The End of Warm Compute

Warm infrastructure exists because true cold creation was too slow. Setsuna changes that premise.

88 ms p50

16 GIB FIRECRACKER MICROVM · REQUEST TO VERIFIED READY · MEDIAN OF 1,000 TRUE-COLD LAUNCHES

01 · THE CATEGORY THESIS

Warm compute is a workaround for slow cold starts.

For decades, infrastructure paid an operational tax to make a machine look instant.

01

Snapshot the machine and restore it later.

02

Keep machines booted in a warm pool.

03

Preinstall, preboot and hold idle inventory.

04

Build a second cold path for misses, invalidation and failures.

02 · THE HIDDEN PRODUCT

The workaround became an infrastructure product of its own.

Once creation is slow, teams must operate state artifacts, pool inventory and two lifecycle paths.

01

Capture, version, distribute and invalidate snapshots.

02

Secure guest-memory and machine-state artifacts.

03

Size pools for peaks; refill, health-check, drain and recycle them.

04

Handle stale state, compatibility and cold-path misses.

05

Debug behavior that differs between restored, pooled and fresh environments.

03 · WHY AGENTS BREAK THE WORKAROUND

Agents turn occasional cold starts into a core systems problem.

Thousands of short-lived, parallel, untrusted environments now sit inside interactive and training loops.

- One user task can trigger dozens of execute-inspect-iterate steps.
- Eval and RL rollouts fan out into thousands or millions of environments.
- Burst demand makes pool sizing structurally inaccurate.
- Generated code raises the value of a clean hardware-isolated boundary.

40%

of enterprise applications expected to include task-specific agents by end of 2026

SOURCE · M1

$206.5B → $376.3B

AI-agent software spend, 2026 → 2027

SOURCE · M2

$7.84B → $52.62B

AI-agents market, 2025 → 2030

SOURCE · M4

DIRECTIONAL THIRD-PARTY MARKET EVIDENCE FOR THE AGENT CATEGORY — NOT A SETSUNA TAM.

04 · THE NEW PREMISE

Setsuna makes fresh creation faster than many restored or pooled service paths.

A 16-GiB Firecracker microVM reaches verified readiness without snapshotting or warm inventory — every launch, every time.

01 New Firecracker process

02 New KVM virtual machine

03 Fresh guest kernel boot

04 Verified networking, memory and workspace

05 Destroy

88 ms

MEDIAN REQUEST-TO-VERIFIED-READY · 1,000 TRUE-COLD LAUNCHES · 16 GIB GUEST

05 · PROOF

The public record had no comparable documented true-cold result this fast.

Every published run — sequential true-cold launches per host and guest shape, timed from request to strict-ready. Raw samples and checksummed bundles are downloadable on the proof page.

HOST	GUEST	LAUNCHES	P50	P95	P99	MAX	RUN
Azure Standard_D32ads_v7	32 vCPU · 64 GB	1,000/1,000	293 ms	309 ms	319 ms	375 ms	Aug 27, 2026
Azure Standard_F8amds_v7	4 vCPU · 32 GB	1,000/1,000	126 ms	132 ms	135 ms	145 ms	Aug 29, 2026
Azure Standard_F16ads_v7	8 vCPU · 32 GB	1,000/1,000	145 ms	152 ms	154 ms	159 ms	Aug 29, 2026
Azure Standard_F8amds_v7 CANONICAL	2 vCPU · 16 GB	1,000/1,000	88 ms	95 ms	99 ms	102 ms	Aug 29, 2026
Azure Standard_E16ads_v7	2 vCPU · 16 GB	1,000/1,000	95 ms	102 ms	104 ms	112 ms	Aug 27, 2026
Azure Standard_E8ads_v7	2 vCPU · 16 GB	1,000/1,000	96 ms	104 ms	107 ms	113 ms	Aug 24, 2026
Azure Standard_D8ads_v7	2 vCPU · 16 GB	1,000/1,000	97 ms	105 ms	109 ms	129 ms	Aug 27, 2026
Azure Standard_F8amds_v7	4 vCPU · 16 GB	1,000/1,000	98 ms	102 ms	106 ms	119 ms	Aug 29, 2026
Azure Standard_F16alds_v7	8 vCPU · 16 GB	1,000/1,000	117 ms	121 ms	123 ms	134 ms	Aug 29, 2026
Azure Standard_E8ads_v6	2 vCPU · 16 GB	1,000/1,000	191 ms	241 ms	252 ms	276 ms	Aug 27, 2026

+ 12 MORE PUBLISHED SHAPES ON SEMSWITCH.COM/SETSUNA

CANONICAL RUN · AZURE STANDARD_F8AMDS_V7

88 ms p50

P99	MAX	LAUNCHES
99 ms	102 ms	1,000/1,000

MANAGED SANDBOX CREATE, INDEPENDENTLY MEASURED — DIFFERENT CLOCKS

E2B	~717 ms	C15
Modal	~2.4 s	C15
Fly Machines	~2+ s	C14
Vercel Sandbox	~1.9 s	C15

Setsuna's timer is host-local request-to-strict-ready. Public managed-service numbers may include network and control-plane spans. Never collapse the two into one ranking.

06 · ARCHITECTURE DELETION

From stateful workaround stack to create-run-destroy.

Setsuna can remove entire systems whose only job was hiding a slow fresh start.

DELETED FOR CLEAN-SLATE JOBS

- ~~Snapshot capture, versioning and restore.~~
- ~~Warm-pool sizing, refill, assignment and draining.~~
- ~~Cold-miss behavior divergence.~~
- ~~Persisted guest-memory artifacts in the normal path.~~

→

WHAT REMAINS

01 Create a fresh microVM.

02 Run the job.

03 Destroy it.

04 Keep images, persistent volumes and caches only where the workload actually needs them.

07 · SECURITY + SPEED

Freshness without the historical latency tax.

Setsuna reduces persisted runtime state and reuse paths while retaining a Firecracker/KVM boundary.

- ✓ Hardware-virtualization isolation.
- ✓ Fresh kernel and guest userspace per execution.
- ✓ No resumed memory/CPU/device state.
- ✓ No pooled guest waiting for reassignment.
- ✓ No customer state inherited from a prior execution.

08 · WHERE IT WINS

Make warm infrastructure optional — not universally forbidden.

Setsuna is strongest where the desired state is disposable and clean; snapshots remain valuable when continuation itself is the product.

SETSUNA

Untrusted code execution, eval/RL rollouts, CI/test branches, one-shot tools, compliance-sensitive clean-slate jobs.

SNAPSHOTS / STANDBY

Exact state continuation, long-lived coding sessions, forking from initialized state, heavyweight app/model initialization.

HYBRID

Persistent data outside the VM; fresh compute around each execution step.

09 · ECONOMICS AND CHURN

The economic gain is idle capacity avoided plus engineering systems deleted.

Setsuna’s value compounds across compute, storage, operations and product velocity.

01

No idle warm guest fleet for clean-slate workloads.

02

No guest-memory snapshot storage and distribution in the normal path.

03

Fewer invalidation, compatibility and pool-health incidents.

04

One lifecycle path to test instead of warm-hit and cold-miss behavior.

05

Faster agent interactions and rollout completion.

10 · CATEGORY POSITION

Setsuna is not another sandbox dashboard.

It is the runtime breakthrough that lets agent platforms simplify their own infrastructure.

01

Hosted API for developers who want the complete service.

02

Dedicated capacity for predictable performance.

03

BYOC for enterprise control and data boundaries.

04

OEM/runtime integration for existing agent and sandbox platforms.

11 · GO TO MARKET

Sell the deletion of warm infrastructure.

Start with teams that already maintain snapshots, pools or expensive environment setup and can measure the pain.

01 Paid architecture and benchmark evaluation

02 Integrate behind the customer’s existing sandbox interface

03 Measure latency, failure paths, cost and code removed

04 Convert to a hosted, dedicated or BYOC contract

05 Publish customer-validated case studies

12 · MOAT

Every qualified host and workload compounds the system.

Setsuna turns hardware-specific systems work into a growing profile and evidence asset.

01

Private per-host/per-shape runtime profiles.

02

Qualification registry and reproducible evidence.

03

Failure knowledge and production behavior across heterogeneous CPUs/clouds.

04

Customer workload contracts and integration depth.

05

Brand authority around measurable true-cold performance.

13 · THE RAISE

$1.25M to make warm compute optional for production agent workloads.

Fund one senior systems hire, production control plane, paid pilots and customer-shaped qualification.

01 18 months of runway.

02 3–5 paid design partners.

03 Production API + BYOC pilot.

04 Validated concurrency/reliability and unit economics.

05 Next round raised on revenue and production workloads.

SEMSWITCH, INC. IS RAISING

$1.25M

INSTRUMENT

post-money SAFE

RUNWAY

18 months

PRIMARY CONTACT

investors@semswitch.com

APPENDIX A · USE OF FUNDS

Illustrative 18-month allocation.

Large enough to stop fundraising again immediately, still inside the current pre-seed range for technical infrastructure.

Founder compensation + benefits	Remove survival pressure; preserve full-time founder focus.	$190K
Senior systems / infrastructure engineer	One exceptional operator for runtime, productionization and reliability.	$300K
Later technical capacity / specialist contracting	Second hire or targeted kernel/cloud/security help after first customer signal.	$180K
Cloud, qualification and preview infrastructure	Non-Spot qualification, production hosts, evidence automation and observability.	$150K
Legal, accounting, insurance and security	Corporate hygiene, contracts, IP, SOC 2 preparation where demanded.	$80K
Design partners, travel and tools	Pilots, customer discovery, integrations and targeted GTM.	$50K
Reserve	Capacity shocks, delayed sales cycles and protection against under-raising.	$300K
TOTAL		$1.25M

TARGET NEXT-ROUND EVIDENCE

- ✓ Research preview converted into a reliable production API and restricted CLI.
- ✓ 3–5 paid design partners, with at least one production deployment or BYOC pilot.
- ✓ Production concurrency, reliability and cleanup contracts validated under customer-shaped workloads.
- ✓ A credible unit-economics model at realistic utilization.
- ✓ A repeatable qualification pipeline across the host families customers actually demand.
- ✓ A commercial signal such as $250K+ contracted ARR or an equivalent committed pilot pipeline — a target, not a claim.

APPENDIX B · HOW SETSUNA MAKES MONEY

A focused entry wedge; four monetization paths in diligence.

Hypotheses to validate with buyers — not published price commitments.

Paid technical pilot

WHO PAYS
Agent startups, labs, enterprise innovation teams

COMMERCIAL SHAPE
$10K–$25K scoped evaluation

WHY IT FITS
Fastest path to first revenue and proof of willingness to pay.

Hosted usage-based compute

WHO PAYS
Agent platforms and developers

COMMERCIAL SHAPE
Platform fee + per-second CPU/RAM/session usage

WHY IT FITS
Validated by E2B, Daytona and Runloop pricing.

Dedicated non-Spot capacity

WHO PAYS
Teams needing predictable capacity/performance

COMMERCIAL SHAPE
Monthly minimum + active host-hour or committed capacity

WHY IT FITS
Simple early operating model; customer receives a dedicated qualified host pool.

Enterprise BYOC

WHO PAYS
Regulated or high-volume enterprises

COMMERCIAL SHAPE
Annual platform license + support + usage/host fee

WHY IT FITS
Customer pays the cloud bill; SemSwitch sells control plane, runtime and qualified profiles.

OEM / runtime licensing

WHO PAYS
Agent platforms, clouds, CI/eval vendors

COMMERCIAL SHAPE
Annual minimum + per-launch or per-host royalty

WHY IT FITS
Setsuna becomes embedded infrastructure instead of competing for every end user.

Research and evidence.

Source ids in slide footers map here. Market forecasts are directional signals, not a claim that Setsuna captures the forecast market.

S1 SemSwitch — "Anatomy of a True Cold Start"

S2 SemSwitch — Setsuna benchmark explorer and proof bundles

S4 SemSwitch — canonical public proof: Azure Standard_F8amds_v7, 2 vCPU · 16 GB, 1,000 consecutive true-cold launches

T1 Firecracker specification — ≤125 ms InstanceStart to /sbin/init (1 vCPU / 128 MiB)

T2 Firecracker project — KVM microVM isolation, minimal device model, jailer

T3 Firecracker snapshot support — guest memory + VMM/KVM state; operators secure snapshot files

T4 Firecracker snapshot versioning — memory dump, VMM/KVM/vCPU state, CPU compatibility constraints

C1 E2B infrastructure architecture — "a sandbox is a resumed snapshot"

C2 E2B pricing — Pro $150/month plus per-second CPU/RAM usage

C3 E2B Series A — $21M round, $32M total (vendor-reported)

C4 Daytona pricing — per-second billing; enterprise BYOC

C5 Daytona Series A — $24M

C6 Runloop pricing — Pro $250/month plus usage

C7 Runloop funding — $7M seed

Research and evidence.

C8 Blaxel seed — $7.3M

C9 Blaxel — 200–600 ms creation from template vs sub-25 ms standby resume

C10 Modal — unpacking sandbox startup latency (warm-pool guidance)

C11 Modal — sandbox pool example (size, TTL, readiness, claim logic)

C12 Modal — cold-start guidance (min_containers, buffer_containers)

C13 Modal — sandbox snapshot limitations

C14 Fly.io — suspend/resume: full VM state; ~2+ s cold start

C15 LogRocket — AI agent sandbox platform comparison, August 2026

M1 Gartner — 40% of enterprise apps to include task-specific agents by end of 2026

M2 Gartner — AI agent software spend $206.5B (2026), $376.3B (2027)

M4 MarketsandMarkets — AI agents market $7.84B (2025) to $52.62B (2030)

F2 Carta — State of Pre-Seed Q2 2026

F3 Heavybit — typical initial pre-seed check $500K–$3M

E1 Azure F8ads_v7 retail price tracker — $0.686/hour, South Central US, Aug 25 2026