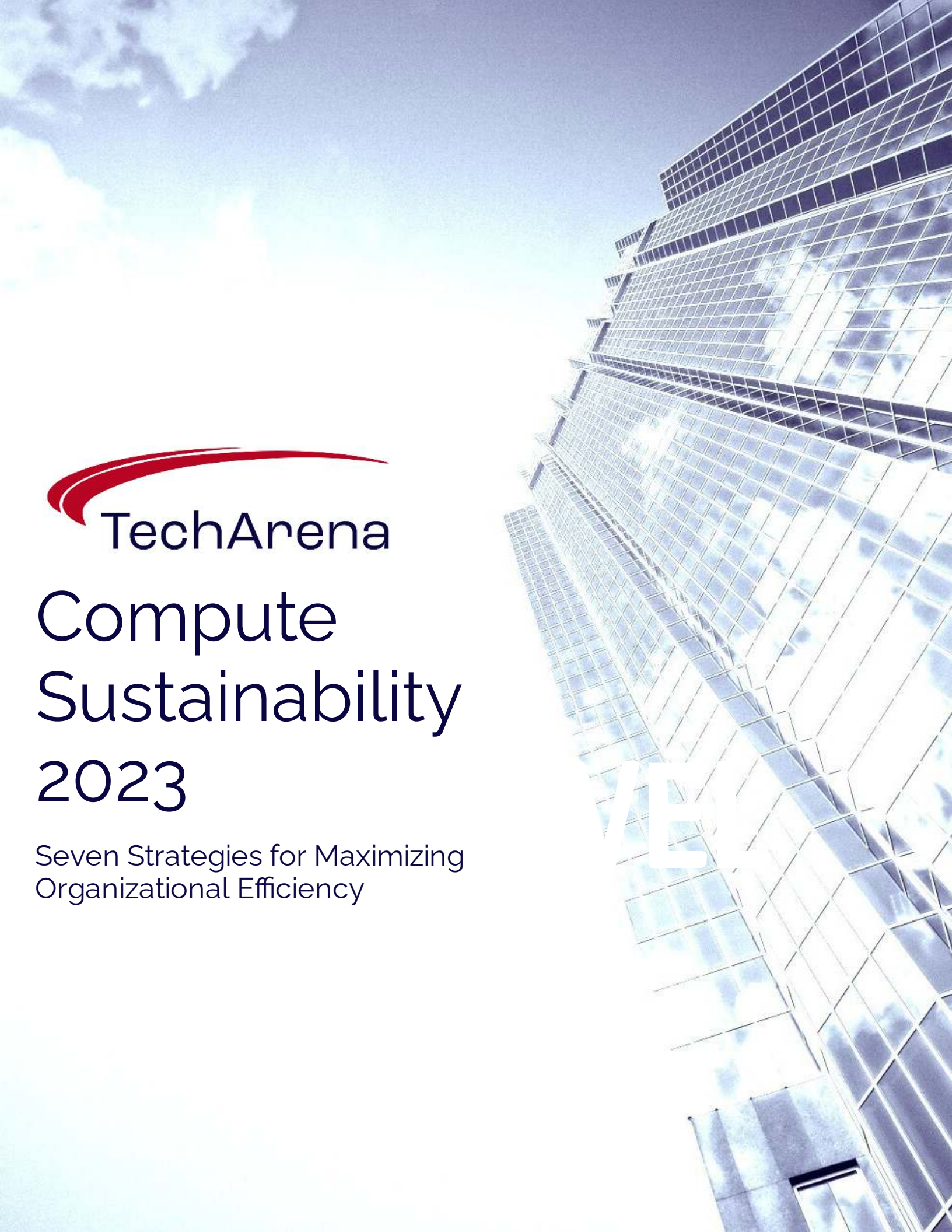


TechArena

Compute Sustainability 2023

Seven Strategies for Maximizing Organizational Efficiency

> **- It's warming, it's us, we're sure, it's bad, we can fix it.**
>
> *— Kimberly Nicholas, Associate Professor of Sustainability Science, Lund University*

There is no issue more pressing to our planet and humanity than the climate crisis. With global temperatures scaling, sea ice melting, and every region of our world facing crises as far reaching as rampant wildfires, coastal erosion, loss of arable land, and disruption of deep-water currents we are at the brink of unrecoverable damage to the planet and existential threat to human survival. In Jonathan Koomey and Ian Monroe's book, "Solving Climate Change, a Guide for Learners and Leaders", the authors detail an actionable path for organizations to enact in order to be part of the solution. They argue that technology exists today to pivot society to a carbon negative state and restore planet health. They further surmise that if governments and private sector entities work together with urgency, we will still have opportunity to recover. They also call out technology as a strategic tool in this fight placing tech sector contributions at the heart of its success.

— The climate targets that we're setting right now will have profound implications on our future. We really need to make those climate targets much more aggressive, aligned with the reality of climate science. There's been a movement of setting net zero by 2050 targets which would have been effective if we had started acting a couple of decades ago based on what the science said then. But now we have a couple more decades of pollution that's built up in the atmosphere that will stay there for over a thousand years without us actively removing it. The warming is happening faster, and we need to decarbonize even faster. Our target should be getting beyond net zero to net climate positive.

- Ian Monroe, President and CIO, Etho Capital

Reading tech sector corporate social responsibility (CSR) reports uncovers a swath of commitments towards carbon neutral and carbon negative status within the next decade. Digging under the covers, organizations are committing to scope 1 (carbon sources owned by an organization directly), 2 (indirect emissions from the purchase of energy), and 3 (indirect emissions in the upstream and downstream value chain) carbon emissions reductions as well as other resource oversight including water replenishment. While large cloud players arguably receive the most attention for these commitments, broader industries are also seeing an uptick in climate commitments tied to programs like the UN-backed Science Based Targets initiative.

As more organizations are taking action towards climate commitments, the management of IT infrastructure and investment must be viewed through a sustainability lens. This starts with a strong tie between IT leadership and corporate responsibility teams and extends into development of an IT plan for measuring all resource utilization as roll up for broader corporate programs. The process of mapping IT energy consumption as part of broader CSR efforts often uncovers decision making paradigms that have historically rewarded IT performance over other considerations.

— When you look at the global impact of data centers, the share of CO2 emissions just caused by data centers looks to be almost three- and one-half percent. It sounds like a small number but it's not. It's larger than aviation, it's larger than shipping, it's larger than rice cultivation.

Bev Crair, Senior Vice President, Enterprise Intelligence and Resiliency, Oracle

It's important to note that there is dissonance between industry quotes of data center energy consumption that you'll read in this report at 3% - 3.5% and published reports including ones from Dr. Koomey measuring global data center power draw at 1% and all IT infrastructure draw at 3-4%. The TechArena recommends reading Koomey's article on the topic to get a sense why this dissonance persists. In consideration of all experts quoted in this report, we have left distinct figures referenced to the source. What's not in debate is that Moore's Law is slowing down, servers are requiring more energy, and workloads are demanding more processing placing data center operators facing escalating challenges with lowering emissions. We expect that in the months ahead this dissonance will coalesce in a new report forecasting energy draw that will align expert thinking on the topic.

The TechArena top seven strategies for compute sustainability report is based on agreement with Koomey and Monroe that computing, and in particular data center compute power, is critical to collective organization ability to drive broader sustainability innovation. What's presented is a summary of discussions with industry experts on the keys to sustainable infrastructure manufacture, deployment and operation and underlying tenets that should be integrated into every organization's IT sustainability plans. Through these interviews, we discovered an industry very much focused on rapid innovation of compute architectures, software code, power and cooling paradigms, and standards that will fundamentally change compute utility/watt while moving to lowering embedded energy and broader resource utilization. Anyone working on plans for sustainable IT, data center energy consumption, IT infrastructure procurement, or software development can benefit from the strategies presented.

While the industry has been working for well over a decade in earnest on compute efficiency, the urgency grew exponentially over the past two years with the pandemic creating more demand for data center compute and the Ukraine war and other events raising energy prices and constraining supply around the world. With large data center providers like Google, Microsoft, Amazon, Meta and more committed to broad sustainability goals, the Open Compute Project (OCP) stepped forward last fall naming sustainability as a lead organizational priority and chartering an initiative to fill the gaps across lowering embedded carbon in silicon, unleashing circularity through infrastructure reuse and recycling, and measuring what matters beyond PUE.

> **- The energy consumption in 2011 for data centers was roughly one percent to one and a half percent of the total energy consumed in the world. By 2030, data centers are expected to consume anywhere between three to 13% of the total energy consumed in the world, which is massive growth. In fact, the numbers are expected to be anywhere between 7,000 terawatt hours to 30,000 terawatt hours. This growth in energy comes with a direct implication on carbon. I felt Open Compute as a community had a moral obligation to address this because we were contributing to that energy consumption. We were building devices which go into data centers which contributes to all of that. If not us, who would be the community who would address this working on the data center gear?**
>
> *— Dharmesh Jani, Infrastructure Product Lead, Meta and Sustainability Initiative Chair, Open Compute Project*

This report is very much influenced by the OCP initiative's work with a goal to take early contributions of the group's output into suggested action plans, and we thank the organization for their engagement.

Strategy 1: Maximize Compute Utility/Watt and Tackle Embedded Carbon by Demanding Efficient Silicon Innovation

The first objective of a sustainable data center environment should be focused on maximizing compute utility per watt of power consumed to ensure that energy investment is optimized for organizational work. Many IT teams are evaluating compute engines based on Arm architecture as power sipping alternatives to the well-established x86 architecture offered by Intel and AMD in order to lower the total energy demand fueling compute.

> **- What Arm is enabling in the market is increasing the core count to such a high number that you're able to effectively get servers that are single socket that have equivalent performance as you would get in the past with two or four socket systems. That alone is a fantastic TDP savings. We're really excited to continue down this path of delivering more solutions that don't explode the TDP budget in systems.**
>
> *— Eddie Ramirez, Vice President of Marketing, Infrastructure Business, Arm*

This energy efficiency advantage is a major reason why earlier this year, for example, HPE delivered platforms featuring Ampere's Arm based processors targeted at the heart of the enterprise and Oracle announced optimization for Ampere systems for Oracle database environments. The large cloud players are not alone in seeking more efficient compute.

> **— From day one efficiency and sustainability have been core to our mission. As data centers consume more and more power, efficiency has become core to how clouds actually build out compute. That's why we have focused on maximum performance in an efficient manner.**
>
> *— Jeff Wittich, Chief Product Officer, Ampere*

That does not mean, however, that the x86 crowd is standing still. With the release of its latest Xeon microprocessors, Intel announced a 14X performance efficiency improvement based on embedded acceleration technology. Intel has also shown a light on the growing importance in embedded carbon management relating to customer sustainability commitments. The company proudly boasts that its Xeon processors are manufactured with ninety-three percent renewable energy providing customers a more transparent, lower embedded carbon profile than competitive offerings.

> **- We've built more energy efficient CPUs that can deliver up to 14X performance per watt, we have new features like an optimized power mode that can save twenty percent power or up to 140W since we're talking about energy.**
>
> *— Jennifer Huffstetler, Chief Product Sustainability Officer, Intel Corp.*

AMD, for their part, has focused on delivering efficient high-performance cores, and argues that acceleration is overhead. They have also leaned into chiplets within their Xen architecture to drive increased design efficiency to their processors.

- Our ability to drive beyond the radical limit of silicon by using chiplets enables us to package an order of magnitude more silicon than a monolithic die delivering improved performance per watt. We're saving a lot of energy efficiency, so that's really the best lever we have today to drive energy efficiency deeper levels of integration, core scaling and breaking radical limits on silicon design.

— Robert Hormuth, Corporate Vice President, Architecture and Strategy, Data Center Solutions Group, AMD

The other key trend that could disrupt embedded carbon significantly is modular server design where processor and memory modules are swapped out independently from motherboards and chassis. This alternative approach to infrastructure lifecycle management places focus on upgrades to components that move the needle from a performance per watt perspective while reducing unneeded embedded carbon consumption with non-performance related platform elements.

> — I've been very public about my commitment to modularity. Modularity is how we reduce the embodied carbon in each new server generation. So instead of what we've done for the last twenty plus years in this industry, where with every new server, it's a new rack, it's a new motherboard, it's all new components. How do we identify the subset of the system that is getting more efficient? The CPU, the memory subsystem, possibly the peripherals, but not every component. Like your baseboard management controller is not getting significantly more efficient for one gen to another gen.Why do I need a new one? Why can't I reuse all of that portion of the chassis and just swap the compute module with the memory? That's actually that much more efficient. The first step is to have modular sub-components well-defined for interoperability and only swap what you need to swap. That greatly reduces your embodied carbon footprint.
>
> *Rebecca Weekly, Vice President, Infrastructure Engineering, Cloudflare*

So what are organizations to do with this silicon environment for mainstream workloads? The TechArena suggests starting by balancing evaluation of compute platform performance with performance efficiency metrics like SPECPower and demanding real-world application performance efficiency measurements from vendors. This includes evaluation of alternative architectures like Arm pitted against AMD and Intel x86 alternatives.

IT leaders are also encouraged to demand published embedded carbon reports for all compute systems and silicon. This transparent accountability will place more pressure on the industry to optimize manufacturing for sustainability and choose renewable energy and other carbon reducing practices wherever possible. It will also pave a path for IT integration into corporate CSR reporting.

> **— We're trying to standardize the format of the data exchange so that someone who's trying to do scope three supply chain and process, what is embodied in the part by the time it is sold. We really want a way for whoever's in possession of that information to make it available to whoever's downstream in receipt of the device and trying to track that as part of the total footprint of what they're doing.**
>
> *— Eric Dahlen, Intel Steering Committee Member, Open Compute Project*

Finally, rapid evaluation of modular platform designs and integration of modular requirements into RFPs will help pivot the market to this commonsense delivery of modern infrastructure. The integration of the CXL standard, which the TechArena has written about extensively this year, is ushering in more flexible system design making modular hardware a practical reality. IT organizations demanding a move to modular systems and resetting refresh rates independently for components and chassis can dramatically lower embedded carbon and help control infrastructure costs.

Strategy 2: Look Beyond Compute Infrastructure and Re-Fresh and Re-Tier Storage

Compute sustainability focuses much of its attention on servers and CPUs, but we'd be remiss in a full analysis of infrastructure without looking at storage. With data center storage expected to grow at over 26% CAGR through 2026 according to Technavio, organizations are deploying storage at an accelerated rate as corporate data stores mount. All of this storage must be powered and cooled impacting IT sustainability.

While storage has never been the predominant energy consumer in the data center, new technology alternatives in the form of QLC NAND powered SSDs deliver a breakthrough in GB/watt offering incredible density through their configurations and bandwidth, latency, and endurance characteristics that position the technology as a shoe in for data hungry workloads from AI/ML data pipelines and analytics to cloud storage and content delivery networks. In fact, according to NREL, SSD technology offers up to 40X the efficiency of spinning disc alternatives and generates far less heat than traditional hard drives placing less stress on cooling infrastructure. When considering sustainable SSD solutions, higher density solutions also pay off in terms of reduced footprint, higher scale-in-place, more efficient utilization of cooling methods, and reduced disposition impact.

Some have argued that the embedded carbon represented in SSDs offset energy savings, and it's true that manufacturing currently consumes approximately double the carbon of hard disk drives when compared for comparable capacity. However, the picture is more complex given the plethora of heavy metals and other toxic compounds that are used in the manufacture of hard disks as well as the longer lifespan offered by SSDs which feature refresh cycles of 10 years on average vs 5 years of average HDD alternatives. SSD vendors are going farther in this regard providing tools for checking SSD longevity while in the field and opening the door for second life applications. Given these considerations, SSDs more than level the sustainability playing field.

- Higher density SSDs from Solidigm enable more data to be stored in less space. The implications for sustainable and affordable data storage from core to edge are enormous, from a smaller physical footprint, to greater power efficiency, to reduced end-of-life disposal impact.

- Greg Matson, Vice President, Data Center Strategic Planning and Marketing, Solidigm

With this technology advancement, the TechArena recommends re-assessing organizational storage tiering and retiring antiquated power-hungry spinning disks in exchange for an all-flash environment featuring modern SSD powered storage arrays. Additionally, within this re-tiering, organizations should consider moving older SSDs to cold tier storage in favor for high performance alternatives. This requires implementation of a reuse plan that involves proper data sanitation to ensure zero increase in attack surfaces for data integrity.

Strategy 3: Re-Engineer Code, Balancing Efficiency and Performance

An often-overlooked area for compute sustainability is the inherent inefficiency of cloud architectures. With redundancy a core tenet of the architecture, developers are driven to land applications on hardware estimating for peak load with several instances of applications running redundantly within a cluster. This inefficiency can result in servers sitting at 20% of performance capacity waiting for peak moments to come and consuming energy for all of that idle time. Advances such as containerization of workloads and serverless computing does help with this inefficiency, but more work must be done.

> **— Maybe the issue here is how we designed cloud compute. We essentially designed it in such a way that it forces people to have to take this sort of wasteful perspective by default.**
> *- Matt Butcher, Founder and Chief Executive Officer, Fermyon*

In fact, it's estimated that up to 20% of all data center energy consumption is rooted in software inefficiency. Not all of this can be explained by cloud computing overhead. Developers have historically focused app development goals on performance and outcome rather than efficient completion of work. Enter green coding, a growing movement within the software development community. Green coding adherents embrace building applications that are carbon efficient, energy efficient, hardware efficient, consume energy at the lowest carbon intensity possible, maximize hardware's energy efficiency, and minimize network traffic. These new design principles can not only change software energy and carbon consumption, but change the requirements of underlying hardware enabling more flexibility in purchase decisions.

TechArena recommends that all organizations begin their efficient code journey by looking at the efficiency of cloud stacks and consider containerization or serverless compute functions to lower total software demand and improve server capacity wherever possible. IT leaders can discuss green code development with vendors and integrate green coding principles into software RFPs. Moreover, these same principles can be integrated into internal application development efforts and made a standard best practice of in-house application development teams.

Strategy 4: Improve Power and Cooling for Modern Infrastructure Demands

We've looked at changes to hardware infrastructure and the software stacks running on hardware. The next target focuses on sustainable power and cooling of data centers. Much work has been done to enable more efficient cooling and power conversion within data centers with advances across power supply technology, consideration of DC powered data centers, hot aisle, cool aisle configurations, hot temperature and humidity data center standards and free air cooling. These advancements have driven down the "overhead" cost for well operated data centers.

However, with the advent of artificial intelligence and deployment of a greater abundance of acceleration technologies such as GPUs, the requirements for advanced cooling technologies have accelerated. Some data center operators are looking to deploy 100-200 KW racks, and it's argued that free air cooling can be effective only to approximately 50 KWs. Many organizations are opting for liquid cooling as an alternative to free air-cooling approaches as heat can be removed directly from silicon and not just dispersed throughout the system or larger data center environment. While every environment is different, liquid cooling is estimated on average to drop total power consumption within a data center by approximately 10%. This drop effects total power, as well as compute power as energy utilized by server fans is reduced or eliminated.

- The biggest thing in the industry right now that I think is going to be really interesting to see where it comes out is liquid cooling. There's some debate on how much this is pushing the efficiency side of things from a cooling perspective. Liquid cooling is taking the cooling directly to the chip itself, so you're not losing in that transfer from the air over to the actual rack.

Sara Martin, Associate Principal and Data Center Market Sector Leader, HED

There is some debate if liquid cooling is taking off because of sustainability or because of the increased power draw and heat of AI clusters, but regardless of what's pushing the trend, data center managers should evaluate liquid cooled solutions vs. traditional air cooling as they gain market share and achieve volume economics especially in dense power per rack configurations. One challenge noted in Andy Bechtolsheim's recent OCP Summit keynote is a lack of standards for liquid cooling technologies limiting their scalable deployment. It's the TechArena's view that liquid cooling alternatives including closed-loop dry systems, open-loop evaporative systems, chilled water systems, and immersion solutions utilizing water, mineral oil and proprietary dielectric solutions must be monitored to see what gains market preference.

For those companies who want to take their pursuit of carbon utilization further, the TechArena recommends following advances in heat re-capture for secondary purposes. This technology is nascent, but the value proposition is clear. We burn fuels today and consume electricity to heat environments, yet data centers produce free heat that, if captured, could provide utility elsewhere. Deep Green is an example of a company investing in this vision and has already demonstrated successful results in capturing GPU heat and repurposing to heat water at an approximate carbon neutral result. While this is a small example of proof of concept, given the opportunity in play and the fact that the largest data center operators have deployed proprietary solutions in this arena, we expect technologies akin to what Deep Green has demonstrated to reach broad market commercial viability soon.

Strategy 5: Measure What Matters and Speak a Standards Based Language

When looking at managing data center sustainability, a foundation of standards-based metrics is critical for consistency in oversight and clarity in communication with vendors and partners. The gold standard for many years has been broad utilization of the Power Utilization Efficiency metric (PUE) which is a simple measurement of total power delivered to a data center divided by power consumed by IT equipment. While PUE is an effective north star to ensure that overhead associated with power conversion and cooling infrastructure is minimized, it provides a flawed outlook assuming all power consumed by IT equipment is well spent. There are also different implementations of PUE across data center operators, with some organizations choosing to implement the ISO standard and others utilizing PPUE or partial PUE statistics. Others are measuring theoretical "best case" PUE vs. operational measurements over time.

To supplement PUE, organizations have added the Carbon Usage Effectiveness (CUE) metric to measure carbon output per unit of energy consumed within a data center. This metric measures the relative energy source carbon output rewarding companies who invest in renewable, clean energy with lower CUE ratings. Organizations also achieve lower CUE scores through investment in more energy efficient infrastructure and investment in carbon offset programs, although the latter approach comes with some controversy of the legitimacy of benefit to true carbon savings.

We've discussed implementations of new infrastructure, stack management, security and network capabilities to unlock the full value of edge computing. Each of these represents new challenges and IT skillsets required for organizational development. Combined, a lack of an IT skillset strategy can thwart even the most well thought out and well-intended edge computing objectives.

- Once these are standardized, these metric values can be exchanged across the value chain. They can actually be used for a comparison or to be able to understand one functionality versus the other in terms of its sustainability impact.

— Shruti Sethi, Azure Storage PM, Sustainability, Microsoft and Sustainability Initiative, Open Compute Project

The water usage effectiveness (WUE) metric has been gaining broader interest with data center operators as CRAC units consume increasingly more water, and liquid cooling technologies have become more prevalent. WUE measures liters of water consumed divided by kwh of IT equipment energy. Organizations seeking to lower WUE scores raise temperature and humidity in the data center based on ASHRAE's expansion of guidelines, utilize recycled water, or increase cycles of concentration in water towers. When you consider that large scale data centers can consume upwards to 5 million gallons of water per day equivalent to consumption of 80,000 households, you understand why WUE and the focus on water conservation is so critically important. This becomes even more urgent when you take into account that some of the hottest markets for new data center buildout, no pun intended, are located in arid areas such as California and Arizona.

- PUE has affected a lot of change for our industry, but we have not got all the way to where we need to be. Carbon use effectiveness (CUE), water use effectiveness (WUE), these things are going to help us to move that additional distance.

— Alex Rakow, Sustainability Lead, Cloud Service Providers, Schneider Electric and Sustainability Initiative, Open Compute Project

Of course, carbon is not the only climate impacting emission in the data center, and that is why global warming potential (GWP) is an important element of any leading-edge measurement practice. GWP provides a common foundation to compare the global warming impact of different gas emissions specifically measuring the impact of one ton of emissions of a targeted gas vs. carbon emissions. For example, methane carries a GWP of 27-30 vs. carbon's 1.0 ranking, meaning that an equivalent ton of methane gas will have 27X impact on global warming over a standard one hundred years. Where do other gas emissions come from? The answer is a variety of sources from combustion power resources in purchased power and embedded within purchased products to operational sources, including refrigerant and backup power generator emissions.

> **- We want to make sure that we're accounting for all of the materials that go into the data center, and into the manufacturing of the devices that go into the data center, that we have basically the whole embodied footprint accounted for. And that's really what GWP is about.**
>
> *— Eric Dahlen, Intel Steering Committee Representative, Open Compute Project*

The TechArena recommends utilizing all of these metrics to build an IT sustainability dashboard and setting sustainability goals based on target metrics aligned with corporate CSR objectives. Additionally, organizations are encouraged to work with vendors through transparent discussions on all metric targets to influence up the supply chain to drive down energy and water utilization and greenhouse gas production.

Strategy 6: Demand Efficient AI Democratization as the Great Sustainability Disruptor

This report has been developed for general purpose data center computing, but that view leaves out an elephant in the room, which is the advancement of artificial intelligence. AI promises to be a disruptive tool across industries to provide the technology needed to fill gaps between carbon neutral goals and carbon negative realities. While we are churning through its hype cycle, we're already seeing vast impact in organizational efficiency, work automation, and resource savings. As we stand at the beginning of the AI era of computing, we also see a troubling reality that the infrastructure used to train AI algorithms is driving an exponential increase in data center power consumption.

An NVIDIA DGXH100 GPU utilizes up to 1.2KW of power, a 1.6X jump from the previous generation DGX A100. When you consider that Microsoft used over 100,000 GPUs to train Chat GPT, you can see how the kwh pile up. In fact, many have attempted to calculate ChatGPT's energy load based on published user data with estimates at up to 23M kwh per month. The result is massive across the data center landscape. Large data center operators are retooling data centers for larger power delivery per rack, placing more pressure on infrastructure, power and cooling systems, and power backup technologies. The race to advance AI also has greenfield data center development with pedal to the metal.

- Right now, the trend is more and more demand for more and more data center space, and a lot of that does come down to AI. If you look at a lot of the hyperscaler projections, which is really the people who are driving the market, they're looking towards doubling, tripling their capacity in the next five to 10 years. As long as all of us are watching Netflix, as long as all of us are scrolling our phones, the market's not going to go backwards anytime soon. Let's put it this way. The more we see where AI lands within corporate America and even just our day-to-day lives, we'll really get a sense of how big this is going to be.

Sara Martin, Associate Principal and Data Center Market Sector Leader, HED

What does this mean for IT operators? Today, supply constraints have limited access to NVIDIA GPUs to only the largest cloud players creating an uneven playing field for AI advancement. Many enterprises also lack the skillsets required to build high performance computing style AI clusters for algorithm training, further widening the gap between large players and mainstream computing.

> **- We have a market that, for all intents and purposes, is dominated by a single player, who's put a lot of work into the whole ecosystem to have the hardware and software to drive everything around AI. Here at Tenstorrent, we think that there's a better way to do it. I think the major thing to think about when you think of Tenstorrent is, we have a solution that is built grounds up for AI. Tenstorrent's mission is inference and training. Everything on the same silicon, same software stack.**
>
> *— David Bennett, Chief Customer Officer, Tenstorrent*

We're sitting in the perfect storm for a Clayton Christensen style disruptive innovation, and the industry and VC community is investing heavily in startups to disrupt not only the single vendor lock on AI training, but also on advancing more efficient compute architectures and math to drive down the resource costs associated with the technology. Companies like Cerebras, Graphcore, Lemurian Labs and Tenstorrent are introducing alternatives to the marketplace.

- We need to re-think, re-imagine accelerated computing for this workload, and we need to make computing more accessible so that more people can come in, more architectures can get trained, and it's not just five companies in the world that have the compute to train this. Because when you have that, the incentive to progress and deliver something that's great isn't there. A lot of the learnings, the failures of these models don't surface. I want to make all the startups out there, all the tier two clouds that are starved for compute, I want for them to have a voice in this and have enough compute to train AI models and deploy them at scale without breaking the planet.

— Jay Diwani, Founder and CEO, Lemurian Labs

Silicon innovation is just the start. Lemurian Labs has invented a new math system for AI to drive efficiency in code and they are not alone in thinking about AI software efficiency and historic framework choices that have made AI a weighty workload. Enter Fermyon, the company that is disrupting web assembly and serverless computing, led by a team that has contributed to the delivery of cloud technologies like Helm and Kubernetes. Fermyon recently released its serverless AI inference solution with a goal of dramatically reducing cost of inference through time slicing. The result is that the GPU is held only for the time needed for workload completion ushering in much more efficient resource utilization. They've partnered with Civo to bring serverless AI services to the world, which is notable as it's access to rare AI inference services and it offers a more sustainable and efficient approach.

So what does the TechArena recommend an enterprise to do in the face of massive change? We'll be writing more extensively on AI in the near future, but for this report on sustainability, we recommend evaluating AI silicon alternatives for both training and inference. This includes Arm and RISC-V alternatives, x86 for inference where the vast majority of workload deployments focus today, and new acceleration technologies from the startup community. Secondly, we recommend integrating sustainability in organizational growing skillsets in AI workloads. A key point across interviews is that AI is rapidly moving from a specialty area for software development to an arena where all software developers will engage. Taking the broader green coding principles outlined earlier in this report and applying to AI is essential for all organizations. Finally, IT organizations should demand a sustainable delivery of AI cloud instances from vendors and build these principles into RFPs. This may feel like it flies in the face of trends of higher energy consumption and massive performance requirements, but calling for sustainable approaches is one of the strongest influence points to drive cloud providers to think differently.

Strategy 7: Look Around Corners to Plan for a Sustainable Future

There is a truth in computing circles that different sectors of the industry aim their strategic planning at different timeframes. Cloud operators and DevOps teams thrive on a culture of nimble and rapid deployments, which focus on near term feedback and rapid iteration. Silicon operates in opposition with product planning forecasted five years and beyond to align with the tooling required to deliver chips to market. While CPU manufacturers have worked for years to pull in these cycles to deliver more nimbly to customers, the requirements of sustainability is requiring the industry collectively to embrace a longer range view to peer around corners. This requires building new processes, approaches, and in reality, entire functions for some organizations.

IT organizations seeking to transform operations to a sustainable future and move to carbon neutral and negative will need to embrace a similar approach. Oracle is one company who recently took this path, establishing an organization to forecast future challenges like sustainability in order to meet customer requirements before customers have even considered that they have them. This approach is letting Oracle more accurately scope future power, cooling, infrastructure and software trends to best meet their customer demands.

> **- Some of the questions I'm looking at are questions of resilience, questions of sustainability, questions of energy utilization. How do we shrink our footprint but still provide the capability and scale our customers require to continue moving forward?**
>
> *— Bev Crair, Senior Vice President, Enterprise Intelligence and Resiliency, Oracle*

The TechArena advises first evaluating corporate culture to determine where on the time continuum the organization stands for strategic long-range planning. This can be determined based on other organizational challenges present in industry dynamics. Next, if an organization lacks long term planning and pathfinding functions, it's recommended to found a team to begin forecasting macro trends across IT sustainability topics including energy prices, greenfield and brownfield data center trends, and compute demand. Armed with this data, IT leaders will be best able to accurately target capacity, infrastructure, software, and operational technologies that deliver a long-term sustainable future. Moreover, organizations will be able to use IT as the strategic tool required to innovate towards broader sustainable CSR goals.

Summary

After speaking to dozens of industry experts and leading thinkers in academia and research, the TechArena has concluded that while there is incredible work to be done to drive true compute sustainability, there are efforts with enough industry might to effect real change in the years ahead. Hyperscalers hungry for AI supremacy will invest in any technology possible to deliver more efficient AI training and inference, and this investment will trickle down to broader market impact. Transparency in embodied carbon and new metrics will force deeper changes across industry players. Enterprises running on-prem data centers stand to benefit from these efforts.

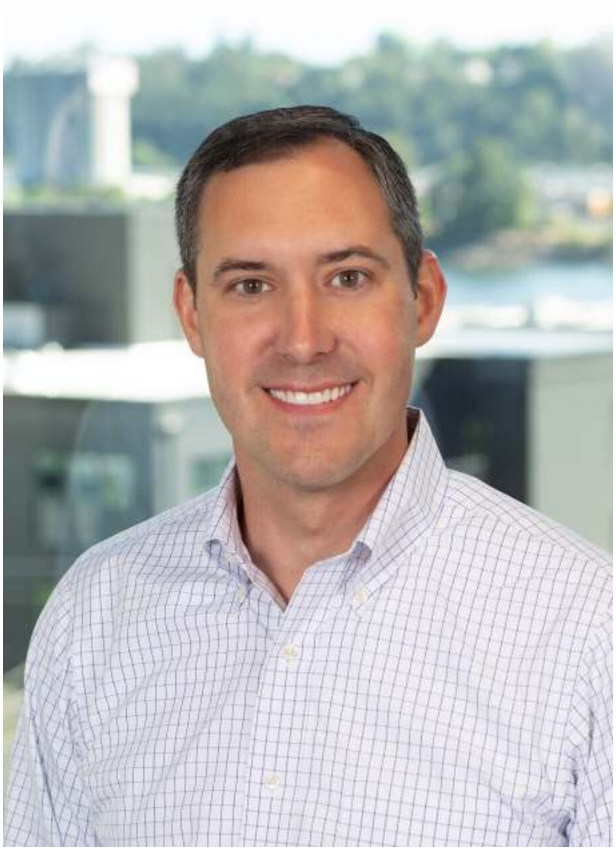
The important lesson to me at the highest level is that IT is a key enabling technology that has broad and deep implications for changes in the economy as a whole. The investment in IT's consumption of three percent of the world's electricity allows us to optimize the other ninety-seven percent of electricity and all the fuels consumed around the planet. That gives the sense of the kind of lever that IT is for solving the climate problem.

— Jonathan Koomey, President and Founder, Koomey Analytics

Industry Experts

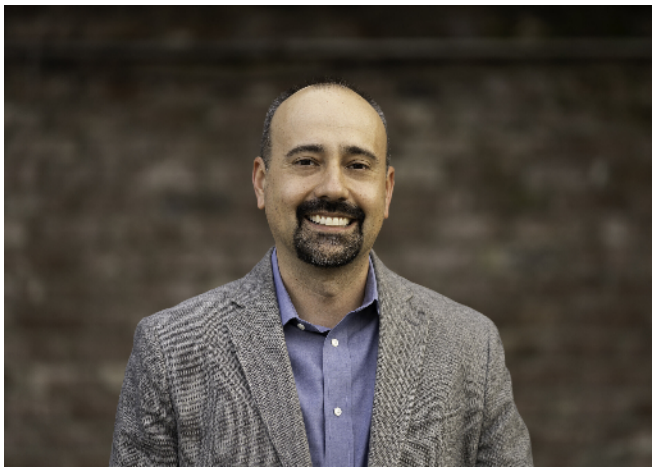
Robert Hormuth is Corporate Vice President, Architecture and Strategy of the Data Center Solutions Group (DSG) at AMD. Robert has 35 years in the computer industry, joining AMD in 2020 after 13 years with Dell where he was CTO of the Server Business unit, 8 years with Intel and 11 years at National Instruments. At AMD Robert is charged with creating a long-term system level vision for DSG and identify the technical requirements/implications to the DSG portfolio. Robert has a B.S. in Electrical and Computer Engineering from The University of Texas at Austin and currently holds 30+ patents.

Listen to Robert's TechArena Interview



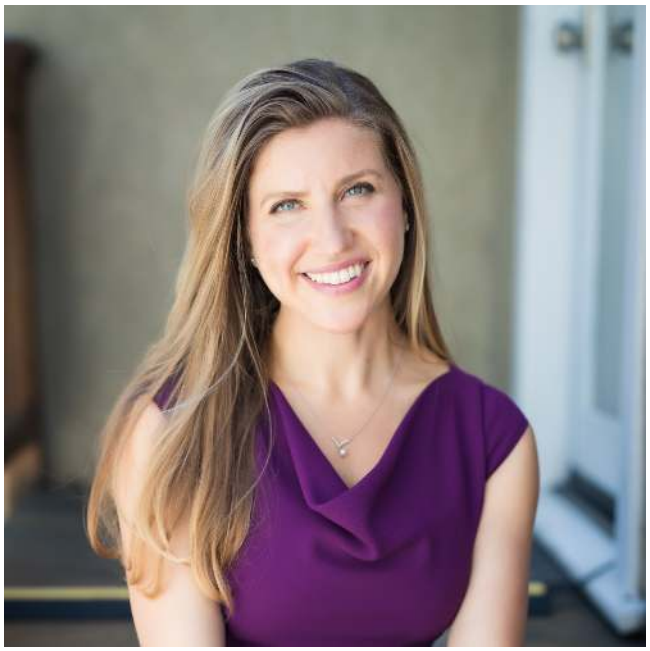
Jeff Wittich is the chief product officer at Ampere. Jeff has extensive leadership experience in the semiconductor industry in roles ranging from product and process development to business strategy to marketing. Prior to joining Ampere, he worked at Intel for 15 years where he was responsible for the Cloud Service Provider Platform business and the product development team responsible for 5 generations of Xeon processors.

Listen to Jeff's TechArena Interview



Eddie Ramirez is Vice President of Marketing and Ecosystem Development, Infrastructure Line of Business, for Arm. Eddie leads a Go-to-Market team at Arm responsible for helping partners innovate and grow through the adoption of Arm-based solutions into data center, networking and edge markets. He also manages a team responsible for ecosystem development activities and is focused on creating a rich and vibrant ecosystem of hardware and software partners. Prior to Arm, Eddie held various executive leadership roles in product management, segment marketing and applications engineering. He has over 20 years of experience in storage and networking having had successful campaigns at AMD, Marvell, Sandforce/LSI, Seagate and Western Digital. Eddie has bachelor's degrees in both Management Science and Electrical Engineering from Massachusetts Institute of Technology (MIT).

Listen to Eddie's TechArena Interview

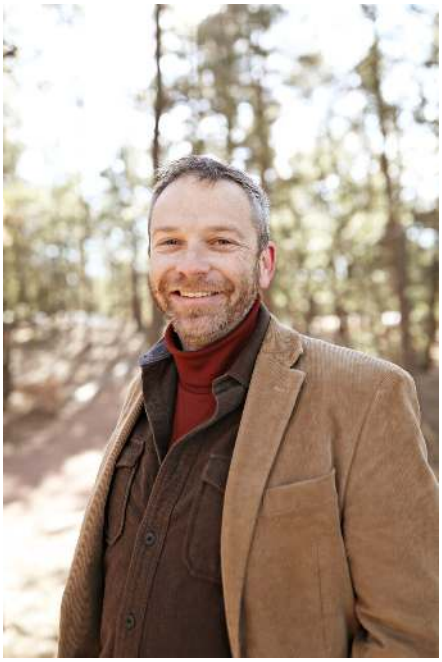


Rebecca Weekly is the Vice President of Hardware Systems Engineering at Cloudflare leading the team that delivers >18% of the world's Internet traffic. Rebecca is on Fortune's 40 Under 40 2020 technology list, and is on Business Insider's Cloudverse100. In her "spare" time, she is the lead singer of the funk and soul band, Sinister Dexter. She has two amazing little boys, and loves to run .

Listen to Rebecca's TechArena Interview

Ian Monroe is Etho Capital's President and Chief Investment Officer, as well as a lecturer on climate solutions and sustainable investing at Stanford University. Prior to Etho Capital, Ian founded Oroeco, an award-winning web and mobile platform that rewards users for helping solve climate change. Ian has also served as an advisor for several international sustainability certification standards, including Climate Neutral, Science Based Targets, and The Roundtable on Sustainable Biomaterials (RSB), and he has been awarded as an Echoing Green Climate Fellow, an Institutional Investor Rising Star, and one of TechRepublic's top 8 leaders in cleantech. Ian has consulted and spoken about scaling sustainability solutions for the United Nations, World Bank, the U.S. Department of State, and a wide range of companies and nonprofits, drawing inspiration from his research at Stanford, his experiences working on supply chains and sustainable technology in over two dozen countries around the world, and his family's small farm in California.

Listen to Ian's TechArena Interview

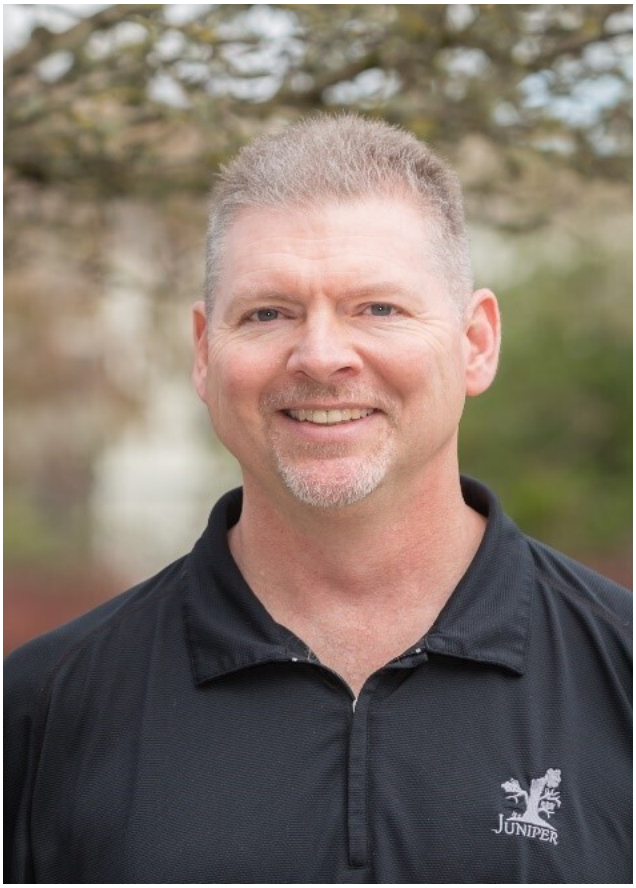


Matt Butcher is co-founder and CEO of Fermyon, the serverless WebAssembly company in the cloud. He is one of the original creators of Helm, Brigade, CNAB, OAM, Glide, and Krustlet. He has written and co-written many books, including "Learning Helm" and "Go in Practice." He is a co-creator of the "Illustrated Children's Guide to Kubernetes" series. He holds a Ph.D. in Philosophy.

Listen to Matt's TechArena Interview

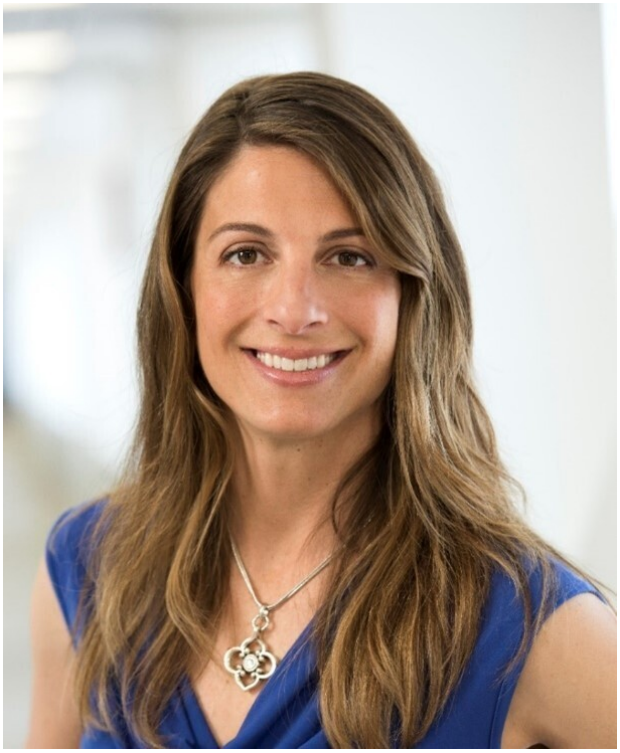
Sara Martin is an Associate Principal at HED, one of the oldest and largest architecture and engineering firms in the country, with 400+ employees and annual revenues above $100M. With nearly a decade's experience, Sara is an expert in delivering optimized, replicable, and HED's resilient facilities for mission critical clients, and she plays a key role in the success of some of most high-profile data center projects, programs, and key confidential client relationships. Sara is also an active member of the 7x24 International community and serves as Vice President of the 7x24 New England Chapter.

Listen to Sara's TechArena Interview



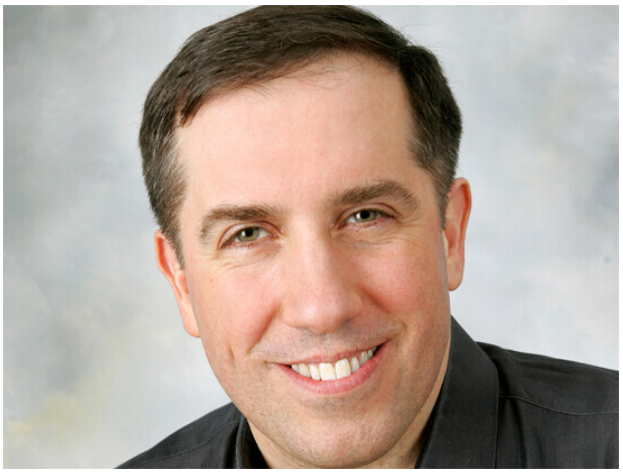
Eric Dahlen has over 30 years experience in server component development, most of it spent on chipset projects and their associated technologies. He is Intel's OCP steering committee rep for sustainability.

Listen to Eric's TechArena Interview



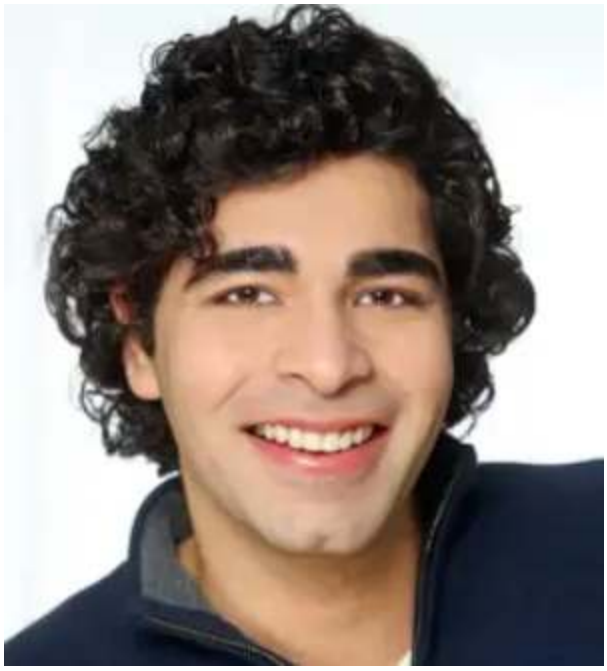
Jen M. Huffstetler is Chief Product Sustainability Officer and Vice President/General Manager of Intel Future Platform Strategy. In this role, she is responsible for driving the integration and execution of the corporate-level product strategy to drive future growth across client, cloud, network and edge to deliver sustainable computing for a sustainable future. This work is built upon Intel's industry-leading sustainable semiconductor manufacturing. Huffstetler brings a diversity of business, marketing and engineering experience from her 20+ years at Intel spanning semiconductor manufacturing, client, network and datacenter product management across CPUs, GPUs, DIMMs, and systems. Huffstetler holds a bachelor's degree in chemical engineering from the Massachusetts Institute of Technology, and an MBA from Babson College, F.W. Olin Graduate School in Corporate Entrepreneurship. She is also a certified Executive Leadership Coach from Hudson Institute, Korn-Ferry Interpreter, and Birkman Certified.

Listen to Jen's TechArena Interview



Jonathan Koomey is a researcher, author, lecturer, and entrepreneur who is one of the leading international experts on the economics of climate solutions and the energy and environmental effects of information technology. Dr. Koomey was a lecturer in Earth Systems, School of Earth, Energy, & Environmental Sciences at Stanford University from November 2016 to June 2018, and for four years before that he was a Research Fellow at Stanford's Steyer-Taylor Center for Energy Policy and Finance. He has also held visiting professorships at Yale University (Fall 2009), Stanford University (2003-4 and Fall 2008), and the University of California, Berkeley's Energy and Resources Group (Fall 2011). He was a Lecturer in Management at Stanford's Graduate School of Business in Spring 2013. For more than eleven years he led Lawrence Berkeley National Laboratory's (LBNL's) End-Use Forecasting group, which analyzed markets for efficient products and technologies for improving the energy and environmental aspects of those products. The group developed recommendations for policymakers at the U.S. Environmental Protection Agency and the U.S. Department of Energy on ways to promote energy efficiency and prevent pollution. Koomey is also a Research Affiliate of the Energy and Resources Group at the University of California, Berkeley. Dr. Koomey holds M.S. and Ph.D. degrees from the Energy and Resources Group at the University of California at Berkeley, and an A.B. in History and Science from Harvard University. He is the author or coauthor of ten books and more than two hundred articles and reports on energy efficiency and supply-side power technologies, energy economics, energy policy, environmental externalities, and global climate change. He has also published extensively on critical thinking skills.

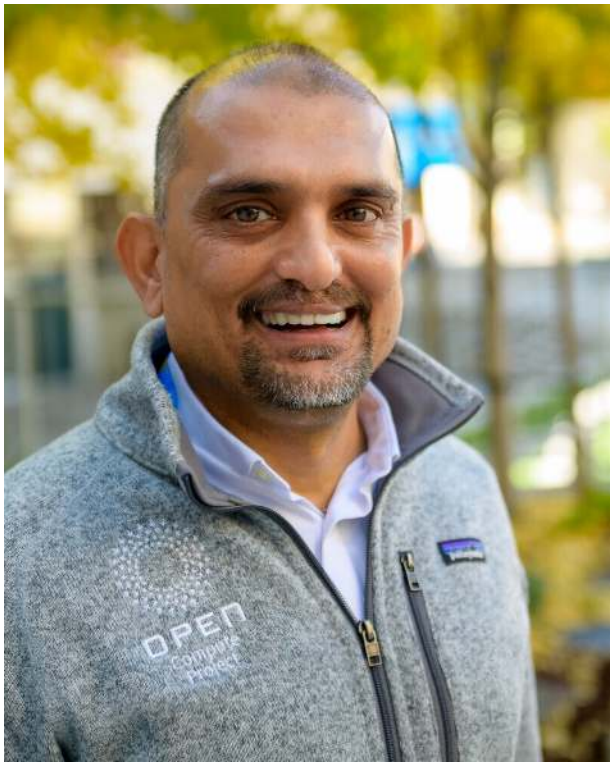
Listen to Jonathan's TechArena Interview



Jay Dawani is the Co-Founder & CEO at Lemurian Labs, a technology startup that specializes in artificial intelligence (AI) and machine learning (ML). Jay has over 10 years of experience in the AI/ML field, and has held previous positions as an AI Advisor at NASA's Frontier Development Lab, an author for Packt Publishing, and a speaker at the Conference on Complex Systems. In addition to their work in the AI/ML field, Jay also serves as a Technology Advisor for the SiaClassic Foundation. Jay Dawani holds a Bachelor's Degree in Applied Mathematics from Western University. Jay also has a certification from Coursera in Deep Learning Specialization.

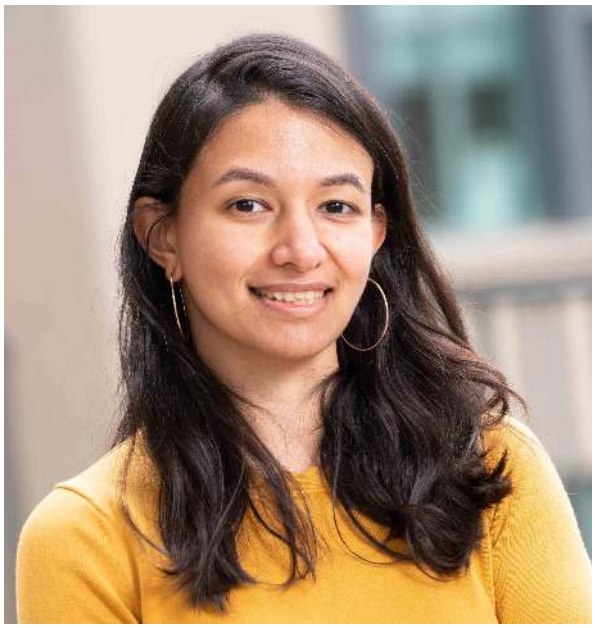
Listen to Jay's TechArena Interview.

.



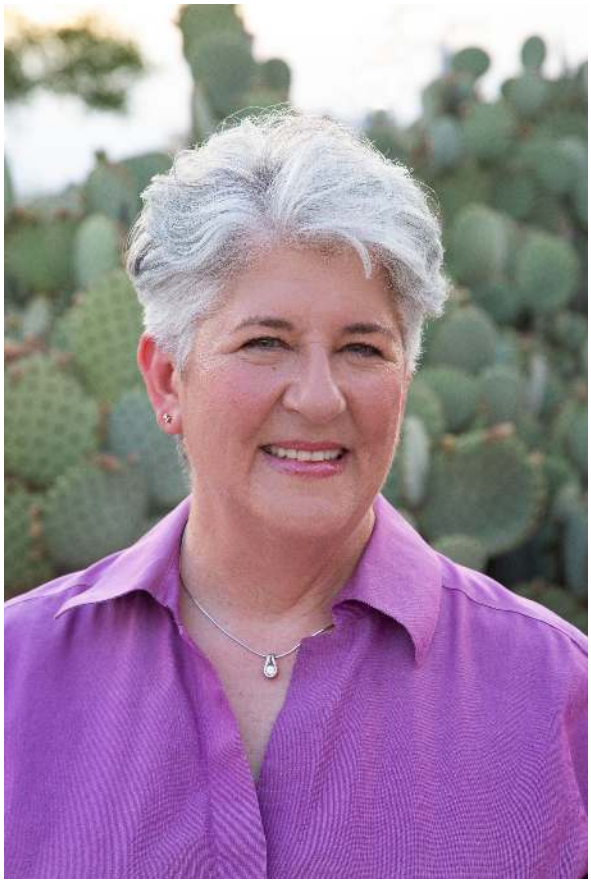
Dharmesh Jani ('DJ') has been an active contributor to the Open Compute Project (OCP) since 2012. At Meta, DJ has led the infrastructure ecosystem and partnerships for the past 5 years. He is responsible for leading OCP and other open technologies, working with stakeholders inside and outside the company. DJ currently is Chair of OCP Incubation Committee and on the board of directors for the Universal Chiplet Interface Express (UCIe) consortium. DJ led the Sustainability Initiative at OCP starting in in 2019, conceiving, championing, and launching the effort. He led the initiative in its first two years, helping it grow into a full OCP project in 2022. DJ also launched the chiplet initiative in OCP as the Open Domain Specific Architecture (ODSA). Over his 20+ year career, DJ has held leadership roles in engineering, product management, and business strategy roles at 4 startups and 3 Fortune 500 companies such as Corvis Systems, Infinera, Meta and Intel

Listen to DJ's TechArena Interview



Shruti Sethi is the Steering Committee Rep for OCP – Sustainability Project, passionate about driving efficiency and sustainability in Data Center technology. Shruti has deep experience in both computing and storage systems. She has industry experience working extensively on Graphics power management, workload management, setting performance targets and Data Center storage hardware. She is currently most vested in the intersection of Data Center Technology and Sustainability, driving the next wave of initiatives in this domain for AZURE STORAGE.

Listen to Shruti's TechArena Interview



Bev Crair is the senior vice president of enterprise intelligence and resiliency for Oracle Cloud Infrastructure. Prior to coming to Oracle, she was the Vice President responsible for all product development for Lenovo's Data Center Group. Bev specializes in organizational design, development and dynamics with a focus on building high performance teams, business strategy and planning, new technology development and implementation.

Listen to Bev's TechArena Interview

Alex Rakow co-leads the OCP Sustainability Project. He is the Sustainability Lead for the Cloud & Service Provider segment at Schneider Electric, where he works with data center operators to advance their sustainability and energy performance. He has 12 years of experience working on sustainability and energy management with both private and public clients, including the National Park Service and the Environmental Protection Agency. He is the author of two books: Energy Resilient Buildings and Communities from CRC Press, and Powering Through: Energy Resilience from Grid to Government from Routledge Press. He has a BS from Cornell University in Environmental Science, and an MA from Johns Hopkins University. Alex is a Certified Energy Manager, and is a Working Group Chair for the iMasons Climate Accord.

Listen to Alex's TechArena Interview

Greg Matson is the Vice President of Strategic Planning and Marketing at Solidigm. With over 20 years experience in the technology arena including leadership roles at Intel, Greg is leading Solidigm's market engagement leveraging the company's heritage in storage innovation.

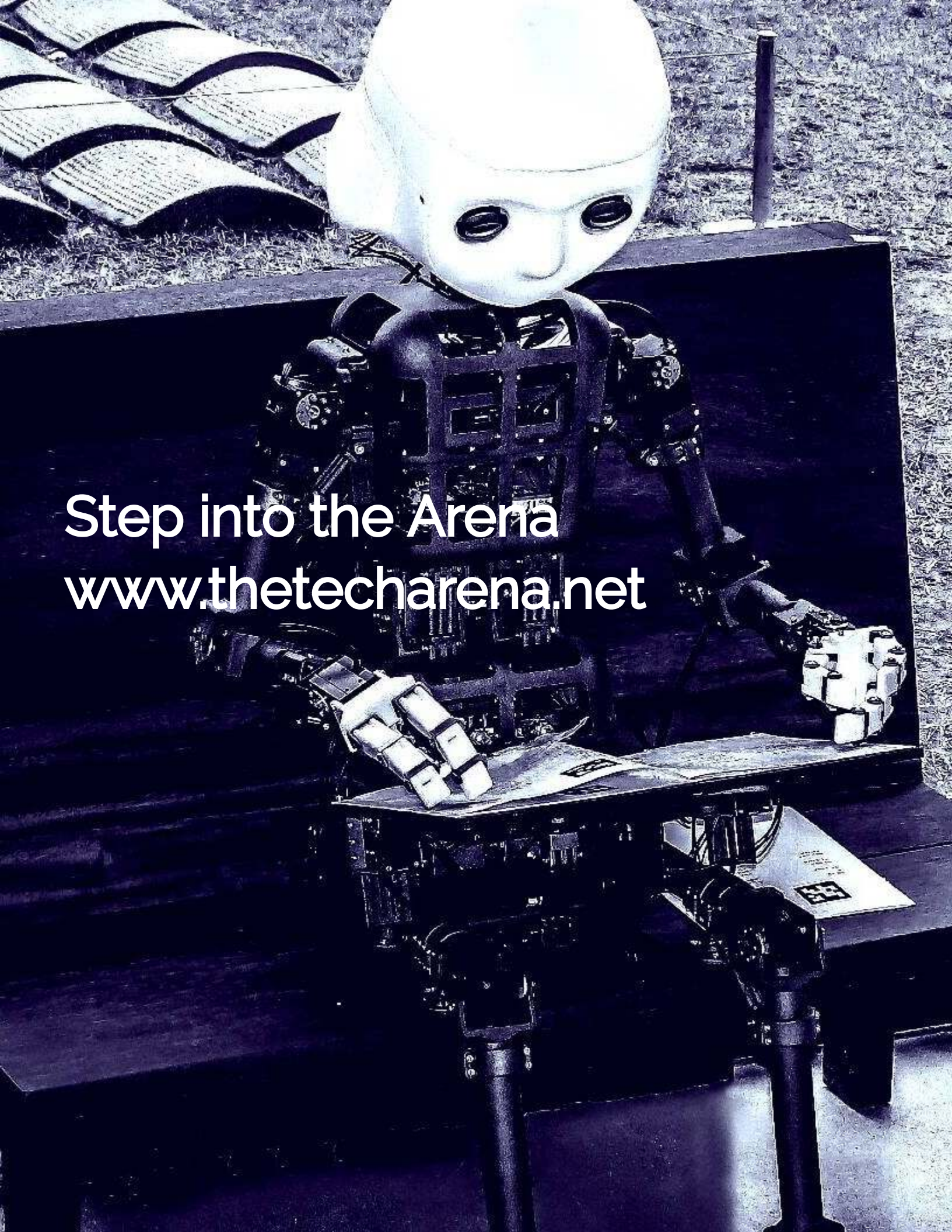
Listen to Solidigm's TechArena Interview

David Bennett is the Chief Customer Office (CCO) of Tenstorrent Inc. Tenstorrent is a North American based AI, Machine Learning, and RISC-V hardware and software startup. David has over 15 years of experience in the technology industry with executive leadership positions across multiple geographies in critical global and regional roles. Before this role, David was CEO of NEC Personal Computers, President and representative director of Lenovo Japan, and a vice president at Lenovo Group. He was responsible for all of Lenovo Japan and NEC's operations, growth, and profitability for all aspects of the business in the country. He is also a visiting lecturer at Yamagata University and a Member of the Board of Sanrio Company Ltd (The Hello Kitty Company), a $2B global company with some of the most recognized IP in the world. Before joining NEC/Lenovo, David was AMD's Corporate Vice President and GM of OEM Accounts Worldwide, including businesses that span consumer, commercial, graphics, and enterprise platforms. Before that, he was Mega Region Vice President for the Asia Pacific and Japan region, worldwide Sales GM for HP and Lenovo, and GM of AMD's Canadian business.

Listen to David's TechArena Interview

Allyson Klein's insatiable curiosity for what inspires innovation drove her to found her first podcast over a decade ago. She has reached over ten million listeners over one thousand interviews with the industry's brightest stars. Allyson created the TechArena to shine a light on leading edge innovation featuring inspirational stories from the industry and under-recognized voices. Allyson leads the TechArena from Portland, Oregon working with both tech titans and breakout stars. Formerly, Allyson served as vice president of global communications and marketing at Micron and general manager of data center, edge, 5G, and AI marketing at Intel. She'd love to connect to chat about the TechArena, the Oregon Ducks, and all things tech.

Check out Allyson's TechArena publications.


Step into the Arena
www.thetecharena.net