FOUNDATIONAL CONCEPTS: CONCEPT NOTE

Resident Tripwire

A per-host NG-RC monitor for novel compromise

Start Here: The Short Version

A tiny model lives on each computer. The model learns what normal looks like for that one machine. The model raises a flag the moment the machine starts to behave in a new way. The flag gives the owner an early warning and a clean record of the moment.

What it can do

- It learns the normal pattern of one host from a short record of normal data.
- It forecasts the next values in each stream.
- It flags a sharp change in the pattern at the moment the change starts.

Where people can use it

- Server teams place the monitor on each host in a small fleet.
- Small sites run the monitor on single machines with no central service.
- Test labs place the monitor on lab hosts during intrusion tests.

Problems it can help solve

- New attacks carry no known signature for a rule or a feed to match.
- Central monitors lose a host after an intruder cuts the network link.
- Single metric thresholds miss attacks that move several signals at once.

The Concept

Sense. One NG-RC model per host reads the metric streams of that host. The model trains on normal data only and scores each new value against a forecast.

Judge. One small resident agent holds the policy for the host. The agent takes a state snapshot at a flag and acts only inside owner-set rules.

Escalate. One fleet layer fuses flags from many hosts into one act or escalate call. A human receives the call with the evidence packets from each flagged host.

How It Works

1. Collect streams: process activity, outbound connections, DNS behavior, sign-ins, file change rates, processor and memory load.
2. Train the model on the first days of normal data only.
3. Forecast the next values of each stream.
4. Score the forecast error for each stream.
5. Treat a sharp error rise as a sign of a new pattern.
6. Snapshot the process tree, open connections, and recent file changes at that moment.
7. Send one compact evidence packet to the fleet layer.
8. Retrain only inside owner-approved change windows.
9. Hold a maintenance mode during deploys and patch runs.

Worked Example: A Web Server Tripwire

A blue team places one resident monitor on a production web server. A red team attacks the same server with an AI-driven intruder. The monitor reads six signals as one joint pattern: latency distribution, TCP retransmits, process fork rate, memory churn, TLS timing, and disk I/O wait. Normal traffic teaches the model the joint rhythm of the six signals. A probe, a webshell drop, or a quiet exfiltration bends several signals at once. The forecast error trips at the start of the bend, and the monitor escalates the moment with a snapshot. The new behavior matches no signature in any feed.

The strongest form runs the model below the operating system. A TPM chip seals the model weights and attests integrity. The compute runs on a DPU, a baseboard controller, or a CPU enclave. An attacker who owns the operating system falsifies its reported metrics, so the sensor reads platform counters directly.

Why Novel Attacks Fit

A new zero-day has no signature. Attack tools from AI builders change shape for each target. Behavior change on the host stays as the stable signal. The model needs zero attack examples in training.

End State

Goal. One resident monitor sits on each host of a small fleet. Each monitor trains on its own host and runs with no network link.

Proof target. Lab tests use replayed intrusions and live test intrusions. The tripwire flags the intrusion window with zero attack examples in training. False alerts stay under an owner-set daily budget.

Product at the end. Resident Tripwire, version 1, holds a host agent with a collector, an NG-RC model, a residual scorer, and a snapshot and escalation module. A fleet escalation API completes the product.

Limits

A patient attacker inside the normal envelope trips no flag. Real change looks novel, so the design needs change windows. The sensor names no culprit and gives no cause. Response actions follow owner-set policy only.