted at.

But we've got to do things differently with new number formats, with domain specific architectures, and then connect them together in ways that are actually usable. And so, the connecting together is where we get to modular design and every market will need a different mix of accelerator capabilities. That's, CPUs, I keep saying, they're great because they can do everything. They just don't do anything especially well and going forward. So we need to take

those CPUs and pair them up with stuff that's really good at doing particular tasks, and the right mix of those domain specific accelerators is going to vary market to market.

So the companies that have the vision and the cross-disciplinary skills to understand how to connect these, how to work with the industry, to establish standards on connectivity, and how then to integrate them with leading edge package technology, is going to have a huge advantage in the future as we work to overcome these energy challenges.

Patrick Moorhead: So Sam, this is the most interesting conversation I've had on energy and semiconductors. I don't say that lightly just because you're on our stage here, but even I can understand what you said and as you know, I don't have an engineering background and I think it's important for everybody, even if you don't have an engineering degree, or not a technical fellow to understand and appreciate where we're headed, because it really is going to dictate our future. And with technology being such a part of the fabric of our lives every day, we better get this right, and I'm glad we have people like you on the case, and I'm glad we have companies like AMD on the case. So, I appreciate you coming on and thank you. We'd love to have you next year, too.

Samuel Naffziger: Super fun to be here, Pat. I just love innovating. I love technology challenges, and I guess, most gratifying is that we really are making people's lives better. We're providing the capabilities to solve the world's hardest problems, and the fact that we have to innovate on new dimensions, physics, and metallurgy and software library routines, that just makes it more exciting. So yeah, really good to talk about it.

Patrick Moorhead: Thanks again.