Compute & Deployment for AI Employees: From Model Specifications to Task-level SLOs

Select cloud, edge, endpoint and data-center combinations using real concurrency, context, retrieval and tool execution

Online technical whitepaper

6 min read

Published / updated
2026-11-22

Intended readers

Infrastructure leaders, architects, operations and security teams, and AI procurement professionals

EXECUTIVE ANSWER / DIRECT ANSWER

AI hardware selection should begin with task-level service objectives: under a specified model, context, concurrency and tool chain, do time to first token, generation speed, success and cost per task meet business needs? Peak accelerator compute or memory capacity is only a constraint; neither proves user experience, stability or project outcomes.

01 / KEY QUESTIONS

Key questions answered in this paper

Q1

Why not procure on peak GPU compute alone?

Agent workloads also depend on the model, quantization, context, memory bandwidth, batching, retrieval, networking, tools and orchestration. They require end-to-end measurement.

Q2

Is on-premises deployment inherently safer and cheaper?

No. Security depends on the full architecture and operation. Cost includes procurement, power, cooling, operations, upgrades and idle capacity, alongside data and risk requirements.

Q3

What is the most important PoC output?

Reproducible task-level latency, quality, stability, resource and cost evidence under fixed versions and data conditions—not a one-time demonstration.

02 / Full paper

01

Describe the workload before the device

An AI employee workload is a task chain of retrieval, planning, tools and review—not one model inference.

Knowledge Q&A, long-form research, coding and multi-role collaboration stress computing differently. Input length affects prefill, output length affects decoding, tool calls introduce waits and dependencies, retrieval adds storage and vector work, and multi-role flows can duplicate context and create concurrency peaks. Record typical and peak shapes for each task class.

CAICT industry research continues to examine coordination among models, compute, data and applications. For procurement, this means a joint model–system–task view: specify model and precision, serving framework, knowledge base, VROS Agent runtime, Workers, network and storage rather than treating hardware as an isolated box.

Workload variable	Conditions to capture	Impact
Model & precision	Parameter scale, quantization and context limit	Memory, compute and quality
Request shape	Input/output length, concurrency and peaks	Queueing, throughput and tail latency
Agent flow	Turns, retrieval, tools and retries	Total duration and cost per task
Service objectives	Interactive wait, batch window and availability	Topology and redundancy

Chapter sources [1] [2] [4]

02

Cloud, edge, endpoint and data center form a portfolio

Placement follows data boundaries, latency, elasticity, operating capability and continuity—not ideology.

Desktops and workstations fit development and small interactive workloads. Office or edge nodes can serve local knowledge and low-latency tasks. Data centers centralize service and resource governance, while cloud provides elasticity and managed capabilities. Hybrid designs are valid, but data, models, logs and identity crossing environments need consistent governance.

AIBotStore’ s public approach considers VROS Agent integration with FusionXpark, TokenBox, FusionOne AI and FusionServer as multiple compute forms. These are candidate deployment forms, not default promises for any model, scale or performance. Devices, compatibility and available capabilities require project validation.

- ✓ Sensitive data and stable boundaries: assess local or private-cloud paths.
- ✓ Variable demand: assess elastic resources and peak overflow.
- ✓ Edge scenarios: test disconnected operation, remote administration and physical security.
- ✓ Hybrid deployment: map model, knowledge, event and log flows between environments.

Chapter sources [1] [5] [3]

03

The serving stack turns hardware into experience

The same device can produce different quality and throughput under another runtime, batching or quantization policy.

The serving stack includes model weights and precision, inference engine, memory management, continuous batching, KV cache, queues, routing, observability and model updates. Techniques such as context caching or speculative decoding require model- and SLO-specific validation; they are not unconditional gains.

VROS Agent needs clear interfaces to model serving, with timeouts, retries and circuit breakers. Retrieval and Worker sandboxes need separate resource budgets so tool work cannot exhaust inference capacity. After engine or quantization upgrades, retest answer quality, tool-call formats and long-context stability as well as speed.

- ✓ Pin model, runtime, driver and dependency versions for reproducible tests.
- ✓ Separate interactive and offline batch queues to prevent contention.
- ✓ Set resource and health measures for model, retrieval, Workers and storage.
- ✓ Evaluate every optimization against quality, latency, throughput and failure behavior.

Chapter sources [4] [6] [2]

04

Measure TTFT, TPOT, throughput and tail latency

Tokens per second alone hides first-token wait, queueing and the slowest requests. Metrics need test conditions.

TTFT measures the wait to the first output token; TPOT measures intervals between subsequent tokens; throughput describes tokens or requests processed per unit time; tail latency exposes high-percentile waits. Agent evaluation also needs end-to-end task duration, tool waits, retries and final deliverable acceptance. Comparisons are invalid without input/output lengths and concurrency.

MLPerf Inference provides standardized inference benchmark methods for data-center and other scenarios, useful for understanding test design. Third-party device results do not become AIBotStore performance commitments. Project acceptance must be independently measured on the specified hardware, software, model, data and task set.

Metric	Question answered	Conditions to report
TTFT	How soon does the first output appear?	Input length, concurrency, queue and percentile
TPOT	Is continuing generation fluid?	Output length, decoding and batching
Throughput	What total processing capacity is available?	Quality constraint, request shape and window

Metric	Question answered	Conditions to report
Tail latency	How do the slowest requests behave?	P95/P99, peaks and timeout policy
Task-level result	Does complete work finish on time and pass?	Retrieval, tools, retries and human review

Chapter sources [4] [7]

05

Local inference still needs security and operations

A device inside a facility does not remove model, API, supply-chain or human risk.

Protect model weights, knowledge indexes, event logs and caches; restrict management and inference planes; and treat external documents and tool output as untrusted. Worker sandboxes, network egress, secrets, image and dependency review, backup, recovery and log retention belong in one security design.

Operations should also inspect temperature, power, memory, disk, queues, model loading, node drift and service degradation. NIST AI RMF and post-deployment monitoring methods can organize risks and runtime issues, but they do not certify a device or establish compliance. High-impact actions retain human confirmation and a working stop path.

- ✓ Keep identities and secrets out of model context, images and public logs.
- ✓ Separate administrative and inference privileges; approve and audit remote operations.
- ✓ Retest model, plug-in, driver and system updates in an isolated environment.
- ✓ Define failover, degradation, recovery and human takeover before failure occurs.

Chapter sources [3] [8] [5] [7]

06

Close procurement with TCO and PoC

The meaningful cost unit is an accepted task, not a device price or theoretical token rate.

TCO includes devices or cloud resources, networking and storage, power and cooling, software adaptation, model and knowledge maintenance, monitoring, security, spares, upgrades and personnel. Benefits must use actual workload, quality and human baselines; target productivity or unverified savings are not facts.

Freeze PoC conditions and cover typical and peak tasks, sustained operation, exception recovery and change retests. Deliver a compatibility matrix, task quality, latency distribution, resource curves, failures, cost model and operating responsibilities. If thresholds are missed, narrow scope, change architecture or stop rather than hiding behind averages.

- ✓ Define service objectives and acceptance data before candidate configurations.
- ✓ Compare deployment options under identical models and tasks.
- ✓ Report averages and percentiles while retaining failures and human rework.
- ✓ Contract for expansion, upgrades, exit and asset handover.

Chapter sources [4] [5] [7] [1]

---

03 / CHECKLIST

Pre-project checklist

Use these items to structure a PoC plan, procurement questions or acceptance checklist.

01	Profile models, context, concurrency, retrieval, tools and retries
02	Select cloud-edge-endpoint placement by data, latency, elasticity and operations
03	Pin serving-stack versions and test quality together with performance
04	Report TTFT, TPOT, throughput, tail latency and accepted task outcomes
05	Secure weights, knowledge, caches, logs, Workers and management interfaces

06 Calculate cost per accepted task and run sustained PoC under real conditions

04 / FAQ

Frequently asked questions, answered directly

01 **What is the difference between TTFT and TPOT?**

TTFT measures the wait for the first output token; TPOT measures the interval between subsequent tokens. Both need concurrency and text-length conditions to be meaningful.

02 **Can third-party MLPerf results serve as project acceptance?**

Not directly. MLPerf informs standardized inference testing, but a project still needs measurements on its own models, data, versions, tasks and service objectives.

03 **Does TokenBox or another hardware platform guarantee fixed performance?**

No. Adaptation and performance depend on the device, model, precision, serving stack and task conditions, and must follow compatibility testing and project acceptance.

04 **Is on-premises always better than cloud?**

There is no universal answer. Local, cloud and edge deployments involve different data, elasticity, latency and operating trade-offs, best compared through task-level SLOs and TCO.

05 / REFERENCES

References and public sources

Chinese references follow GB/T 7714 style and English references follow APA style. Accessed November 22, 2026. External material informs industry and technical methods; it does not imply endorsement or certification of AIBotStore.

[1] China Academy of Information and Communications Technology. (2026). Artificial intelligence industry development research report [Report].
https://www.caict.ac.cn/kxyj/qwfb/bps/202602/P020260202487301304903.pdf →

[2] China Academy of Information and Communications Technology. (2026). AI Native research report 1.0: Enterprise transformation framework and practice [Report].
https://www.caict.ac.cn/kxyj/qwfb/ztbg/202610/P020261022558735347833.pdf →

[3] China Academy of Information and Communications Technology. (2026). Artificial intelligence security governance blue book [Report].
https://www.caict.ac.cn/kxyj/qwfb/bps/202601/P020260109784447548497.pdf →

[4] MLCommons. (2026). MLPerf Inference: Datacenter scenarios and metrics.
https://mlcommons.org/benchmarks/inference-datacenter/ →

[5] National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0).
https://www.nist.gov/itl/ai-risk-management-framework →

[6] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1).
https://doi.org/10.6028/NIST.AI.600-1 →

[7] Rao, A., Keller, A., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., Staheli, D., & Bergman, A. (2026). Challenges to the monitoring of deployed AI systems (NIST AI 800-4).
https://doi.org/10.6028/NIST.AI.800-4 →

[8] OWASP GenAI Security Project. (2025). Agentic AI—Threats and mitigations.
https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/ →

Scope statement

This whitepaper supports technology selection and project planning and does not replace legal, compliance, safety or investment advice. Actual VROS Agent, AIBotStore, hardware-integration and controlled-evolution scope depends on the delivered version, project contract and acceptance. Industry task examples do not represent delivered customers or achieved outcomes.