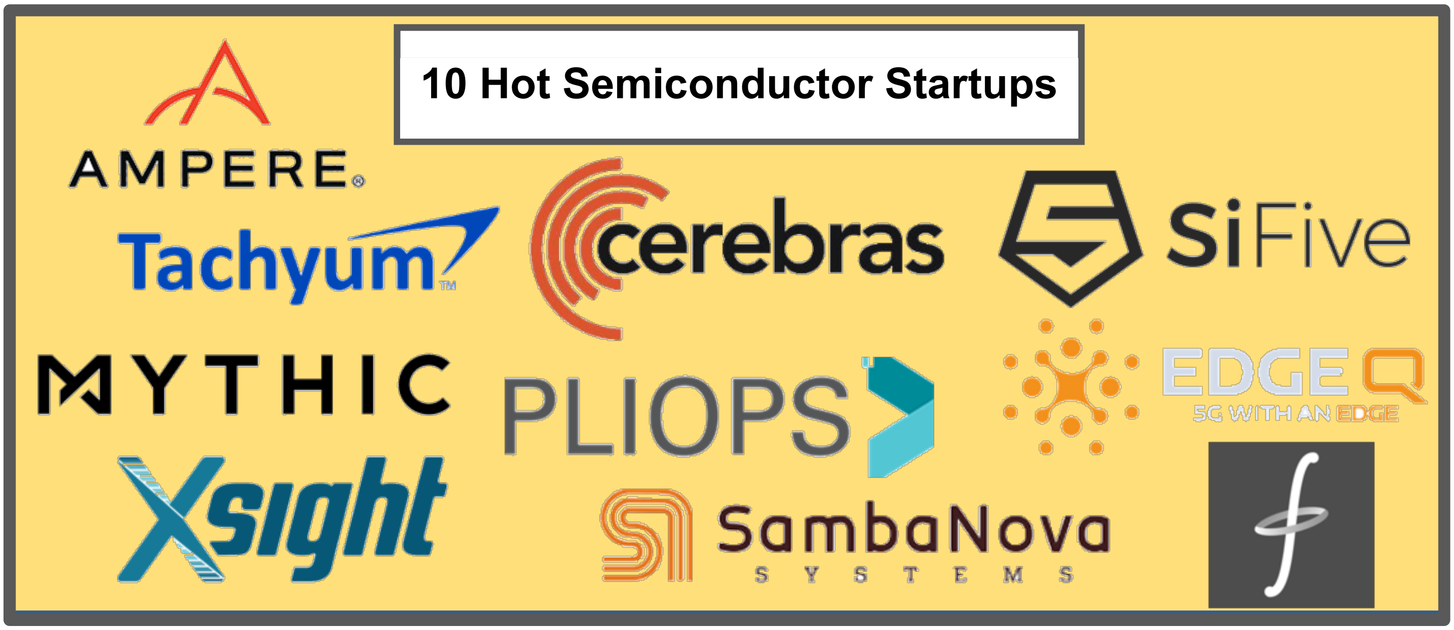


CRN picks for the “10 Hottest Semiconductor Startups of 2021 (So Far)” are Ampere Computing, Cerebras Systems, EdgeQ, Fungible, Mythic, Pliops, SambaNova Systems, SiFive, Tachyum, and XSights Labs. We take a look at each awardee below.

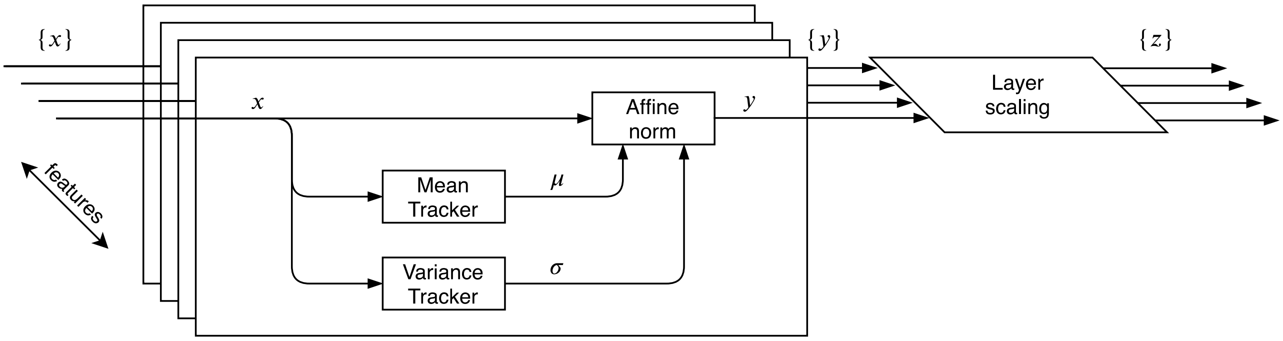
Ampere Computing; Renee James, CEO

> 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.

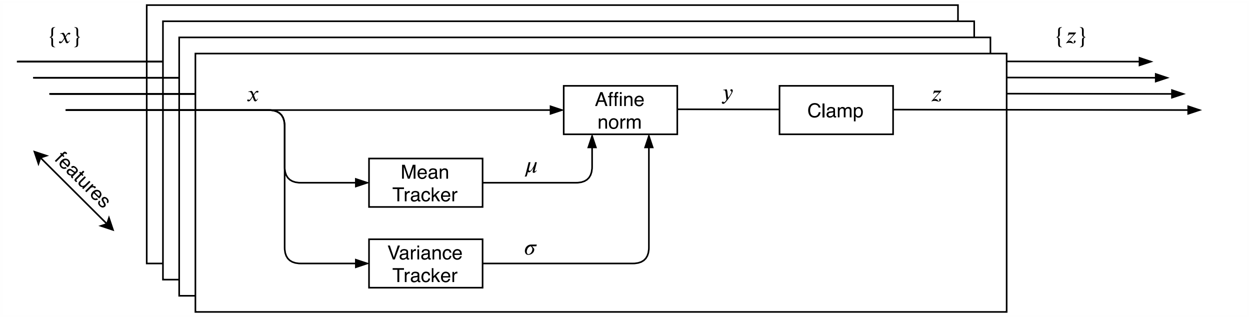
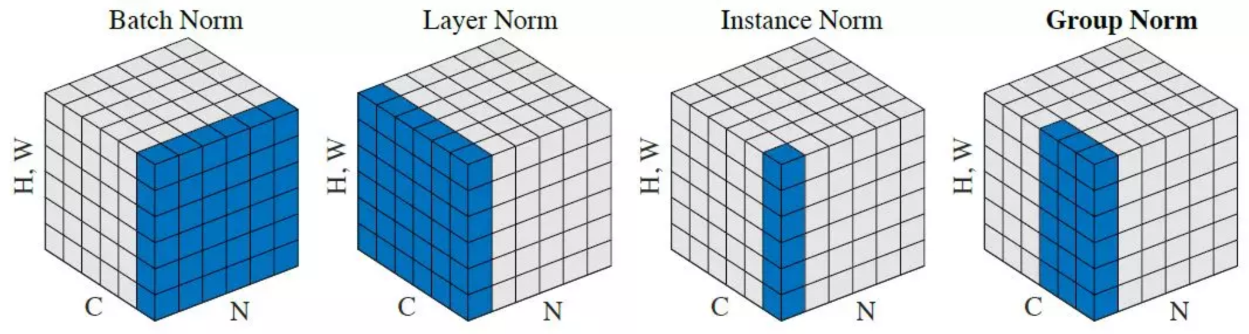


Figure 2: Online Normalization with activation clamping. Adaptive bias and gain excluded for simplicity.



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.

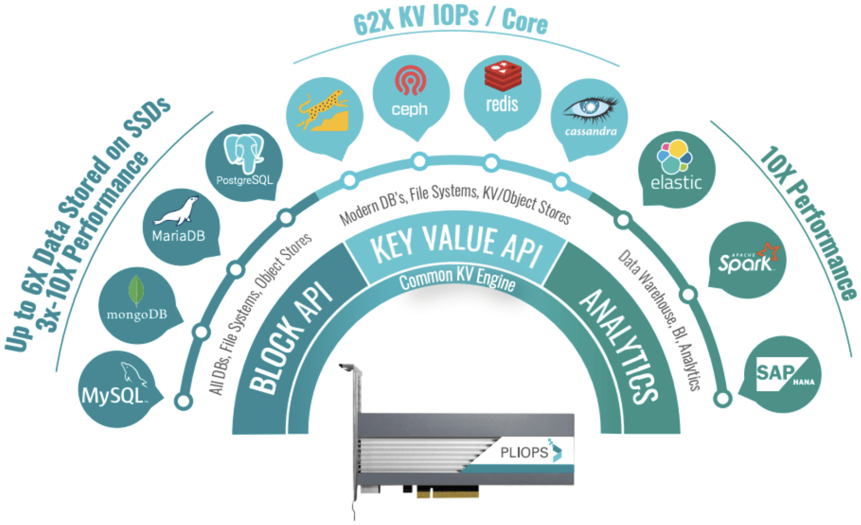
Mythic; Mike Henry, CEO

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.

Karen Heumann

G2M Communications