HFS

HOT TECH

HFS Services-as-Software™ Hot Tech: Autonomy.Computer

Author and program lead:

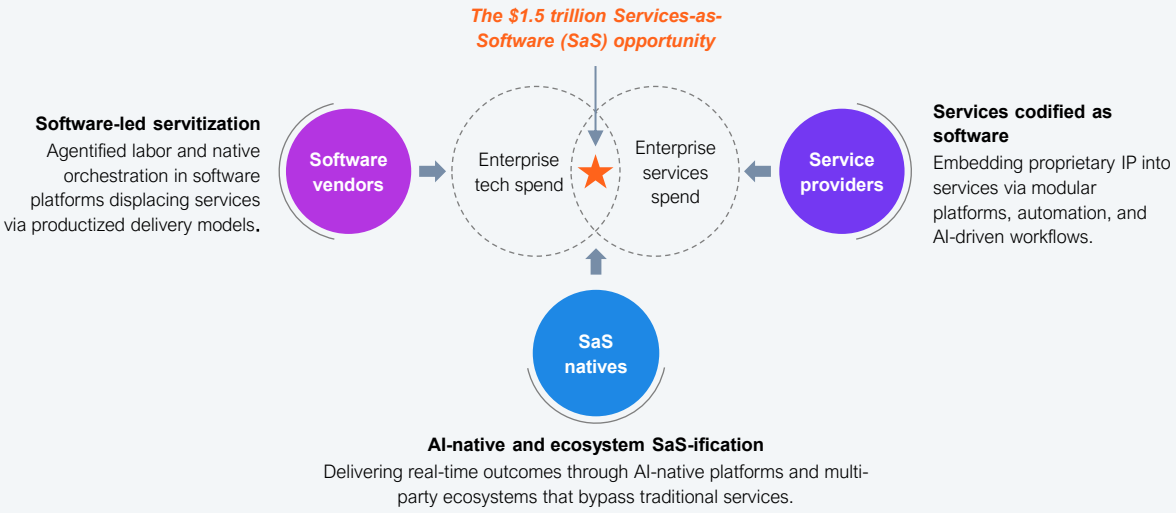
David Cushman, Hot Tech Editor-in-Chief, Executive Research Leader

Executive summary

HFS Hot Tech is an exclusive group of emerging players, each with a differentiated value proposition aligned with creating momentum toward Services-as-Software™. If you're a CTO or Head of AI watching agent pilots succeed in sandboxes but fail in production, this is for you.

HFS Research coined "Services-as-Software" (see Exhibit 1) to encapsulate a concept reshaping how the world will consume technology services and software. This emerging category will disrupt traditional services and software models, absorbing significant revenue from both, and create a new total addressable market worth $1.5 trillion.

Exhibit 1: HFS Hot Techs offer a range of approaches toward Services-as-Software



Source: HFS Research, 2025

HFS analysts regularly speak with numerous exciting startups and emerging players. We designate a select few as HFS Hot Tech based on their offerings' distinctiveness, ecosystem robustness, client impact, financial position, and—in this case—their impact on Services-as-Software.

HFS Hot Techs may not have the scale and size of more established rivals, but they have the vision and strategy to impact and disrupt the market. In the rapidly changing AI-led operations space, enterprises realize they can't be everything to everyone.

Enterprises consuming third-party services, service providers, and technology providers need a smart ecosystem to succeed and survive in the future. HFS Hot Techs are service and technology providers handpicked by our analysts to help you flesh out your ecosystem with offerings that solve today's complex business problems and exploit market opportunities.

HFS Hot Tech organizations display truly differentiated offerings and out-of-the-box thinking that can be inspiring and useful. This report profiles one of the HFS Hot Techs selected through our rigorous five-step assessment. The HFS Hot Tech designation remains in place for one calendar year. Every Hot Tech joining our program remains listed on our exclusive and searchable database. https://www.hfsresearch.com/hv/

Stop burning cash on agent architectures that can't scale

AUTONOMY

Your AI agents work in demos. Then you scale to thousands and watch costs explode while performance collapses. The problem is that you're building agents like microservices when they should be state machines. Autonomy.Computer's actor-based architecture runs 5,000–10,000 concurrent agents per container, cutting a multi-day run to 15 minutes while securing every interaction with cryptographic identity.

The San Francisco-based firm is part of a new wave of emerging tech firms focused on the shift from “agents that work” to agents that operate efficiently and securely while controlling resources to deliver commercial viability.

What does this mean in practice? One enterprise customer needed to process 4,000 image and video files in a complex analysis pipeline. Using Autonomy.Computer, a multi-day run was cut to just 15 minutes.

Another found that a loan approval workflow built with big tech vendors worked in simple demos but fell over when confronted with messy real-world data, dynamic states, and high concurrency. Autonomy.Computer’s deep-work agents orchestrated the full lifecycle, collecting documents, running checks, interacting with internal APIs, and processing thousands of concurrent cases. Each case had its own durable memory in a system that behaved like a team of diligent loan officers with great recall and audit trails.

The head of engineering at a financial services company told us, “In less than an hour, we built a customer end-to-end solution. I wish we had other platforms that were this easy to deploy to.”

Microservice agents waste 90% of runtime waiting, not computing: here's the math that proves it

Any startup or enterprise team can sign up, build, and deploy agents directly on the platform, with no special onboarding.

Autonomy.Computer operates in the agentic execution layer of the AI stack, above model providers such as OpenAI and Anthropic, alongside model-centric and agent abstraction tools like LangChain and Crew AI, and below workflow SaaS and industry-specific solutions. It focuses on runtime, secure communication, and the scaling of continuous autonomous workflows.

The team believes the architecture of most current agent systems is fundamentally inefficient because builders treat each agent as a standalone container, spinning up thousands of isolated processes. Instead, an agent should be viewed as a long-lived state machine (which remembers what it has already done) rather than a microservice (which has to reload context for every request).

Microservice-style agents spend most of their time waiting for an LLM response, a tool call, or an external event. So “traditional” designs waste resources by keeping entire containers or serverless functions “alive” while they’re idle—and burning money. Such agents don’t actually compute continuously; they’re mostly waiting.

5,000—10,000 agents per container: Why actor models deliver 100x density over traditional architectures

Autonomy.Computer's architecture is based on the "actor" model. Each agent is implemented as an actor: a lightweight, stateful process that communicates asynchronously with other actors. Thousands of these can run concurrently inside a single small container because they use minimal CPU while idle.

This means Autonomy.Computer can support fleets of 5,000–10,000 concurrent agents per container. This actor-based approach treats agents as persistent digital entities that communicate and evolve rather than as disposable function calls.

Every agent breach costs $4.45M on average—cryptographic identity stops it before it starts

Security is a design principle. Each agent has its own cryptographic identity, its own mutually authenticated communication channels, and its own isolated execution environment.

This means agents prove "who" they are to each other with verifiable identity at every interaction. Sensitive data never moves through shared pipes, tools connect through encrypted portals, and all cross-cloud or on-prem links are secured end-to-end.

Net-net, each agent acts like a fully verified, access-controlled digital employee, thereby reducing the risk of data leakage.

HFS' take

If you're still running agents as isolated microservices, you're overpaying by 100x and exposing every workflow to security gaps your auditors will catch before your customers do. Autonomy.Computer's actor-based architecture isn't a nice-to-have optimization—it's the foundation for running agents that actually generate ROI instead of burning OpEx. The enterprises moving first on this shift will lock in 18-24 month advantages in both cost structure and time-to-market. Waiting means watching competitors deliver Services-as-Software contracts you can't match.

Vendor factsheet

- **Founded:** 2025
- **Headquarters:** San Francisco, CA (USA)
- **Key executives:** Matthew Gregory, Co-founder & CEO, Mrinal Wadhwa, Co-founder & CTO
- **Employees:** Not disclosed
- **Funding:** Not disclosed
- **Partners:**
  Identity: Okta, WorkOS,
  Inference and GPU: AWS, CoreWeave, Lambda,
  Models: OpenAI (via inference partners),
  Data: Databricks, Snowflake, Box,
  Systems integrators: MasterBorn, expanding systems integrator ecosystem
- **Clients:** Mix of vertical AI startups and large enterprises, with early traction via Box-led partner projects in financial services and life sciences
- **Core domains:** Agentic runtime and agentic sandbox, Actor-model concurrency, Secure non-human identity, Agent orchestration, NeoCloud and Services-as-Software enablement
- **Solution set:** Autonomy.Computer platform-as-a-service, Actor-model compute engine, Autonomy.Computer framework and CLI, Non-human identity and secure messaging, Agent context, memory, and state, Tooling and orchestration for LLMs, data, APIs, MCP, PrivateLink for cross-cloud and on-prem, Serverless runtime
- **Industries:** Financial services (loan approvals, compliance), Life sciences and pharma (regulated document translation), Recruitment and HR, Procurement, Security and IT operations, Broader document- and media-heavy vertical AI
- **Go-to-market:** Self-serve platform-as-a-service with bottom-up adoption, Land-and-expand in enterprises, Partners and SIs use Autonomy.Computer as the execution layer in their own solutions

The HFS Hot Tech report author and program lead

David Cushman
Hot Tech Editor-in-Chief,
Executive Research Leader

David is an Executive Research Leader at HFS and Editor-in-Chief of the HFS Hot Tech program. He also leads our OneOffice™ Emerging Technology Practice, is our strategic lead on Generative AI, Web3, and metaverse, and covers automation and employee experience. He is a published author (*The 10 Principles of Open Business*, Palgrave-Macmillan), a former tier-1 consulting director, and a digital strategy and innovation expert with leadership experience in start-up, scale-up, and enterprise digital transformation.

HFS

About HFS

- **INNOVATIVE**
- **INTREPID**
- **BOLD**

HFS Research is a leading global research and advisory firm helping Fortune 500 companies through IT and business transformation with bold insights and actionable strategies.

With an unmatched platform to reach, advise, and influence Global 2000 executives, we empower organizations to make decisive technology and service choices. Backed by fearless research and an impartial outside perspective, our insights give you the edge to stay ahead.


www.hfsresearch.com


hfsresearch


www.horsesforsources.com


www.horsesmouthpodcast.com