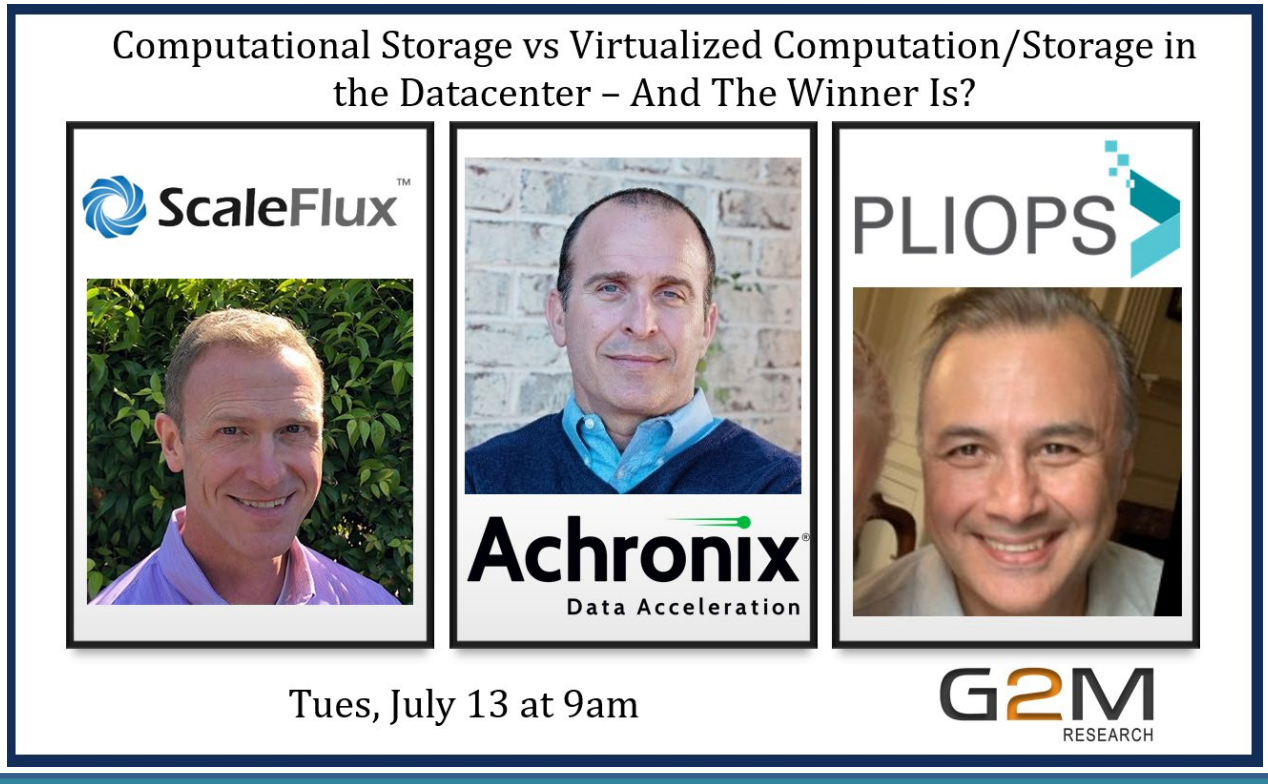
Highlights

At the end of each newsletter, we provide a list of events, conferences, tradeshows that are coming up in the next month or two. It was a struggle to find much to list this time around. And, even with the return of some conferences, most companies are electing to keep them virtual for another go-around. The uncertainty and concerns regarding outbreaks of the virus due to new strains and/or a rapid reopening as well as the differences in how countries are handling travel is making it challenging for event planners to commit time, energy, and expense to traditional conferences and other in-person networking. Many companies are simply not committing one way or the other yet, with a “check back later” on their website. We are looking forward to connecting in-person soon but can also appreciate this conservative approach, particularly if it means once things fully open, they do not shut down again.

Our 2022 webinar schedule will be posted soon. Let us know of any topics you would like to see and we are always willing to host custom webinars too.

Cheers! Mike Heumann



HPE - Will Revenue Catch Up with These Smart Purchases?

Zerto

Hewlett Packard Enterprise (HPE) is purchasing Zerto for $374M in a bid to stay competitive with Dell. The question is whether these smart purchases will curtail the drip, drip, drip of revenue decline that HPE's storage business has been experiencing. Q1 of this year, HPE showed storage revenues up 5% from the prior year, after 6 successive quarters of decline. The Zerto acquisition is the fifth of a storage/storage networking company by HPE in the last 10 years. HPE recently acquired HyperX, the gaming division of Kingston Technology, for $425M. In March 2017 HPE paid $1B cash for Nimble Storage, Inc, in addition to assuming or paying out unvested equity awards valued at $200M. Before that, HPE acquired SGI in August 2015 for $275M, and 3PAR in September 2010 for $2.35B.

Zerto's software and tools help organizations recover in minutes from cyberattack, ransomware, and other unplanned disruptions bringing data back to the original state just a second before the attack or disruption. Hewlett Packard noted that Zerto easily replicates and migrates data between VMware's VMW vSphere and Microsoft's MSFT Hyper-V environments and natively to Amazon's AMZN Amazon Web Services and Microsoft Azure.

We featured, Andy Fernandez, Senior Product Marketing Manager for Zerto, and his predictions for 2021 in our March Enterprise Storage Newsletter. The future, a beginning, the snake eating its tail.

Micron Nailed 2021 Predictions

Micron; Raj Hazra, Senior Vice President & General Manager of the Compute and Networking Business Unit, gave his predictions for 2021 to Storage Newsletter, as follows. Seems like he nailed much, if not all, of it.

In 2021, the prevalence of remote work – even post-pandemic – will continue accelerating capabilities in the cloud. Companies will look to create preparedness for a new normal whether it be more IT solutions for a flexible workforce, larger data stores to fuel continued growth of online commerce, or resilient IT systems to address any future health care crises. This will drive unprecedented demand for agile IT infrastructure, multi-cloud solutions and pervasive connectivity to power edge-to-cloud use cases. While we see great opportunity for memory and storage to fuel increasingly data-centric cloud services, we will also see a rise in data center operators evaluating disaggregated, composable systems to better scale for coming enterprise demands and data growth.

Boundaries between memory and storage will blur: 2021 is going to see AI-as-a-service become mainstream, intelligence migrate to the edge, and 5G come to life. This is going to propel fundamental changes in the way server systems are architected. Memory will extend into multiple infrastructure pools – and will become a shared resource. And the lines between storage and memory will blur. You'll no longer think "*DRAM for memory and NAND for storage*". Instead, faster NAND will create the ability to use it as memory, and applications will grow in their sophistication to utilize resourcing in innovative ways. In 2021, we'll also see enterprises seeking new kinds of solutions such as storage-class memory and memory virtualization to further unlock the value of AI and exploding volumes of data.

More pressure for an energy-efficient cloud: The move toward composable infrastructure will be critical in reducing over-provisioned resources, and thus, mitigating the rising environmental impact of IT. Information and communication technology is already predicted to use 20% of the world's electricity by 2030. As companies look to incorporate sustainability into business strategy and reduce Opex for compute-intensive workloads such as AI and high-performance computing, we'll see escalating demand for energy-efficient architectures, enabled by composable infrastructure.

AI will become more accurate and more ubiquitous, and, we'll start to see it filling in more gaps for simple tasks where people would traditionally say "*Why would I ever use an AI algorithm for that?*" For example, grocery stores might tap AI-enabled cameras that periodically check to see if a shelf is empty and if so, alert a clerk to restock. In a post-Covid world, we'll see more businesses adopting AI for use

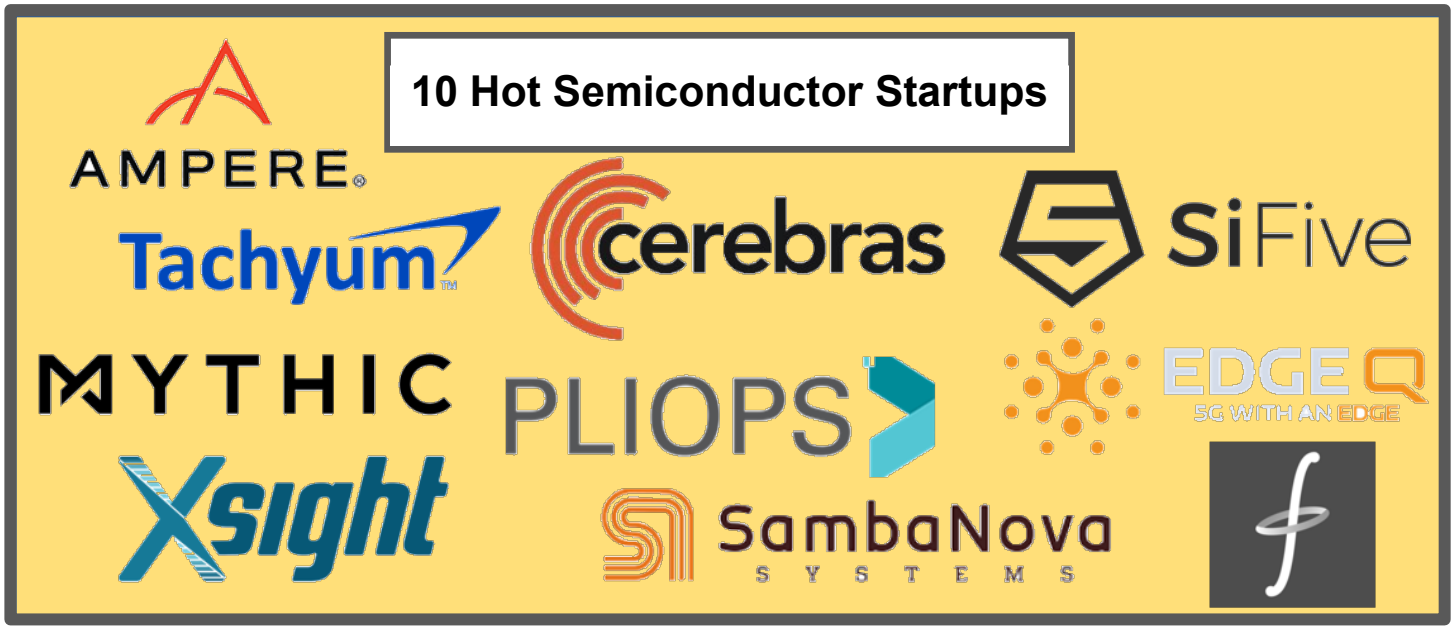
cases like these to create these contactless experiences. We'll also see AI moving into infrastructure such as data centers and telecom base stations as neural network algorithms become more adept at workload and system error correction and recovery.

The rise of edge data centers: There are lots of startups that are focused on building edge data centers that look like transport containers that sit in metro areas to enable content – like your Hulu videos – to be closer to the consumption. We'll see the adoption of these edge data centers in the next few years, as enterprises and consumers look to tap massive amounts of data for insight and faster services closer to the source.

High-bandwidth solutions for high-compute at the edge are becoming a requirement. With fully autonomous solutions, the amount of compute performance needed for cars is reaching data center levels; in ADAS and autonomous driving, cars need hundreds of tera operations per second. This is some of the highest levels of performance in the industry today, rivaling what you find in data centers.

Given this, in 2021, we can expect to see embedded players increasingly turning to creative options for low-power, high-bandwidth memory and storage. For instance, requirements are exceeding capabilities of standard PC DRAM and low-power DRAM, and instead driving the need for capabilities of graphics memory like GDDR6 and or HBM. We'll see these increasingly adopted in cars which need fast, high-performance memory.

In 2021, look for more usage of object stores, for storing structured and unstructured data, files, blocks, objects – all in one repository. AI's large data sets need proximity to where processing is happening. So, rather than viewing it as a large cold store, object stores are going to be able to do AI-type workloads, which means large sets of data can be accessed with high bandwidth and low latency. As a result of the rise of data-intensive AI workloads, we'll see the need for high-performance NVMe storage also increase, since this high-performing object store resides on flash-based storage, as opposed to the traditional cold storage. This could lead to faster adoption of Gen4 NVMe SSDs to enable more powerful object store for AI.



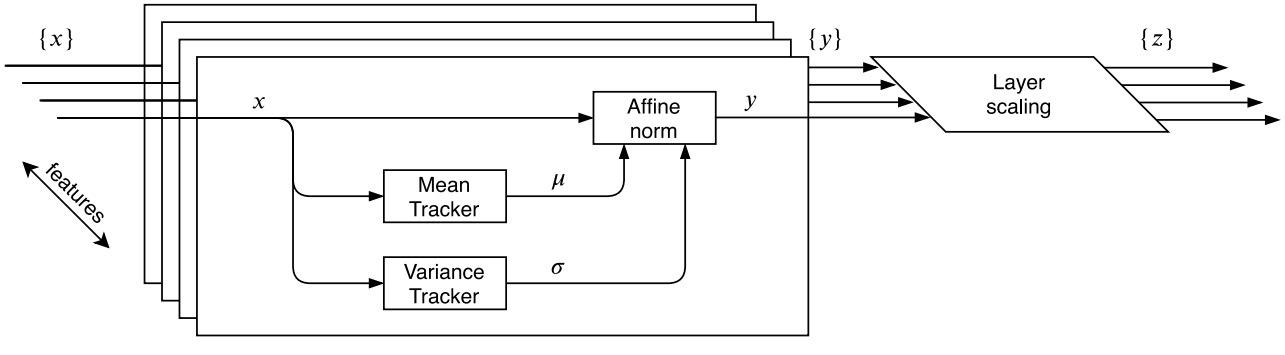
CRN picks for the "10 Hottest Semiconductor Startups of 2021 (So Far)" are Ampere Computing, Cerebras Systems, EdgeQ, Fungible, Mythic, Pliops, SambaNova Systems, SiFive, Tachyum, and XSights Labs. We take a look at each awardee below.

Ampere Computing; Renee James, CEO

> Ampere is moving toward a full custom microarchitecture core design from the ground up, in their view to achieve better performance and better power efficiency in datacenter workloads compared to Arm's Neoverse "more general purpose" designs. Ampere's move away from reliance on Arm's next-gen cores and reliance on their own design show incredible confidence in their custom design. Ampere's Altra processor shipping now has 80 cores and operates on much less power per core than rival Intel and AMD chips. The Altra Max processor has 128 cores and is going to ship later this year. And, Ampere's next-generation processor is projected to be sampling on a 5-nanometer manufacturing process (where the width between circuits is 5 billionths of a meter) in the first half of 2022.

Cerebras Systems; Andrew Feldman, CEO

> Cerebras Systems boasts greater compute density, more fast memory, and higher bandwidth interconnect than any other datacenter AI solution along with space efficiency and the simplicity of using a single device:
>
> 850,000 AI-Optimized Cores (123x more)
>
> 40 GB On Chip SRam (1000x more)
>
> 220 Pb/s Interconnect Bandwidth (45,0000x more)
>
> 20 OB/s Memory Bandwidth (12,800 more)



Of particular interest, is their commitment to correcting errors related to lack of, and/or, errors in normalization for training neural networks. Online Normalization uses moving average statistics on the forward and backward pass and adds layer scaling (Figure 1) to guard against the effects of errors in the statistical estimates. Layer scaling divides out the root mean square (RMS) of the activation vector across all features to prevent exponential growth of activation magnitudes.

Activation Clamping is an improvement over the original layer scaling that performs equally well with the added benefit of being less computationally expensive.

Figure 1: Online Normalization with layer scaling. The incoming feature is represented as x , μ and σ are the moving average mean and standard deviation, and y is the normalized feature. The feature z is the output of Online Normalization. Activation vectors across all features are represented by $\{x\}$, $\{y\}$, and $\{z\}$. Layer scaling introduces a cross-feature dependence by dividing out the RMS of $\{y\}$. Trainable bias and scale are excluded for simplicity.

Layer scaling helps stabilize training by eliminating the compounding of estimation errors. Left unchecked, these estimation errors can lead to the exponential growth of activation magnitudes across layers [1]. We propose simply clamping activations to stabilize training. Activation clamping can be expressed as:

$$z = \text{clamp}(y; c) = \min(\max(-c, y), c)$$

where activations are constrained to the range . A statistically motivated setting for the clamping hyperparameter can be argued given the definition of normalization. The output of the affine norm y should be zero mean unit variance. Assuming a Gaussian distribution, the chances of activations being outside of a few standard deviations shrink at the rate of the complementary error function.[1]

Online Normalization with activation clamping is depicted in Figure 2. If the statistical estimates of Online Normalization are accurate, clamping does nothing to the activation; clamping only modifies activations when there is a large error in the statistical estimates. Furthermore, as the network asymptotically nears convergence and the learning rate is annealed, the error in the statistical estimates approaches zero. For inference, clamping can be removed from Online Normalization.

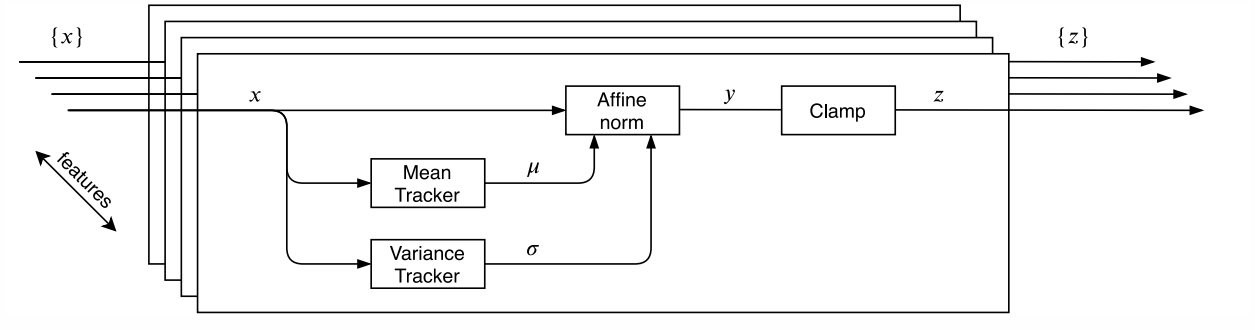
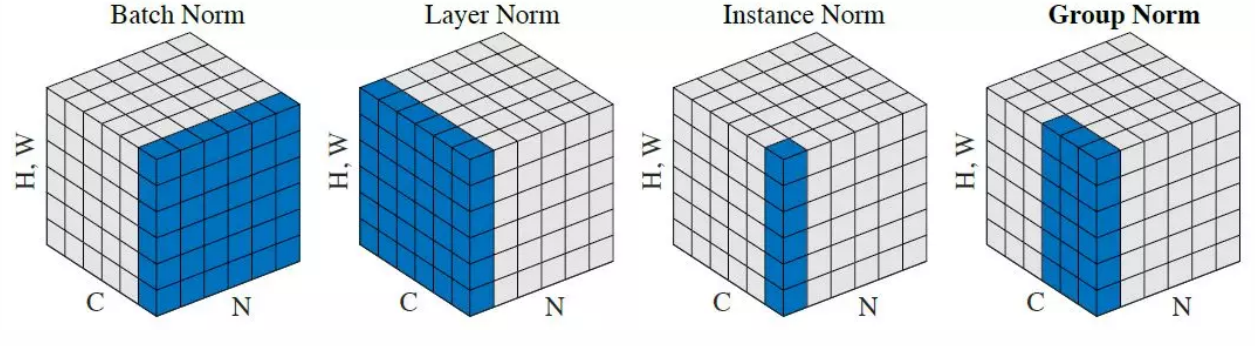


Figure 2: Online Normalization with activation clamping. Adaptive bias and gain excluded for simplicity.



While ML practitioners have differing ideas about normalization, it is generally undisputed that it does, in fact, accelerate neural network training. Normalization, as defined by [1], is a process that z-scores data by subtracting out the distribution mean and dividing out the distribution standard deviation. Without normalization neural networks are functions of their inputs.

EdgeQ, Vinay Ravuri, CEO

> "Our vision at EdgeQ has always been about implementing 5G in a format that is accessible, consumable, and intuitive for our customers. EdgeQ is not only the first company to converge both 5G and AI on a single chip for wireless infrastructure, but we are also able to make those capabilities available in a SaaS model. This fundamentally reduces the initial capex investment required for 5G, thereby removing both technical and economic barriers of 5G adaptation at greenfield enterprises," said Vinay Ravuri, CEO and Founder, EdgeQ. "This pay-as-you-go model ensures that the evolving demands of the market can leverage the full fluidity and elasticity of EdgeQ's 5G-as-a-Service product."

Fungible; Pradeep Sindhu, CEO

> The Fungible Data Center process to pool and deploy resources with greater flexibility:
>
> > Hyperdisaggregate your infrastructure into fluid pools of storage, compute (CPU, GPU) and network enabled by the Fungible DPU.
> >
> > Compose or recompose bare-metal servers with required compute, storage, and network on the fly.
> >
> > Deploy your favorite applications using templates from the marketplace.
> >
> > Consolidate and host your datacenter applications on the same platform.

Mythic; Mike Henry, CEO

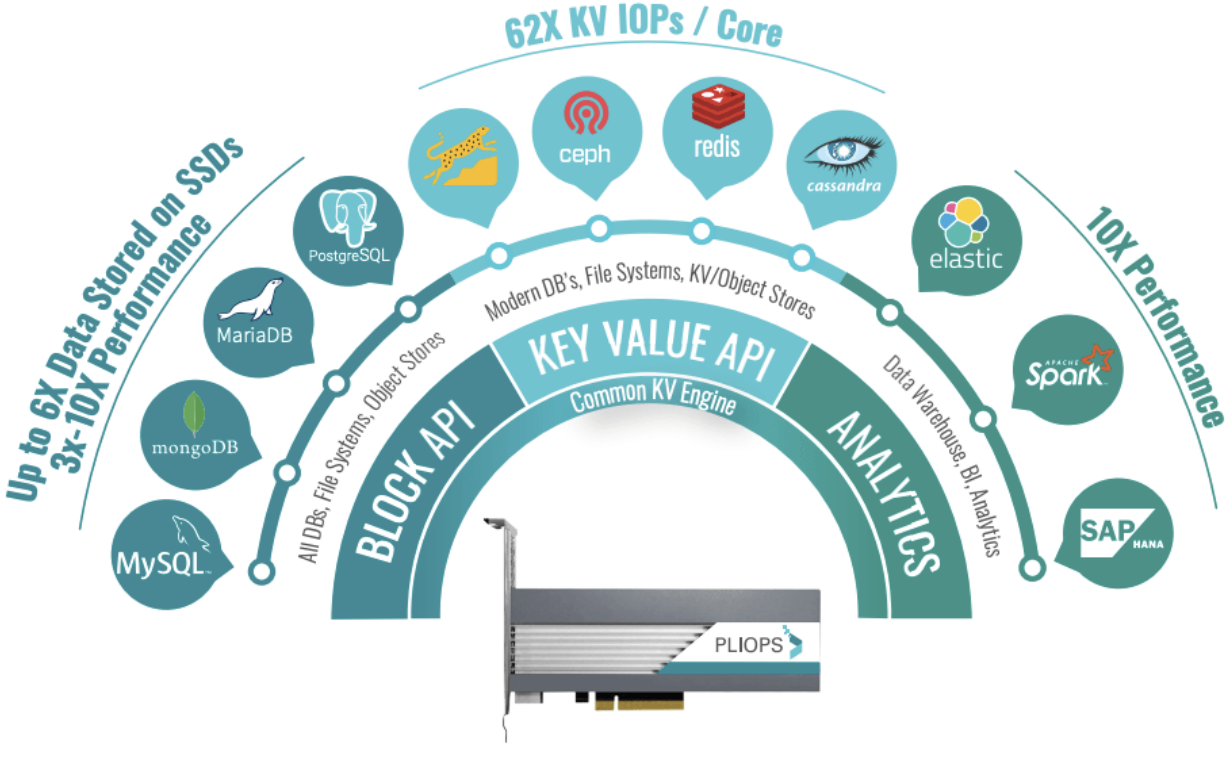
Mythic's approach centers on solving some of the basic, but significant, obstacles to AI use today:

> Mythic Analog Matrix Processors (Mythic AMP™) offer huge advantages in power, performance, and cost. They lower the barriers to innovation, making it vastly easier and more cost-effective to create powerful Edge AI solutions. Mythic AMPs leverage analog computing by performing the calculations required for inference of deep neural networks inside a dense flash-memory array. This represents a significant advantage over typical digital architectures. With Mythic's integrated development environment, AI developers can quickly deploy even the most sophisticated deep neural networks, confident that they will perform effectively – from the data center to the edge device.

Pliops; Uri Beitler, CEO

Pliops Storage Processor highlights: Enables data centers to access data up to one hundred times faster with one-tenth of the computational load and power consumption • Provides better storage scalability, longer-lasting NVMe SSDs, and more efficient CPU utilization • Increases data stored on

SSDs by up to 6x through optimal space reduction and higher SSD utilization • Offloads the computational load required for cloud databases and software-defined storage • Increases throughput of cloud databases such as MySQL and Redis by up to ten times, while cutting the compute load by 80% and network traffic up to 99% • Reduces load latencies by three orders of magnitude and mixed read/write latencies by two orders of magnitude • Easy deployment as a low-profile PCIe card or cloud-based service.



SambaNova Systems; Rodrigo Liang, CEO

SambaNova DataScale lauds "world record-breaking performance metrics" at multi-rack scale when compared to the latest A100 GPUs in four key areas as follows:

1) Performance: World record DLRM inference 7x better throughput and latency than A100. World record BERT-Large training 1.4x faster than DGX A100 systems; 2) Accuracy: World record state of the art accuracy of 90.23% out-of-the-box for high-resolution computer vision compared to DGX A100 systems. World record state-of-the-art accuracy of 80.46% for DLRM recommendation engines compared to NVIDIA A100 GPUs; 3) Scale: World record BERT-Large training and state-of-the-art accuracy at multi-rack scale; 4) Ease of Use: From loading dock to data center, SambaNova DataScale quickly and easily integrates into any existing infrastructure running customer workloads in about 45

minutes. Download thousands of pre-trained Hugging Face Transformer models in seconds on SambaNova DataScale at state-of-the-art accuracy with no code changes required.

SiFive; Patrick Little, CEO

SiFive is providing an open-source alternative to Arm's CPU design business with core designs and custom silicon solutions for AI, high-performance computing and other growing markets based on the open and free RISC-V instruction set architecture. The San Mateo, Calif.-based startup has recently received takeover interest from multiple parties, including Intel, which has reportedly offered $2 billion to acquire the startup. Before the reported takeover interest, SiFive announced that Intel's new foundry business, Intel Foundry Services, will manufacture processors using SiFive's processor designs. Last August, the startup raised a $61M Series E funding round led by SK Hynix, with participation from several other investors, including Western Digital Capital, Qualcomm Ventures and Intel Capital. A month later, the company appointed former Qualcomm executive Patrick Little as its new CEO.

Tachyum; Radoslav Danilak, CEO

Tachyum Prodigy is lauded as "the world's first Universal Processor"-

> Tachyum's Prodigy processor can run HPC applications, convolutional AI, explainable AI, general AI, bio AI, and spiking neural networks, plus normal data center workloads, on a single homogeneous processor platform, using existing standard programming models. Without Prodigy, hyperscale data centers must use a combination of CPU, GPU, TPU hardware, for these different workloads, creating inefficiency, expense, and the complexity of separate supply and maintenance infrastructures. Using specific hardware dedicated to each type of workload (e.g. data center, AI, HPC), results in underutilization of hardware resources, and more challenging programming, support, and maintenance.

Xsight Labs; Guy Koren, CEO

X1 is the industry's first, low power 25.6Tbps (32 x 800G) data center switch with 100G SerDes and is designed from the ground up to address the bandwidth, power, form factor, and radix requirements for current and next generation cloud deployments and hyperscale networks.

X1 introduces a groundbreaking new architecture that achieves new levels of power and silicon efficiencies. It enables cloud service providers to deploy a 25.6Tbps (32 x 800G) in a 1 RU form factor.

X1's architecture incorporates a unique set of features, like application-optimized switching, X-PND™, and X-IQ™ enabling customers' switch deployments to achieve optimized latency and power efficiency.

Utilizing HPC-Scale Storage and AI for Business Intelligence

with sponsors Samsung, Datyra, NVIDIA, and WekaIO

Has your organization explored and/or deployed AI systems for business intelligence yet? (check one)

We have deployed AI for a variety of business applications	38%
We have deployed AI for a couple of business applications	15%
We are performing proof of concept evaluations on AI solutions, with the idea of deploying them in the near future	31%
We are talking to vendors about potential AI solutions	0%
We aren't actively exploring using AI in our organization	15%

What do you see as the greatest challenge for your organization to implement an AI solution? (check all that apply):

Understanding business value we can reasonably expect from AI	27%
Finding the right vendor and/or people to implement an AI solution	20%
Building the right training data set	40%
Affording the hardware required for a meaningful solution	13%
Achieving the right level hardware and software performance	40%
Other issues:	7%

G2M Research Multi-Vendor Webinar Series

Our June 15 webinar "Where Has NVMe-oF Progressed to in 2021" with sponsors KIOXIA (Matt Hallberg) and LightBits Labs (Josh Goldenhar):

NVMe-oF – the conventional way-

> Storage Initiator/Target Use Cases – Classical connection of storage uses (initiators) and storage devices (targets); Provides significantly better performance than SDSI-based protocols
>
> All-Flash Array (AFA) Back-End Us – NVMe-oF replaces SAS/SATA "tree" topologies behind network controllers; NVMe-oF provides significantly more flexibility

NVMe-oF has enabled "unconventional" use cases-

> Scale-Out Flash Storage (SOFS) - Connects servers and storage appliances with NVMe SSDs into a single namespace Provides DAS-like storage performance, but with the ability to manage storage globally Scale performance and capacity linearly
>
> Networked Storage Devices - NVMe-oF on 100GbE (10Gb/s) is faster than a PCIe 4.0 x4 connection (8Gb/s); NVMe virtual namespaces allows SSD partitioning; Potentially eliminates SSD "blast radius" issue

Hear what the experts from KIOXIA and Lightbits Labs had to say here and/or download a pdf of the slides. Our webinar schedule is below- Click on any of the topics to get more information about that specific webinar. Interested in Sponsoring a webinar? Contact **G2M** for a prospectus.

You can view all our webinars and access all the slide deck presentations.

July 13:	Computational Storage vs Virtualized Computation/Storage in the Datacenter: "And The Winner Is"?
Aug 17:	AI/ML Storage - Distributed vs Centralized Architectures
Sept 14:	Composable Infrastructure vs Hyper-Converged Infrastructure for Business Intelligence
Oct 12:	Cloud Service Providers: Is Public Cloud, Private Datacenter, or a Hybrid Model Right for You?
Nov 9:	The Radiometry Data Explosion: Can Storage Keep Pace?
Dec 14:	2021 Enterprise Storage Wrap-up Panel Discussion

Enterprise Storage Events

July 20-21	Gartner Tech Growth & Innovation, Virtual
July 19-23	GDC, Virtual
August 4-6	Storage Field Day, Virtual
August 15-17	Xchange 2021, San Antonio
August 16-19	Data Center World, Orlando
September 20-22	Bio-IT World, Boston
September 28-29	SDC21, Virtual
October 5-7	VMworld, Virtual
October 6-7	P99 Conf, Virtual