softserve

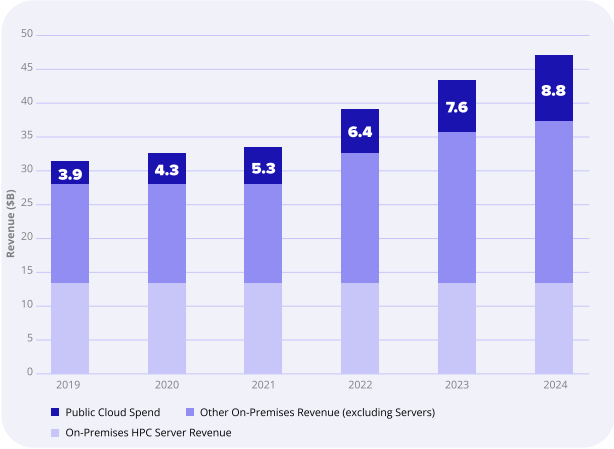
EFFICIENT HIGH-PERFORMANCE COMPUTING MATTERS MORE THAN EVER

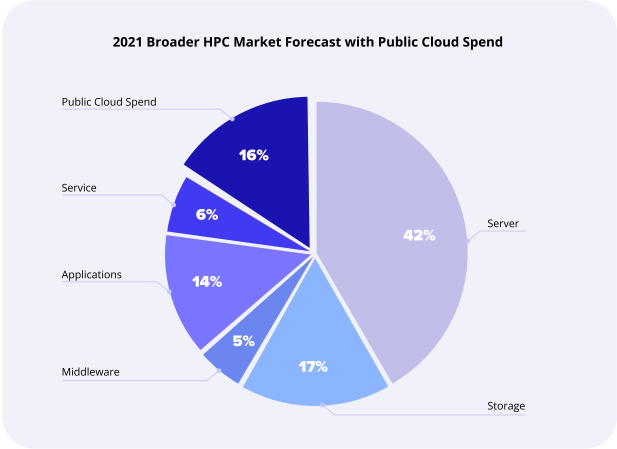
Important data-based fields such as climate modeling, genomic sequencing, AI training, and financial forecasting all share one critical dependency, **computational power**. But as the world's appetite for data grows exponentially, the question is no longer **how fast** we can compute, but **how efficiently**.

According to market forecasts, the global high-performance computing (HPC) industry will continue to expand sharply through 2030, fueled by demand from sectors including government and defense, BFSI, energy, and healthcare.

A COMPLETE HPC MARKET PICTURE

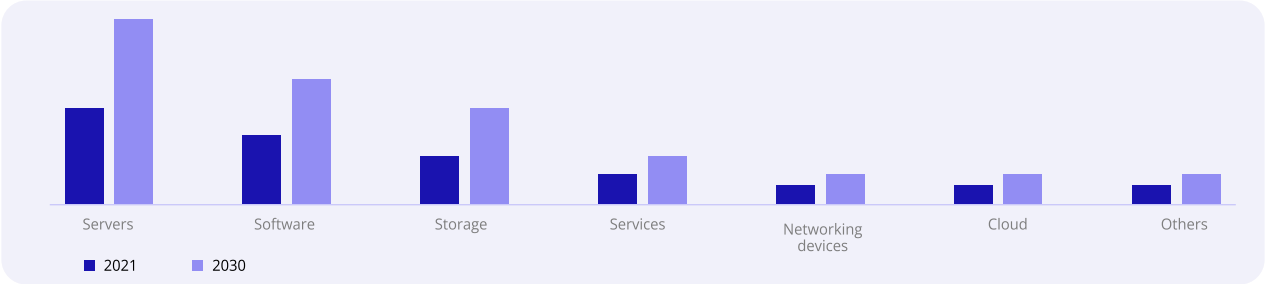
INCORPORATING THE CLOUD TO THE BROADER MARKET FORECAST



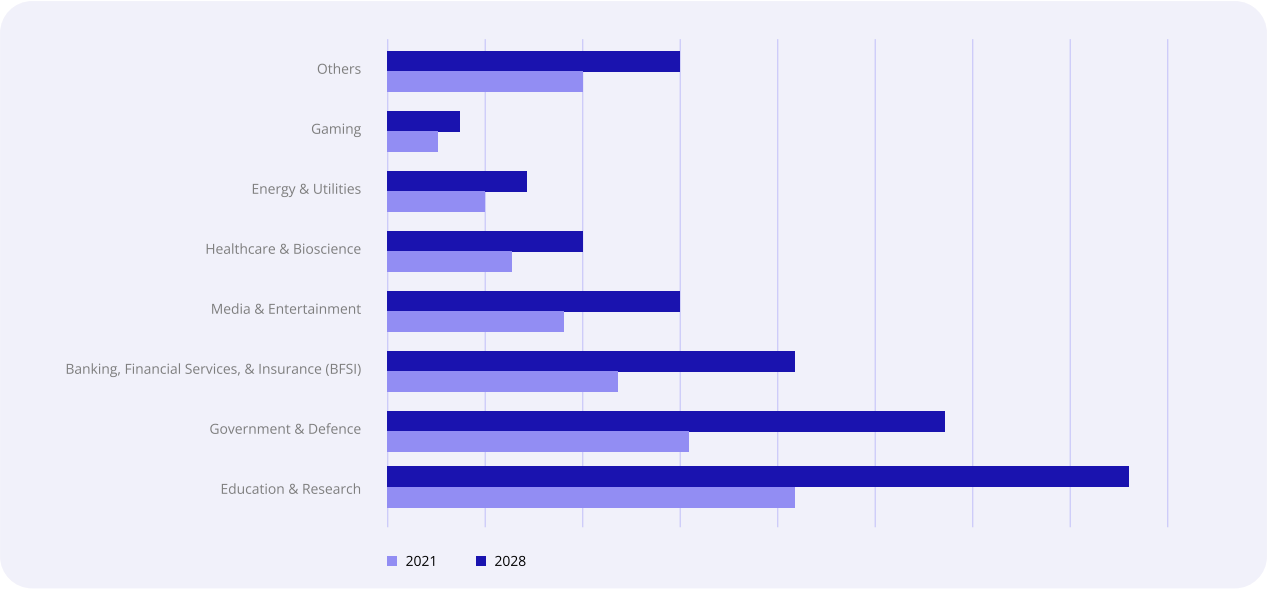


HIGH-PERFORMANCE COMPUTING MARKET

BY COMPONENT, IN VALUE (2021 & 2030)



Source: www.psmarketresearch.com



Global high-performance computing (HPC) market, by application areas, 2021-2028 (USD million)

However, more compute power doesn't always translate to better performance. Businesses often respond to performance bottlenecks by scaling infrastructure (adding GPUs, nodes, and cloud resources) without addressing inefficiencies in data movement, communication, and algorithm design.

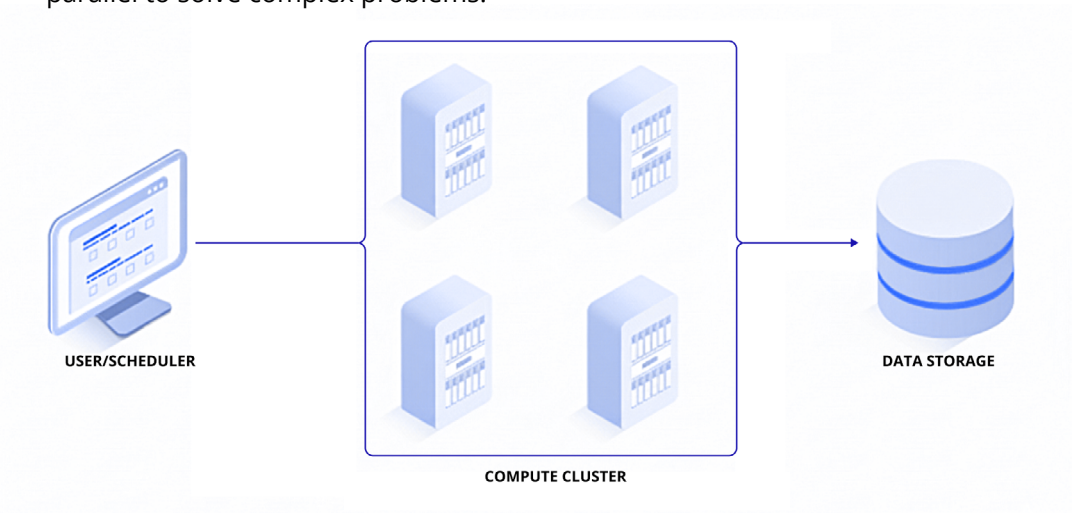
SoftServe R&D has seen this challenge firsthand. When one client's HPC-based physical simulation algorithm hit a scaling wall (and skyrocketing cloud bills), our engineers redefined performance by **rethinking how data moves** rather than adding more hardware. The results? Twice the performance at half the cost.

This is the story of **HPC optimization done right**, and why efficiency, not brute force, is shaping the future of AI, cloud, and data innovation.

SEE MORE

Power performance with HPC

HPC isn't a single supercomputer. It's a **system of distributed computing resources** working in parallel to solve complex problems.

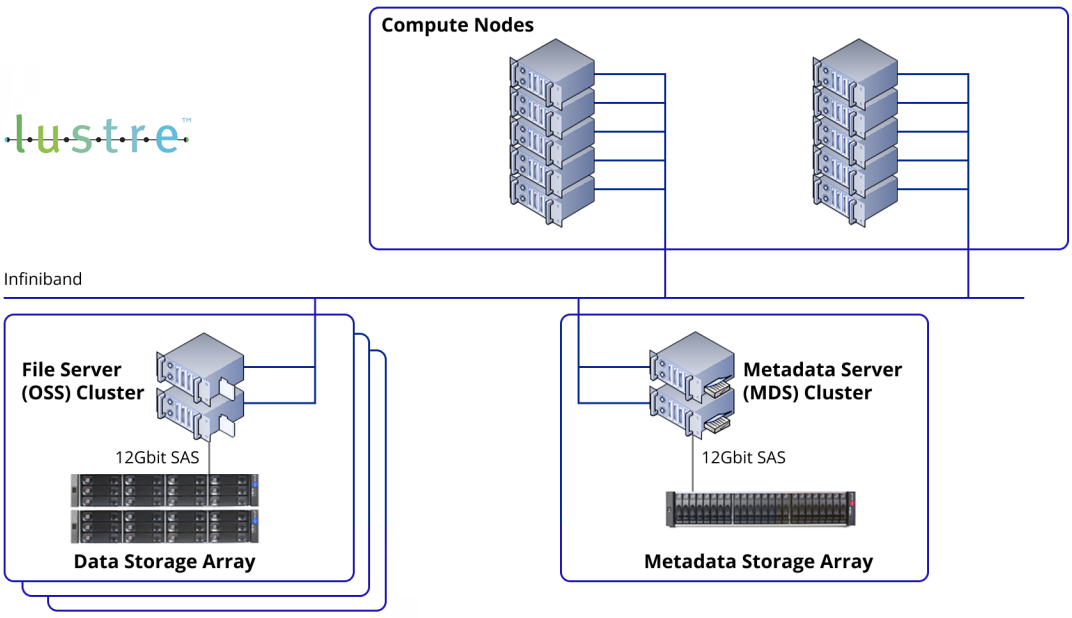


A typical HPC environment includes:

- **Scalable computing clusters** of multi-core CPUs and GPUs (e.g., Intel Xeon, NVIDIA H100/H200, B200, AMD MI250X/MI35X)
- **High-speed interconnects** such as InfiniBand or RDMA over Ethernet
- **Parallel file systems** like Lustre or IBM Spectrum Scale for massive data throughput

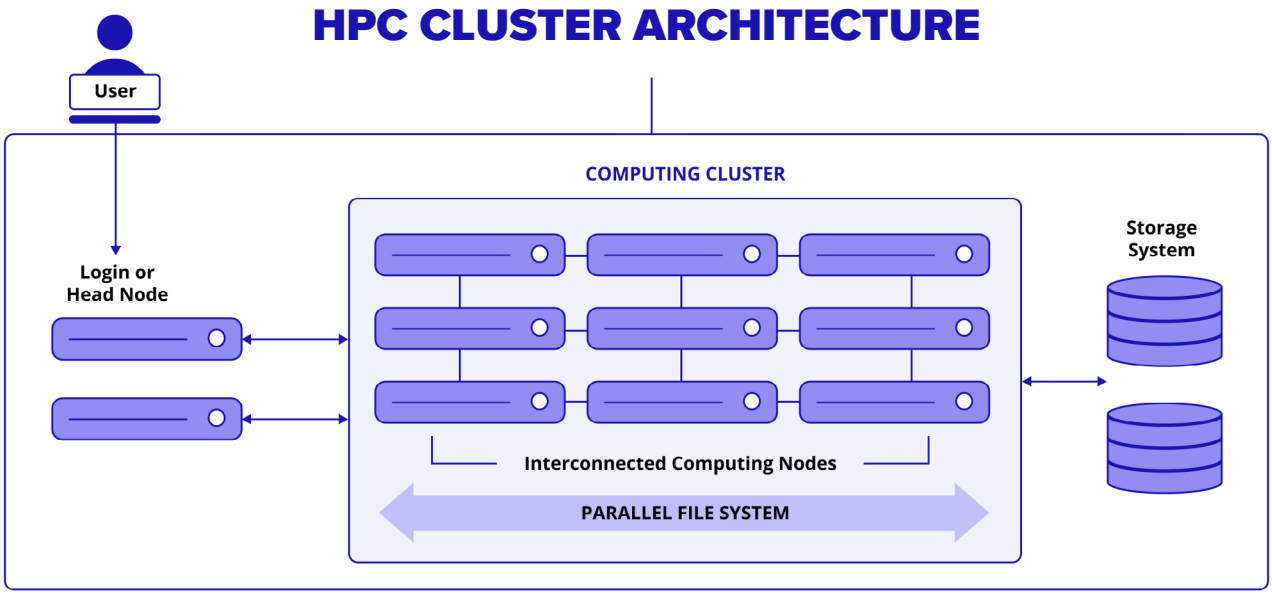
LUSTRE DISTRIBUTED FILE SERVING

FOR HPC ENVIRONMENTS

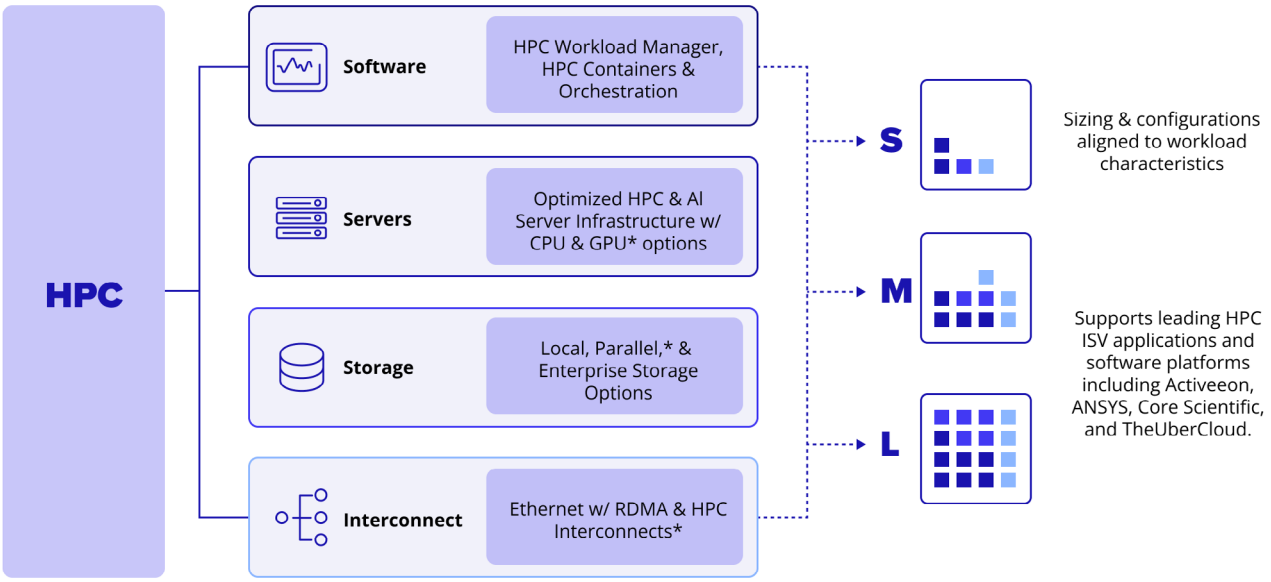


InfiniBand topology HPC cluster — an InfiniBand switch is integrated into each of the classes

- **Sophisticated schedulers** that manage workloads across nodes



Together, these components allow organizations to simulate climate systems, model financial risks, and run AI training at scales traditional computing simply can't match.



But scalability comes with complexity. From GPU memory access to inter-node communication, each layer creates the potential for performance bottlenecks. And as more organizations migrate HPC workloads to the cloud, inefficiencies result in wasted compute time, energy, and cost.

Move beyond scaling

In one R&D case, a client's researchers had built a fully parallel algorithm for simulating physical systems. By the book, it should have scaled perfectly across GPUs — **Amdahl's and Gustafson's laws** both suggested near-linear performance.

Instead, as GPU counts rose, performance flatlined. Costs didn't. A single cloud invoice totaled nearly **$100,000** for one workload iteration.

The issue wasn't the math. It was the movement of data. The algorithm performed unnecessary CPU–GPU memory transfers and inefficient message passing between nodes.

This is a common HPC pitfall. Teams chase **effectiveness** (more compute power) when they should target **efficiency** (better utilization of existing resources).

Rethink efficiency through optimization

Our team took a systematic, top-down approach:

1 Profile before you optimize.

Performance begins with understanding. Using tools like **NVIDIA Nsight Systems and Nsight Compute**, we profiled the entire stack — CPU, GPU, memory, and MPI communication — across multiple input datasets and configurations.

We froze the software environment to ensure reproducibility and began investigating the true culprits behind latency.

2 Streamline communication.

Most inefficiencies stemmed from how data moved between nodes and GPUs.

- **GPU Direct (RDMA):** By bypassing CPU memory and transferring data directly between GPU and NIC, we cut communication latency and improved bandwidth dramatically.
- **NVSHMEM (OpenSHMEM for NVIDIA GPUs):** Integrating communication directly into CUDA kernels simplified the codebase and eliminated redundant synchronization points.
- **Compression:** For data-heavy workloads, introducing selective **lossless** or **lossy** compression (e.g., reducing precision from FP64 to FP32) reduced communication volume without sacrificing critical accuracy.

These adjustments delivered both cleaner architecture and tangible performance gains.

3 Master the memory hierarchy.

As the saying goes, "HPC turns compute problems into memory problems."

Optimizing memory access patterns yielded some of the biggest improvements.

- **Data Layout Optimization:** Transforming from an **Array of Structures (AoS)** to a **Structure of Arrays (SoA)** layout improved cache utilization and memory coalescing.
- **Asynchronous Operations:** Leveraging async read/write between host and device overlapped computation with data transfer, boosting GPU utilization by up to 20%.
- **Smart Buffering:** Implementing double-buffer techniques allowed continuous compute–transfer cycles, minimizing idle GPU time.

4 Fine-tune GPU utilization.

To fully exploit the hardware, we optimized kernel configuration and register use:

- Adjusted grid/block dimensions for balanced workload distribution.
- Controlled register pressure by limiting loop unrolling and caching intermediate data efficiently.
- Maximized shared memory usage across streaming multiprocessors (SMs) to reach ~75% active utilization.

Measurable impact: Efficiency in action

The optimization process led to measurable, business-level outcomes:

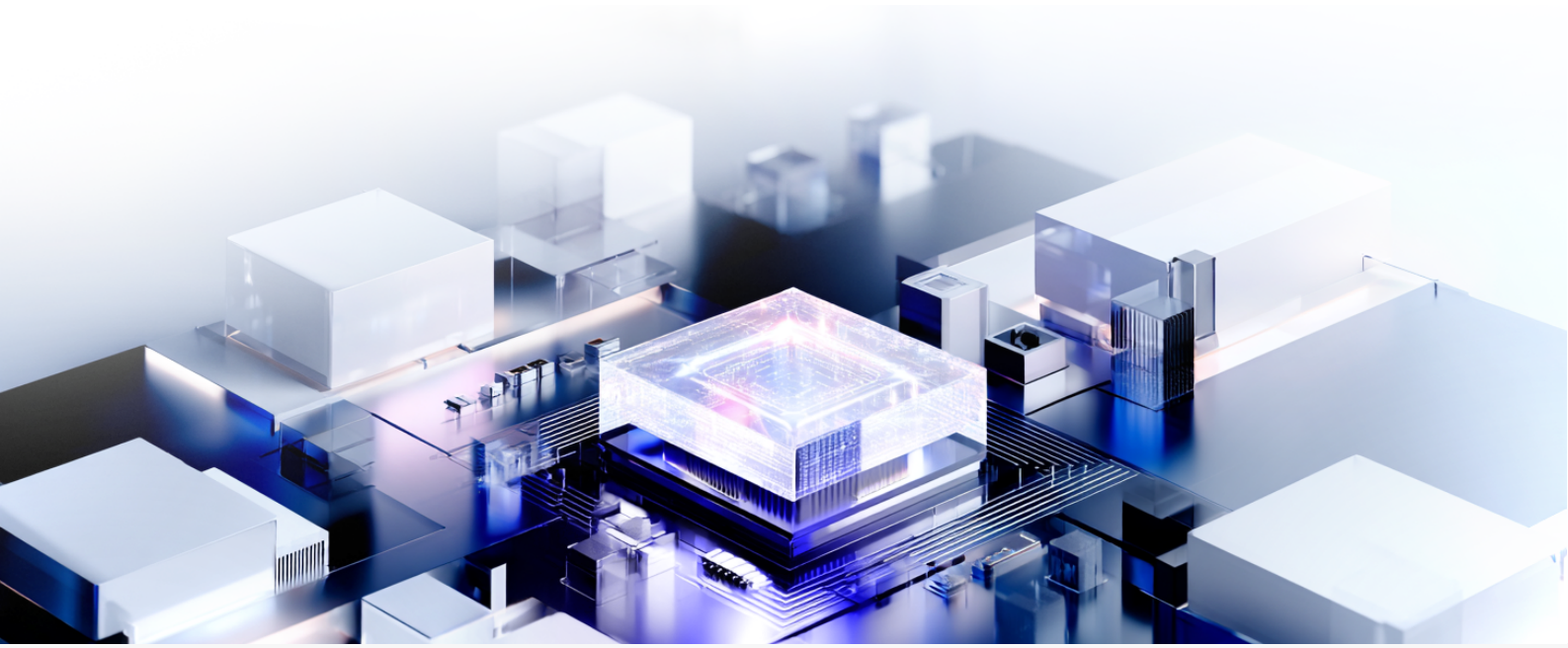
Metric	Before Optimization	After Optimization
Execution Time	Baseline	**>2× faster**
Scaling Efficiency	Plateau after ~8 GPUs	**Linear up to 20+ GPUs**
Memory Footprint	100% baseline	**~50% reduction**
Cloud Cost per Job	~$100K	**~$50K (estimated)**

Beyond the numbers, the project demonstrated a broader truth:
Sustainable performance isn't about adding resources; it's about smarter design.

Challenges

Modern GPUs prioritize AI-focused math formats such as FP8, BF16, and TF32, while traditional IEEE FP32 and FP64 show slower improvement. For HPC workloads, this creates engineering challenges rather than hard limits. Accuracy must be preserved while extracting performance gains. When dense phases occur, refactor MMA paths. In other cases, optimize streaming-core code. AI techniques such as auto-tuning, learned surrogates, and ML-guided preconditioners can assist in designing faster and more accurate solutions.

SoftServe R&D is actively pursuing this applied research to deliver **SoL HPC applications** for our customers — measured by real **time-to-accuracy** gains on today's GPUs.



Efficiency is the new performance

As **Moore's Law slows**, we can no longer count on faster processors to deliver linear gains. Instead, innovation must come from synergy — between **researchers**, **software engineers**, and **hardware designers** — working as one ecosystem.

Energy efficiency and sustainability will increasingly define competitive advantage, especially in compute-heavy fields like **AI and digital twin simulations**.

The future belongs to those who can compute smarter, not just faster.

High-Performance Computing is the foundation of modern digital transformation. From enabling real-time AI to powering sustainable energy modeling, **efficiency is the competitive edge** that defines success.

SoftServe's R&D team continues to push the boundaries of HPC optimization, combining domain expertise, algorithmic innovation, and software craftsmanship. Because in today's world, every teraflop counts, but only the efficient ones truly transform.

Whether you're training large AI models, simulating physical systems, or modernizing cloud-native infrastructure, **optimization is where innovation starts**.

Contact SoftServe to harness HPC's full potential and bridge AI, data, and cloud strategies with cost-efficient, scalable architectures.

CONTACT US

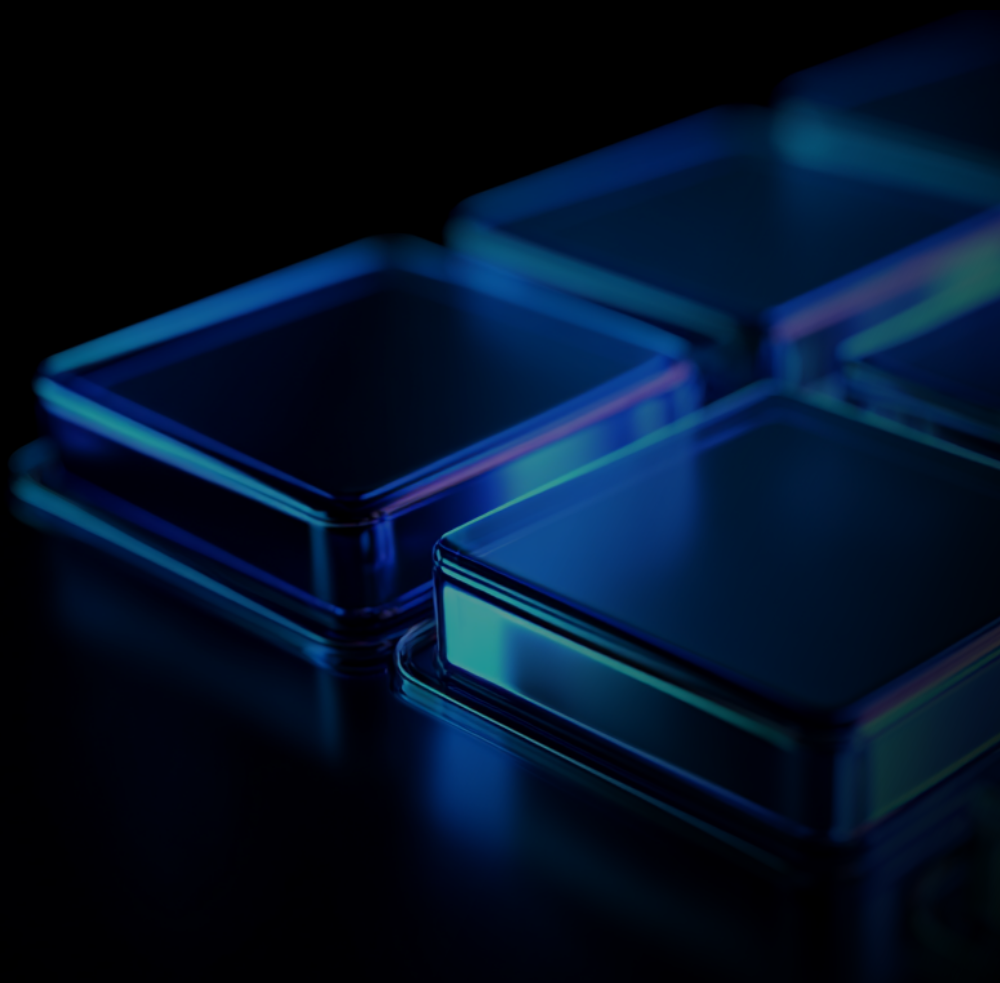
About Us

SoftServe is a premier IT consulting and digital services provider.

We expand the horizon of new technologies to solve today's complex business challenges and achieve meaningful outcomes for our clients. Our boundless curiosity drives us to explore and reimagine the art of the possible. Clients confidently rely on SoftServe to architect and execute mature and innovative capabilities, such as digital engineering, data and analytics, cloud, and AI/ML.

Our global reputation is gained from more than 30 years of experience delivering superior digital solutions at exceptional speed by top-tier engineering talent to enterprise industries, including high tech, financial services, healthcare, life sciences, retail, energy, and manufacturing.

Visit our website, blog, LinkedIn, Facebook, and X (Twitter) pages for more information.



AUSTIN HQ

201 W. 5th Street, Suite 1550
Austin, TX 78701
+1 866 687 3588 (USA)
Toll Free: +1 866 687 3588

LONDON

30 Cannon Street
London EC4 6XH
United Kingdom
+44 203 807 01 41

info@softserveinc.com
www.softserveinc.com

softserve