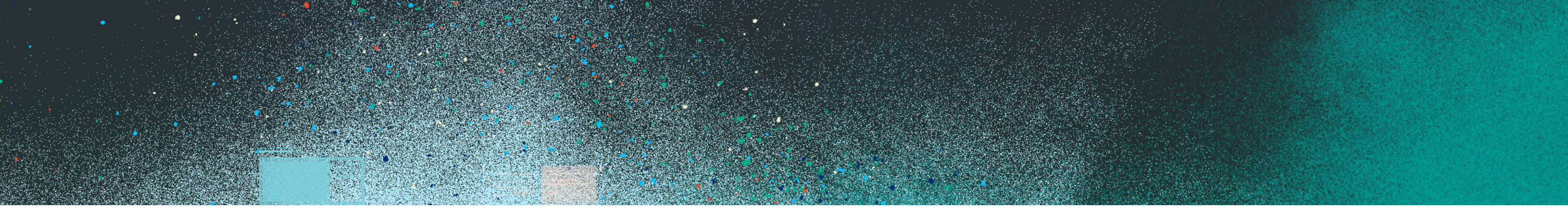
Maximizing GPU Utilization with HPE Alletra Storage MP X10000

AUTHOR

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

MARCH 2026



Executive Summary

The rapid evolution of generative artificial intelligence has transitioned from a focus on training models to wide-spread deployment at scale. Current industrial-scale AI infrastructure has rapidly adopted optimizations to enhance their effective GPU utilization. Now these techniques are becoming adopted within smaller providers and enterprise AI deployments.

As large language models (LLMs) grow in complexity and the demand for long-context windows increases, the underlying hardware architectures are facing unprecedented pressure. Traditional compute-centric models, where the Graphics Processing Unit (GPU) operate as isolated islands of memory and compute, show the inefficient nature of AI inferencing. At the center of this challenge is the Key-Value (KV) cache, a rapidly changing and massive data structure that stores the intermediate states of every token in a session. The management of this cache has become the primary determinant of "effective utilization"—the degree to which a GPU is occupied with the generation of new output tokens rather than the redundant processing of input prompts.

Signal65, working together with HPE and Kamiwaza tested HPE Alletra Storage MP X10000 as a target for maintaining a distributed KV-Cache. Testing was focused on assessing the impact of using high-speed storage for storing KV data, and then measuring the change in the output token generation rates and the time to first token.

Signal65 worked with HPE partner Kamiwaza to run real-world, agentic workloads to provide realistic modeling of the impact of using an X10000 system for KV-Cache offload. Our findings showed a massive improvement in the effective utilization rate of GPUs:

Reduction in Time to First Token (TTFT) **up to 21.5x** vs. not using any KV-Cache

Increase in output token generation rates by **up to 19.4x** vs. no KV-Cache

Benefits vs. memory-only offload show **TTFT reduction of 5.6x and token rate increase of 5.9x**

The results shown focus on relative improvements, since actual token generation rates and TTFT are heavily dependent upon specific prompts, the LLM model and the GPU type utilized. Moreover, our focus throughout this paper is on the relative improvements, which are broadly applicable to different LLM models and GPUs.

Effective GPU Utilization

In traditional data center monitoring, a GPU is often considered "utilized" if its compute engines are active. However, this metric is increasingly misleading in the context of LLM serving. True "effective utilization" refers specifically to the time the GPU spends generating value, producing output tokens rather than performing redundant re-computation of input tokens that have already been seen.

When GPU memory is exhausted, the inference engine must evict the KV-Cache of the least active or least recently used session. When that user returns for a subsequent conversation turn, the system must re-run the entire prefill phase, re-generating the entire KV vectors over again. This “re-computation tax” dramatically reduces effective token generation rates, and particularly the effective output token generation rates, robbing the GPU cluster of the compute cycles that could have been dedicated to output token generation. Testing has shown that reloading KV-Cache from high performance storage can deliver up to 21.5x faster TTFT compared to re-calculating, enabling the GPU to resume output generation almost instantly.

Inferencing Bottlenecks

To comprehend the necessity of storage offloading, it is essential to dissect the two stages of LLM inference: the prefill phase and decode phase. The prefill occurs when a model receives a prompt, it then processes all input tokens in parallel to generate the initial Key and Value vectors for every layer of the transformer architecture. This phase is computationally intensive and benefits from the massive parallel processing power of modern tensor cores.

In contrast, the decode phase is autoregressive and sequential. The model generates one token at a time, and for each new token, it utilizes all previous tokens by retrieving their stored vectors from the KV-Cache. This phase is memory-bandwidth bound. The GPU spends most of its time retrieving data from GPU memory (HBM) rather than performing arithmetic operations. As the context window expands from tens of thousands to millions of tokens, the KV-Cache grows, eventually consuming the remaining HBM and reducing the available context window and limiting additional requests.

The Impact of KV-Cache Scaling

The memory footprint of the KV-Cache is a function of the model's architectural hyper-parameters, the batch size, and the total sequence length. The following table provides a comparison of memory requirements for prominent model architectures, illustrating the scale of the challenge.

Model Architecture	Layers (L)	Heads (H)	Head Dimension (D)	Size per Token (FP16)	Cache Size @ 120K Tokens
Llama-2-7B	32	32	128	~0.50 MB	~62 GB
Llama-2-70B	80	64	128	~2.50 MB	~312 GB
Llama-3.1-70B (with GQA)	80	64 (8 KV)	128	~0.31 MB	~39 GB
Mixtral 8x7B	32	32	128	~1.00 MB	~120 GB

The analytical formula used to derive these values is as follows:

Total Memory = Layers * Heads * Dimensions * Token Size

In an environment where an NVIDIA H100 GPU provides only 80 GB of capacity, it is mathematically impossible to sustain long-context multi-turn conversations for even a small number of concurrent users without frequent cache evictions.

The Impact of KV-Cache Scaling

To maximize effective utilization, the industry is moving toward disaggregated serving architectures. In these environments, the prefill and decode stages are separated into different hardware pools.

- **Prefill Nodes:** Using compute optimized GPUs to ingest prompts and build an initial KV-Cache
- **Decode Nodes:** Optimized for high memory bandwidth and latency-sensitive token generation

With disaggregated inferencing, the ability to rapidly access a unified KV-Cache from multiple inferencing nodes is even more critical. Moreover, HPE Alletra Storage MP X10000 with RDMA enhanced transfers, helps enable cost-effective AI inferencing at scale when using traditional inferencing, and provides significant advantages when new deploying disaggregated inferencing.

HPE Alletra Storage MP X10000

HPE Alletra Storage MP X10000 is a software-defined, scale-out object storage system designed for enterprise AI workloads, including fine-tuning and as a high-performance memory tier for inferencing. The X10000 utilizes a disaggregated architecture which allows each storage controller to access all NVMe SSDs in the cluster simultaneously, maximizing performance while limiting the impact of "noisy neighbor" issues associated with some system designs.

Inferencing Bottlenecks

The X10000 is designed to scale linearly in both performance and capacity, which is essential for the unpredictable demands of AI workloads.

Component	Specification
Node Architecture	2U chassis building blocks with independent controller and storage nodes
Max Capacity	Scalable from 15 TB to over 6 PB of physical capacity per system, not including data reduction
Connectivity	Support for 100 Gbps Ethernet with RDMA capabilities
Read Throughput	Up to 20 GB/s per node for large RDMA-enabled reads
Storage Media	NVMe TLC and QLC SSD options for tiered performance/cost

The disaggregated architecture ensures that as an organization adds more controller nodes to handle increasing request concurrency, the aggregate bandwidth to the KV-Cache grows linearly. This provides a future-proof foundation for scaling inference clusters from hundreds to thousands of GPUs.

Object Storage for AI Workloads

AI workloads are not monolithic, with distinct categories each with different requirements-with training and fine-tuning quite different from Inferencing. Fine-tuning can be characterized by having periods of high read, followed by sustained writes while creating checkpoints. Inferencing also has several distinct components, including supporting RAG workloads, and KV-Cache.

While block or file storage may be utilized for some of these workloads, object storage provides several distinct operational advantages for several components, including the KV-Cache lifecycle.

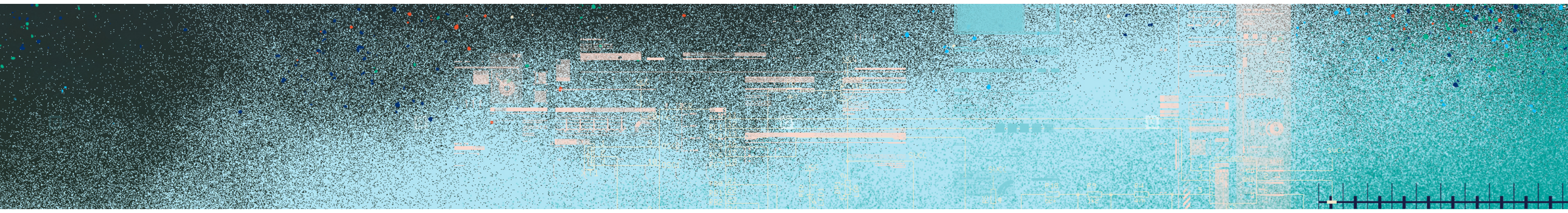
1. **Massive Throughput at Scale:** Object protocols are designed for high concurrency with access to rich metadata to help automated life-cycle policies. When optimized with RDMA, it provides an efficient path for moving large blocks of KV-Cache data between storage and GPU memory.

2. **Autonomous Lifecycle Management**: KV-Cache data is typically transient. A conversation may only remain "warm" for hours or days. Object storage allows administrators to set bucket-level lifecycle policies, such as a Time to Live (TTL) of 7 days. The X10000 automatically manages the expiration and deletion of these objects, removing the need for external data tiering or cleanup tools.

3. **Global Namespace for Cluster-Wide Reuse:** Object storage provides shared and consistent access for hundreds or even thousands of inference nodes. If a user's session is moved from one node to another, the new node can instantly retrieve the existing KV-Cache from the unified repository, enabling inference session mobility between disparate GPUs.

Enhanced RAG Pipelines

For Retrieval-Augmented Generation (RAG) workloads, data Intelligence nodes for the X10000 can enhance raw data with vector embeddings. The HPE Alletra Storage MP X10000 may include optional, integrated **data intelligence node** powered by NVIDIA L40S GPU technology. This node allows the storage system to analyze objects as they are ingested, extracting metadata and generating vector embeddings in real-time. As the storage system receives new documents, it can automatically generate the embeddings required for semantic search and store them alongside the raw data. This creates an "AI Data Platform" that leverages the relevant metadata along with the raw data for active inferencing with RAG.



Signal65 and Kamiwaza Analysis

HPE asked their partner, Kamiwaza to perform real-world testing of the X10000's impact on KV-Cache while running agentic workloads. Kamiwaza utilized their KAMI (Kamiwaza Agentic Merit Index) testing framework to generate unique, agentic inferencing workloads against an HPE inferencing server with 8 x NVIDIA H200 GPUs attached to an X10000 object storage system. All testing was performed in HPE European Solutions & Performance Center with test scenarios designed by Kamiwaza working with Signal65 to review and validate the configurations and outcomes. The results clearly show significant improvement in output token generation rates by offloading the KV-Cache to the RDMA enhanced X10000 system.

The benefits of using GPU Direct Object Storage for the HPE Alletra MP X10000 are shown in the graph in figure 1 below:

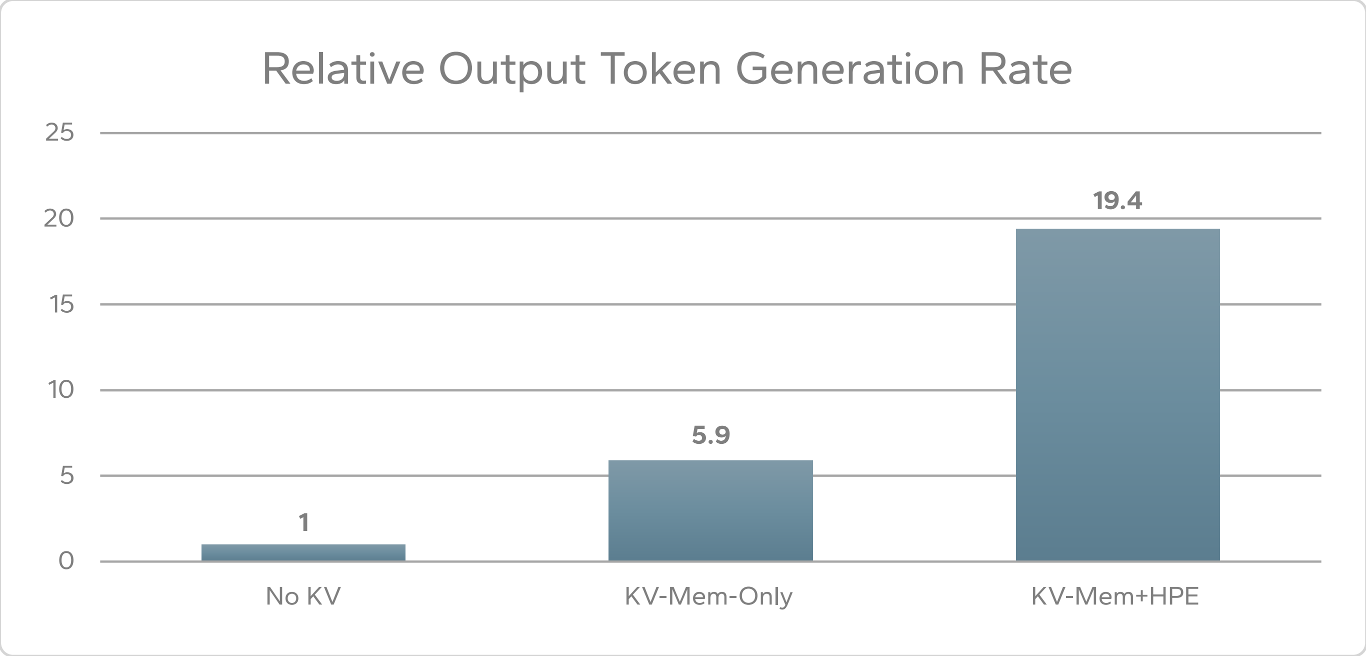


Figure 1: *Relative Token Generation Rate Improvment*

Signal65 Comments *– As inferencing workloads become increasingly the focus of LLM deployments, the use of KV-Cache has a significant impact on the productivity and cost effectiveness of inferencing. The addition of HPE Alletra Storage MP X10000 as a secondary KV-Cache produced over a 5x increase in productivity, whether measured by TTFT or output token generation rates. Clearly, companies using a scalable, secondary KV-Cache like the X10000 will have a significant productivity and financial advantage compared to those who do not.*

Testing Environment

As discussed, real-world inferencing workloads were created that simulated up to 96 simultaneous users generating agentic requests. The KAMI test framework was utilized, with measurements including TTFT, and output token generation rates. Additional details are provided in the Appendix.

Reduction in TTFT

The use of the Alletra X10000 as a secondary KV-Cache showed a reduction of 5.6X in TTFT, compared to a system using memory only KV-Cache offload. For systems without any KV-Cache offload, the benefits of adding memory plus the X10000 showed up to a 21.5X improvement in TTFT. These measurements were taken on the third or later conversation turn during multi-turn conversation inferencing sessions.

These results were obtained using a setup that included system memory along with the X10000 accessed via GPU direct object via RDAM, acting as a secondary KV-Cache. See the Appendix for further details.

Output Token Generation Rates

As outlined above, one of the best measurements of GPU efficiency is evaluating the output token generation rate. During testing by Kamiwaza using the agentic workload generation, we found that the output token throughput increased by 5.9x when compared to using memory only cache offload, and by up to 19.4x compared to a system with no KV-Cache.

Empirical Proof Points from Testing

The real-world testing outlined above validates these theoretical models. In high-concurrency scenarios where multi-turn conversations are the primary workload, the following results were observed:

- **Throughput Gains:** Systems utilizing the X10000 for KV-Cache offloading achieved up to 19.4x higher token output throughput compared to GPU memory-only configurations.
- **Latency Mitigation:** Under heavy load, the tail latency for TTFT remained stable with offloading, whereas the baseline system saw TTFT spikes as re-computation became rampant.

Benefits of GDS: Measuring Performance of RDMA vs. TCP

Internal testing at HPE and during our hands-on evaluation have demonstrated that the RDMA-enabled data path on the X10000 provides a transformative performance boost over traditional S3. The ability to move data directly between storage and a GPU over RDMA is known as GPU Direct Storage (GDS). According to HPE published reports: the X10000 demonstrated up to an 80% performance improvement using RDMA for S3-compatible storage compared to S3 over TCP.

Metric	Traditional S3 over HTTP	RDMA for S3 Storage	Improvement
Throughput	~11 GB/s	~20 GB/s+	1.8x - 2x
Latency	Baseline	80% Lower	5.1x reduction
Latency Variability (P99)	High Jitter	80% Reduction	Improved consistency
Host CPU Utilization	High	Low	Significant Reduction

By reducing the CPU overhead of data movement, GDS ensures that the host's computational resources are available for higher-value tasks, such as request orchestration, result filtering, or other AI related tasks.

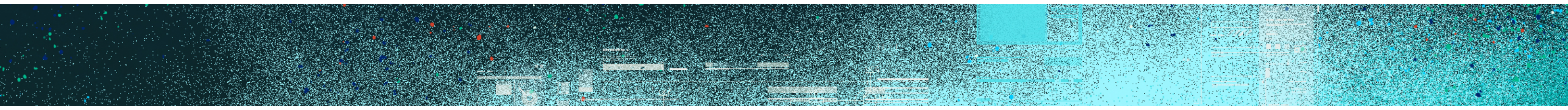
Modeling KV-Cache Usage

To further validate the impact of KV-Cache offloading on GPU effective utilization, Signal65 constructed a queuing theory model. A standard inferencing environment may be modeled as an M/G/1 queue (Poisson arrivals, general service times). However, this only works if there is sufficient retention of the KV-Cache, in short term (memory) plus the long term (storage) tier.

Memory-Constrained Queue

In an LLM inferencing system, a request is only "serviced" if there is both compute capacity to generate the tokens and memory capacity to hold the KV-Cache. If new requests are high (i.e. the arrival rate is high), the aggregate memory demand of all active requests will exceed the physical capacity of our KV-Cache retention. At this point, the system becomes unstable. The queue length grows unbounded, and the effective processing rate goes to zero because the system spends all its time evicting and recomputing caches.

By offloading the KV-Cache to the X10000, the "memory limit" is effectively moved from the limit of the GPU and system RAM to the petabyte scale of the X10000. This ensures the model remains compute-bound, where stability is determined by the tokens it can generate per second, rather than the size of the GPU or system RAM.



HPE and NVIDIA Technology Alignment

The implementation of GPU direct storage via RDMA on the client side is facilitated by the NVIDIA BlueField-3 Data Processing Unit (DPU). The BlueField-3 is an infrastructure-on-a-chip that offloads networking, storage, and security tasks from the host server.

In an LLM inference system, NVIDIA BlueField-3 DPUs serve as dedicated orchestrator for the storage data path. Featuring 16 Arm cores and up to 400 Gbps of NDR InfiniBand or Ethernet connectivity, the DPU manages the RDMA connections to the X10000 and executes the cuObject server-side callbacks in hardware.

The BlueField-3 provides several benefits, including the ability to provide **Zero-Trust Security** without a performance overhead. The DPU can perform line-rate encryption and decryption of data as it moves between the storage array and the GPU memory, ensuring that sensitive conversational data is never exposed in the clear on the network fabric. Furthermore, the DPU can emulate block or file storage through technologies like BlueField SNAP, allowing legacy inference frameworks to benefit from the X10000's object performance without modifying their internal I/O logic.

Direct Data Movement: NVIDIA cuObject and RDMA

The primary hurdle for using object storage in the inference data path has historically been the latency and CPU overhead of the traditional HTTP/TCP networking stack. Standard S3 transfers require the host CPU to handle packet processing, multiple memory copies between the kernel and user space, along with the typical "jitter" of the TCP/IP protocol stack.

GPU Direct Storage for Objects API

To circumvent these limitations, NVIDIA developed the cuObject library, a high-performance suite of libraries that enables direct data transfers between GPU memory and S3-compatible storage using Remote Direct Memory Access (RDMA). The cuObject architecture introduces a clean separation between the control plane and the data plane:

- **Control Plane:** The application uses the standard S3 API (GET/PUT) to signal a transfer. The request is modified with custom metadata tags, which contain the necessary RDMA memory descriptors.
- **Data Plane:** Once the transfer is authorized, the X10000 storage node performs an RDMA_WRITE (for GET) or RDMA_READ (for PUT) directly to or from the GPU's memory. This bypasses the host CPU, the system RAM, and the operating system kernel entirely.

Conclusions: Achieving Effective AI Efficiency

As generative AI continues to mature, companies will judge the success of their AI projects based upon the cost effectiveness of their results. Increasingly, the metric for success will shift from model choice to the efficiency of the inference service. Maximizing effective GPU utilization is not merely a technical challenge; it is a business imperative. Every second a GPU spends re-processing data is a time spent not generating value for the user or revenue for the provider.

TCO Advantages

The economic argument for HPE Alletra Storage MP X10000 is rooted in the high cost of GPU infrastructure. An enterprise can choose to scale their inference capacity in two ways:

1. **Add More GPUs:** This is a linear increase in cost, power consumption, and data center footprint. It "solves" the memory problem by brute force but leaves the GPUs underutilized during the decode phase that produces the output results.

2. **Offload State to the X10000:** By investing in high-performance shared storage, organizations can increase the number of concurrent sessions per GPU. A single X10000 cluster can support the KV-Cache needs of hundreds of GPUs, providing a much more favorable cost-per-token than adding more high-end compute nodes.

HPE Alletra Storage MP X10000, integrated with NVIDIA's cuObject and BlueField-3 technologies, represents one of the first solutions for the KV-Cache memory crisis. By providing petabyte-scale, RDMA-enabled object repository, it allows the GPU to focus on its primary strength: high-speed token generation. Through the decoupling of compute and memory, the X10000 system's automation of data lifecycles via bucket policies enables inference clusters to scale beyond GPU and system memory limits, thereby enhancing the ability to deliver cost-effective AI.

The test results are clear: the most efficient path to scaling AI is not through more computation, but through smarter retention of pre-computed data. By embracing a storage strategy for KV-Cache management, enterprises can break through performance inhibitors and unlock the full potential of their GPU investments.[1]

[1] HPE Alletra Storage MP X10000 Page: https://www.hpe.com/us/en/alletra-storage-mp-x10000.html

Appendix

Agentic Inferencing Stack

- KAMI – The Kamiwaza Agentic Merit Index (aka KAMI) Agentic workload
- vLLM – Inferencing server
- LMCache - KV-Cache software

Kamiwaza Agentic AI Workload

https://signal65.com/research/ai/benchmarking-leadership-in-open-and-proprietary-models/

Inferencing System

HPE ProLiant Compute DL380a Gen12 AI server

- 2 x CPUs
- 1.5 TB System RAM (Used as the primary KV-Cache offload)
- 8 x H200 GPUs (w/ 140 GB HBM)
- 2 x 100 Gb ConnectX-7 RDMA enhanced DPU

Storage Target (Secondary KV-Cache Offload)

HPE Alletra Storage MP X10000

- 4 x X10250 Controller Nodes
- 2x Aruba 8325 backend switches
- 24x 3,84TB NVMe
- 8x 100GbE Frontend

Important Information About this Report

CONTRIBUTORS

Russ Fellows
VP, Labs | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author’s name, author’s title, and “Signal65.” Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION

Signal65 I signal65.com