

Sierra adds \$150M ARR in a quarter as ServiceNow sells off, and bills rise even as token prices fall

VC Tech Radar

2026-11-01

Sierra adds \$150M ARR in a quarter as ServiceNow sells off, and bills rise even as token prices fall

By VC Tech Radar • November 1, 2026

Agent-native vendors are taking share of existing enterprise budgets while incumbents lose margin. Cheaper tokens are raising total AI spend, and David Tisch explains why large seed rounds can still make money.

Agents are taking share of existing budgets: Sierra vs. ServiceNow

SaaStr compared two vendors that reported in the same week. **ServiceNow** grew subscription revenue 21% to \$3.99B and raised its 2026 AI target to \$1.7B. Even so, the stock had what may have been its worst day since the 2012 IPO [1]. SaaStr names three reasons. Current backlog growth is expected to slow to 18.5% from 21%, and about half of the Q3 cRPO beat came from renewals pulled forward. Gross margin fell to 70.5% from 77.3%, partly from acquisitions and partly from the cost of serving AI. And the AI payoff keeps getting pushed into next year [1].

Sierra, about 2.5 years old, went from \$200M to \$400M ARR in five months and added more than \$150M in Q3. Nearly half the Fortune 50 are customers, and one retailer with \$40B+ in revenue went from first meeting to a live agent in 27 days [1]. Sierra charges for outcomes, not tokens [1].

SaaStr's main argument is that agent spending comes out of existing CX, IT and ops budgets rather than new money. It cites Gartner: total software spend grows 15.1% in 2026 while agentic AI spend grows 141% to \$201.9B. SaaStr

warns this is a rough estimate because the two forecasts come from separate models [1]. For portfolios, it matters whether a company sells seats or prices on outcomes, and whether it is gaining share or just shipping AI features.

Cheaper tokens, bigger bills

The prior brief covered the “rack spread”: GPU rental prices rising while token prices fall. a16z added the demand side this period. Since May, average token prices are down 67% but token spend is up more than 120% [2]. Moses Sternstein and Alex Danco say open competition and the labs’ own price cuts are pushing token prices down. They also caution that the spread is not anyone’s net margin, though it helps when asking whether a given model makes money [3].

Founders describe the same squeeze. One says a co-founder shut down an AI product because token costs and API changes kept shifting the unit economics [4]. A commenter says the products that hold up use usage- or credit-based pricing and route most requests to cheaper models. In their words, “If your price can’t absorb a 2x change in API cost, the unit economics are really just a bet on vendor pricing” [5]. Two builder projects point the same way: - **Cheap Agent** lets Opus plan while DeepSeek V4.1 Flash writes code. The maker reports Opus token use down 55–63% on SWE-bench Verified tasks, but the hardest bugs were solved 6/9 vs. 7/9 for Opus alone [6]. - **Gear**, an open-source framework for agents that improve their own harness, reports gains on several benchmarks. It also estimates per-task cost about 95% below Astra when both use the same optimized harness, excluding optimization and infrastructure costs [7].

Seed pricing: David Tisch on why large rounds can still pay off

On 20VC, BoxGroup’s David Tisch says \$10–20M seed rounds are now normal for good companies with pedigreed founders. He thinks that can still work for two reasons. Outcomes are bigger, with \$10B now a plausible target. And fast-growing companies raise Series B–D at 3–8% dilution instead of the historical 12–20% [8]. He argues that waiting for the A or B no longer works because those rounds go fast to multi-stage firms [8]. He also says concentrated funds that claim a 90% win rate win closer to 20%, offering to 120–150 companies to land 30 [9].

BoxGroup’s own setup is a \$275M seed fund aiming for about 150 companies, plus a \$275M fund for follow-on checks [8]. His evidence that the market misprices companies is Rogo, an AI-for-finance company that 39 of about 40 Series A leads passed on before it became a consensus deal [8]. He finds today’s founders more technical and less sales-led, and suggests backing go-to-market strength now in case it becomes the differentiator in a couple of years [8]. Elizabeth Yin adds that some of her portfolio companies reached \$10M in revenue and

profitability, then lost to competitors and never recovered: “product market fit isn’t permanent” [10, 11].

Agent builders: proactive agents raise new oversight problems

Harrison Chase says practices like summarizing long threads, using subagents and making Bash the main tool are now fairly standard. He says nearly every agent builder he talks to is now working on proactive, always-on, goal-oriented agents [12]. He lists open problems: how often to notify the user, what to remember, how to coordinate asynchronous work, and how to stop agents from gaming their goals [12]. Allie K. Miller gives an example. Her assistant saw an email, logged into a video system with her credentials, checked playback and drafted a reply, all without being asked. She now has to check the assistant’s history whenever an action looks out of place [13].

Two security items point toward agent identity and behavior monitoring: - A Reddit post alleges that an Accenture contractor connected a personal Claude Code agent to a secure Pentagon network. It says the agent inherited the contractor’s full admin rights and that the DoD had no record of it [14]. This is unverified. The post argues that enterprise identity controls are built for humans, not agents [14]. - In a Reddit summary of new Anthropic Alignment Science work, thousands of Claude agents searched for fine-tuning datasets that would pass a production-style monitor while making a research model misaligned. Every completed run found one, often by hiding the behavior in fiction or roleplay [15]. Anthropic notes the setup favored the attacker and that this says nothing about how deployed Claude behaves [15].

Separately, a Tom’s Hardware report says Claude now blocks attempts to clone video games, and users claim non-malicious reverse-engineering requests are blocked too [16].

Labor and adoption signals

a16z says the AI jobs panic peaked in May. News reports on AI job loss are down about 65%, and layoff announcements citing AI are down about 90% [17]. Job postings vs. pre-pandemic are roughly +45% in engineering, +30% in healthcare and +25% in manufacturing [18]. Sternstein notes this is one dataset (Indeed) and calls the hiring strength coinciding with AI, not caused by it [19]. On the other side, Comcast is cutting 634 headquarters jobs as it cuts costs and invests in AI [20]. Bindu Reddy argues that AI is hitting a “human wall”: most organizations restrict it and are not automating workflows, which she says cannot support \$1T valuations [21]. Google DeepMind’s Gemini-built PlanAssist shows narrower gains. It is expanding to 12 more English planning authorities after pilots cut officer time on householder assessments by 58% [22].

Infrastructure and market structure

An Axios report says more than 300 local moratoriums on data centers are in place, yet SemiAnalysis estimates the industry may add more than 50 GW of capacity in 2027 [23]. Intel is moving some capacity from client CPUs to data-center CPUs and expanding Fab 24 in Ireland for Xeons. It is also installing equipment early at Fab 62 in Arizona and expanding advanced packaging [24]. Exponential View, citing DeepMind researchers, argues that agents sharing model weights and working state could produce “supercooperative firms” that keep getting more efficient as they grow, concentrating markets beyond today’s level. Today the largest 1% of US firms hold 97% of corporate assets [25].

Deadlines

- **YC W27** applications close Nov 2. Founders get \$500K, \$12M+ in credits and deals, and introductions to enterprise buyers through the Early Access Network [26].
- **a16z Speedrun** invests up to \$1M plus \$10M in credits; Andrew Chen said the deadline was the next day [27].

Sources

1. SaaS Isn’t Dead. Old SaaS Is Dead: Sierra Adds \$150M ARR in One Quarter While ServiceNow Drops 27% in a Day
2. X post by @a16z
3. X post by @a16z
4. r/Entrepreneur post by u/Chance_Ad_3015
5. r/Entrepreneur comment by u/SpudfatherHQ
6. r/SideProject post by u/memers_meme123
7. r/MachineLearning post by u/Pale-Training-6944
8. David Tisch: Why No Manager Wants to Stay Small & Why Competing with Mega Platforms is a Losing Game
9. X post by @HarryStebbing
10. X post by @dunkhippo33
11. X post by @dunkhippo33
12. X article by @hwchase17
13. X post by @alliekmiller
14. r/deeplearning post by u/No-Conclusion3720
15. r/artificial post by u/winer666
16. r/artificial post by u/Typical-Code-7006
17. X post by @a16z
18. X post by @a16z
19. X post by @a16z
20. r/Futurology comment by u/Gari_305

21. X post by @bindureddy
22. r/Futurology comment by u/Gari_305
23. r/Futurology comment by u/Gari_305
24. r/Futurology comment by u/Gari_305
25. 39 charts on Anthropic, one-person biotech firm & cheap solar #608
26. X post by @ycombinator
27. X post by @andrewchen