

STRATEGIC PROFILE & CAPABILITY OVERVIEW

The software stack that makes *AI infrastructure economics* actually work.

Patent-pending middleware. Independently validated performance. NeoCloud-native architecture — engineered to unlock superior economics from the GPUs you already own.



PATENT-PENDING IP

THIRD-PARTY VALIDATED

NEOCLOUD-NATIVE

SOVEREIGN-READY

+33.4 %AVG THROUGHPUT GAIN
AT SATURATION**+91.4 %**AVG TTFT IMPROVEMENT
AT HIGH CONCURRENCY**Up to 50%**COST REDUCTION AI
INFERENCE WORKLOADS**25^{MW}**NEOCLOUD INITIATIVE
NAVI MUMBAI

EXECUTIVE SUMMARY

Cogniware AI is an AI infrastructure technology company headquartered in the Washington, DC area, developing patent-pending middleware that optimizes generative AI inference performance and economics at scale.

The company's core value proposition is straightforward: reduce AI workload cost and complexity through intelligent compute virtualization, parallelization, and multi-LLM orchestration — without tying customers to a single hardware vendor. With third-party validation from Artificial Analysis, a growing patent portfolio, and a 25 MW NeoCloud initiative in India, Cogniware is positioned to capture the software-defined infrastructure layer of the AI stack — where no dominant winner has yet emerged.

HEADQUARTERS

Washington, DC Metro

FOCUS

AI inference middleware
& multi-LLM
orchestration

STRATEGIC ASSET

25 MW NeoCloud · MIDC
Taloja, Navi Mumbai

STAGE

Production-validated ·
scaling commercial
deployment

WHY COGNIWARE — CAPABILITY MATRIX

01

**Cogniware
Middleware**

Cuts AI cost at scale by increasing usable throughput, lowering latency under load, and reducing infrastructure required per workload.

02

**CogniDREAM
Platform**

Improves quality and business value through multi-model reasoning, intelligent orchestration, and more accurate enterprise outcomes.

03

**AI Strategic
Services**

Accelerates enterprise AI readiness — reduces execution risk, proves ROI on the right KPIs, scales teams into production faster.

04

**NeoCloud
Architecture**

Enables superior NeoCloud economics through AI-native design optimized for density, resiliency, sovereign workloads, and long-term flexibility.

05

**Benchmark-
Backed IP**

Supports a credible performance narrative — stronger high-concurrency throughput, resiliency, and user experience under tested conditions.

INDEPENDENT THIRD-PARTY VALIDATION

VALIDATED BY

Cogniware Middleware *outperforms*
NVIDIA baselines under load.

 Artificial
Analysis

MODEL Llama 3.3 70B

HARDWARE 8 × NVIDIA H200

DATE May 2026

~14.7 KPEAK THROUGHPUT ·
TOKENS/SEC**+34% faster** vs NVIDIA
baselines**99.4 %**RESPONSE RATE AT
HIGH CONCURRENCYAA baseline: **86.8%** ·
+12.6pp**<1.0 S**TIME TO FIRST TOKEN AT
SATURATION**+91.4%** TTFT improvement
vs baselineSource: **Artificial Analysis** — independent benchmarking authority for AI inference performance.

VERIFIED · MAY 2026

NEOCLOUD & STRATEGIC INFRASTRUCTURE

Building India's *AI-native sovereign* data center
capability.

Software-driven economics meet purpose-built infrastructure. A25MW NeoCloud anchor in Navi Mumbai delivers density, sovereignty, and execution capability for the next decade of AI workloads.

25 MWNAVI MUMBAI · MIDC
TALOJA**4×**GPU VENDOR
ARCHITECTURES**800G**AI FABRIC · RDMA /
ROCEV2**Software-Driven ROI**

Middleware turns NeoCloud capacity into a premium service — higher GPU utilization, superior revenue per megawatt.

**Sovereign AI Design**

Physical & logical isolation, dedicated controls, data residency for India's regulated sectors — BFSI, government, enterprise.

**Progressive Liquid Cooling**

High-density rack architecture purpose-built for GPU intensive AI inference and training at scale.

**Strategic Partnership**

India market entry with infrastructure execution capability.

EXECUTION, TEAM & STRATEGIC ENGAGEMENT

A team built to *execute* the AI infrastructure software thesis.

EXECUTION CONFIDENCE

IP PROTECTION

Patent-pending **Virtual Compute Nodes & Inference Orchestration** — core defensible moat.

THIRD-PARTY PROOF



Independent validation by **Artificial Analysis** (May 2026) — recognized AI benchmarking authority.

TEAM DEPTH

150+ combined years from Appian, SAP, AWS, Accenture, Capital One & MicroStrategy.

ADVISORY STRENGTH

Two-time IPO CMO (**Kevin Spurway**); Head of Products at Qlik (**Malcolm Ross**); **120+ advisory years**.

COMBINED OPERATING & ADVISORY PEDIGREE

Appian SAP AWS Accenture Capital One MicroStrategy Qlik

STRATEGIC ENGAGEMENT

The AI infrastructure software *moment* is now.

Enterprises and hyper scalers face acritical choice: pay premium rates for undifferentiated inference capacity, or deploy software intelligence that unlocks superior economics from the hardware they already own. Cogniware is that software layer — independently validated, patent-protected, and ready to scale.

INDEPENDENTLY VALIDATED

PATENT-PROTECTED IP

25 MW NEOCLOUD ANCHOR

PRODUCTION-READY

Contact Us

GALAXY TECHNOLOGY SERVICES PTE. LTD.

68 Circular Road, #02-01,
Singapore 049422

GALAXY TECHNOLOGY SERVICES SDN. BHD.

No. 361, Tingkat 1, Jalan Danau Niaga 1,
Taman Danau Kota, Setapak, Kuala Lumpur
w.p. Malaysia 53300

GALAXY TECHNOLOGY SERVICES LLC

No 2520 Adams Pond Lane Apex, North Carolina,
US 27523

GSPARK TECHNOLOGY SERVICES PVT LTD

16 North St, Palayapalayam, Tamilnadu,
India 609101



contactus@gtsg.in



gtsg.in