⚡ SOVEREIGN ENTERPRISE AI ARCHITECTURE & FINOPS

The Sovereign AI Runtime Fabric.

Mastrion GmbH is the runtime layer for enterprise AI. Sidecarless ambient mesh proxies, sub-2.4ms deterministic PII redaction, Radix KV prefix cache deduplication, and declarative FinOps cost routing between local sovereign silicon and frontier models.

PILLAR 01

Ambient Zero-Trust Mesh

Sidecarless eBPF socket redirection with zero app code refactoring.

PILLAR 02

Model Memory Firewall

Sub-2.4ms in-line PII redaction & EU AI Act Art. 12 SHA-256 ledger.

PILLAR 03

Compute Garbage Collection

Radix KV prefix cache reclaiming up to 62.4% redundant GPU tokens.

PILLAR 04

FinOps Cost Arbitration

Dynamic prompt routing with proven 38-day hardware payback.

The Inevitable Abstraction Layer for Enterprise AI

When software evolved from bare-metal C into modern managed runtimes, critics lamented runtime overhead. We did not abandon managed runtimes; we engineered memory bounds checkers and garbage collectors so developers could build safely at scale.

Generative AI has reached this exact inflection point. Enterprises cannot expect engineers to send raw HTTP requests to non-deterministic models while manually policing data egress, model hallucinations, and compliance exposure.

> *"Mastrion is the runtime layer for enterprise AI: the memory bounds checker that redacts confidential data before it leaves the socket, and the compute garbage collector that reclaims up to 62% of redundant GPU cycles."*

<2.4ms	62%	38 Days
In-Line PII Latency	Token Reclaim	Appliance Breakeven

networking.mastrion.io/v1alpha1 · CRD Specification

```yaml
// Declarative AI Ingress Guardrail Policy
apiVersion: networking.mastrion.io/v1alpha1
kind: AIEgressPolicy
metadata:
  name: core-banking-guardrail
spec:
  target: waypoint-ambient-proxy
  isolation: airgapped-local-vllm
  piiShield:
    enforce: true
    entities: [IBAN, SSN, CreditCard, SecretKey]
    action: MaskAndLogHash
  finOps:
    radixCache: true
    costCenter: "Engineering-Core-4810"
  compliance:
    standards: [EU-AI-Act-Art12, BaFin-BAIT, BSI-C5]
    auditSink: "immutable-ledger://compliance-vault"
```

The Enterprise Cloud Trap vs. Sovereign Control

Over 80% of enterprise AI traffic consists of code completions, internal summarizations, classification, and embeddings. Routing these routine tasks to public US cloud APIs creates severe operational failure modes.

Public model APIs expose European enterprises to unvetted prompt caching, data egress penalties under EU AI Act Article 12, and runaway token bills. Mastrion enforces architectural sovereignty by keeping routine tokens entirely inside internal infrastructure while brokering only complex reasoning to cloud endpoints under strict governance caps.

82%
Routine Token Ratio

€0.00
Local Token Cost

100%
Sovereign Airgap

Strategic Comparison Matrix

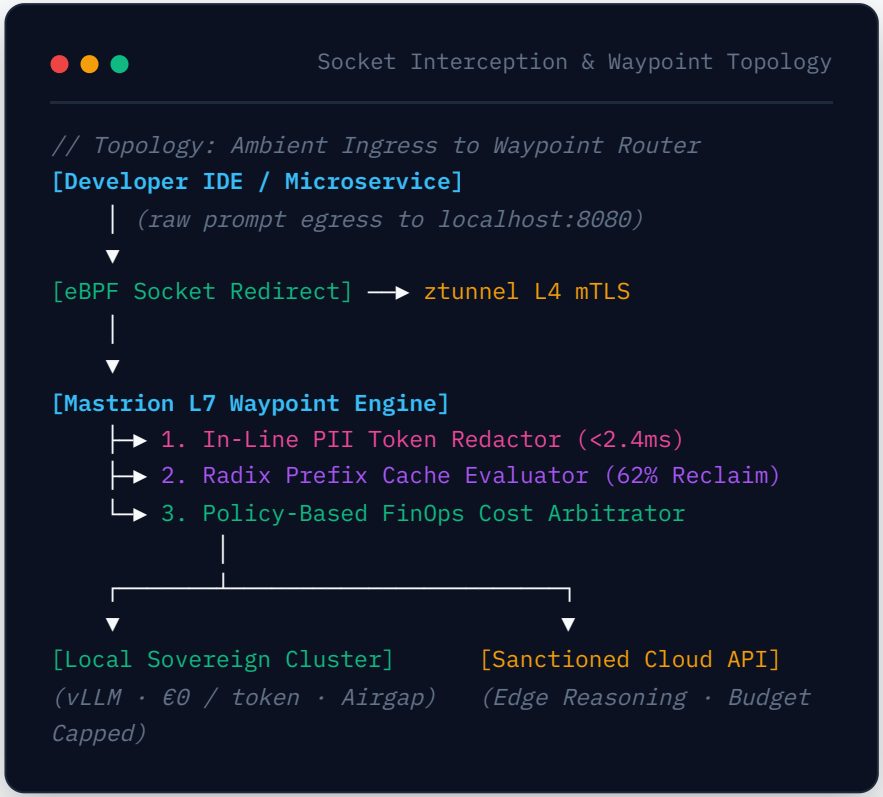
Vector	Public Cloud APIs	Mastrion Sovereign
Data Egress	US Cloud Bound	100% On-Prem / Local VPC
PII Interception	Post-Hoc Logs Only	Sub-2.4ms Socket Masking
Cost per 1M Tokens	€15.00 - €75.00	€0.00 (Fixed Hardware)
Cache Deduplication	Black-box / Provider Tier	62% Radix Reclaim (L1/L2)
EU AI Act Art. 12	Unverifiable Black Box	HMAC Cryptographic Proof
p99 Latency (TTFT)	350ms - 1,200ms	<85ms (Local L40S Cluster)

Ambient Service Mesh & Waypoint Ingress

Traditional enterprise gateways require intrusive SDK refactoring or heavy sidecar proxies that drain memory and complicate operations. Mastrion introduces sidecarless ambient waypoint proxies.

Operating at the Linux socket level via eBPF redirection, Mastrion transparently intercepts model traffic from developer workstations, internal microservices, and automated pipelines. L4 mutual TLS guarantees internal network segmentation, while L7 waypoint proxies execute deep prompt inspection and prefix cache routing.

<0.8ms	0 Code	100%
Mesh L4 Overhead	App Changes Required	eBPF Socket Capture



FinOps Cost Arbitration & 38-Day Breakeven

Stop paying premium frontier pricing for boilerplate tokens. Mastrion evaluates prompt structure, token count, task classification, and data sensitivity in real time to route traffic dynamically.

Routine tasks (code completion, test generation, schema translation, document summarization) run on internal GPU clusters at €0 incremental token cost. Only complex multi-step reasoning from sanctioned roles is dispatched to external cloud APIs, protected by strict cost circuit breakers.

38 Days
Appliance Payback

€222,000
Year 1 Net Savings (50 Devs)

82%
OpEx Cloud Reduction

routeTable.finops.mastrion.io · Routing Policy

```
// Declarative Cost Arbitration Route Table
routeTable:
  - match:
      task: [CodeCompletion, UnitTests, DocGen]
    action: RouteToLocalSovereignCluster
    backend: vllm://local-l40s-pool
    costImpact: "€0.00 / token · 100% Internal"

  - match:
      containsPII: true
    action: AirgapLockdown
    backend: local://finance-weights-32b
    security: "Zero Cloud Egress · BaFin Audit Safe"

  - match:
      task: ComplexMultiStepReasoning
      userRole: PrincipalArchitect
    action: RouteToSanctionedFrontierCloud
    budgetGate: "Max €50/day per user cap"
```

Model Memory Firewall: In-Line PII Redaction

Enterprise developers unintentionally paste confidential data into prompts: customer IBANs, medical identifiers, national identity numbers, and AWS API secret keys.

Mastrion's Model Memory Firewall executes sub-2.4ms deterministic token scrubbing directly inside the waypoint socket. Regulated entities are masked before GPU inference, and a tamper-evident cryptographic SHA-256 hash is emitted to an immutable audit sink complying with EU AI Act Article 12.

<2.4ms	100%	SHA-256
Scrub Latency	Deterministic Masking	Audit Receipts

In-Flight Redaction Inspection Log

```
// In-Line Socket Interception & Redaction Trace
rawPrompt: "Verify client account DE89370400440532013000 born
1984-11-12 with AWS key AKIAIOSFODNN7EXAMPLE"

[PII Scanner Active · Engine: rust-regex-simd]
  ✓ Matched [IBAN_DE]: DE89...3000 ⟶ [REDACTED:IBAN_01]
  ✓ Matched [DATE_OF_BIRTH]: 1984-11-12 ⟶ [REDACTED:DOB_01]
  ✓ Matched [AWS_SECRET_KEY]: AKIA... ⟶ [REDACTED:KEY_01]

sanitizedEgress: "Verify client account [REDACTED:IBAN_01] born
[REDACTED:DOB_01] with AWS key [REDACTED:KEY_01]"

auditReceipt:
  sha256Hash:
"e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855"
  complianceStandard: "EU-AI-Act-Art12 · Tamper-Evident"
  latency: "1.84ms"
```

Compute Garbage Collection: Prefix Caching

Development organizations repeatedly submit the same system prompts, codebase schemas, API documentation, and git diff contexts. Re-evaluating these repetitive tokens wastes expensive GPU memory and compute cycles.

Mastrion's Radix-tree prefix cache indexes prompt tokens across engineering teams. Identical token sequences are resolved from GPU KV cache memory at zero marginal compute cost, slashing Time-to-First-Token (TTFT) by up to 74% and reclaiming up to 62.4% of total cluster capacity.

62.4%	-74%	€0.00
Token Reclaim	p99 TTFT Latency	Reclaimed Token Cost

Radix-Tree KV Cache Execution Proof

```
// Radix Token Tree Deconstruction (2,840 Token Request)
Prompt Composition:
  [Chunk 1: System Policy & Security Prompts]  ⟶ 820 Tokens (CACHED)
  [Chunk 2: Codebase Ast & API Schema Docs]    ⟶ 952 Tokens (CACHED)
  [Chunk 3: Dynamic User Instruction]          ⟶ 1,068 Tokens (NEW)

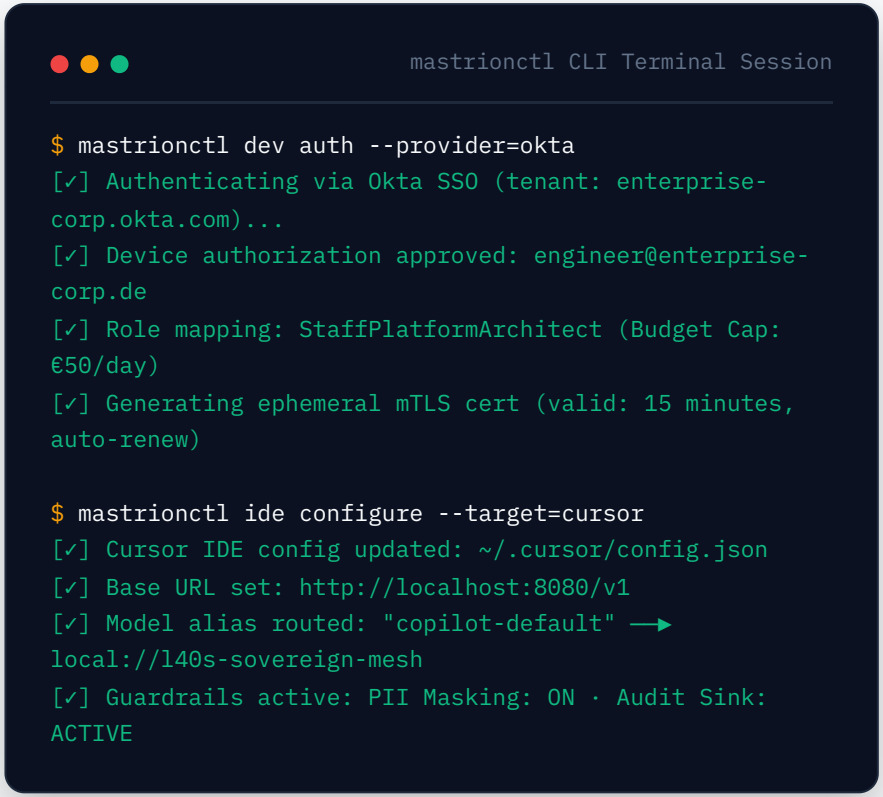
Prefix Cache Resolution:
  totalTokensSubmitted: 2,840
  radixPrefixHitTokens: 1,772 Tokens (62.4% RECLAIMED)
  freshTokensEvaluated: 1,068 Tokens
  ttftWithoutCache: 340ms
  ttftWithMastrionCache: 88ms (74% Faster)
  vramAllocationReclaimed: 3.8 GB GPU Memory
```

Enterprise Developer Experience (1-Click SSO)

Security guardrails fail when they slow down software development teams. Mastrion integrates directly into existing engineering workflows without friction or learning curves.

Engineers authenticate via corporate Okta, Microsoft Entra ID (Azure AD), or Ping Identity using `mastrionctl dev auth`. Standard IDEs (VS Code, Cursor, JetBrains) connect to the local waypoint endpoint using OpenAI-compatible APIs, instantly gaining compliant sovereign inference without hardcoded API keys.

1-Click	0 Keys	15 Min
SSO Enrollment	Stored on Laptops	Rolling Auth Tokens



Turnkey Air-Gapped Hardware Appliances

For defense, intelligence, core banking, and clinical healthcare institutions, public cloud hosting is legally prohibited. Mastrion delivers pre-configured turnkey 2U and 4U sovereign rackmount appliances.

Appliances arrive pre-loaded with immutable Mastrion OS, high-throughput vLLM runtime kernels, and encrypted NVMe weights. Hardware Security Modules (HSMs) provide tamper-evident physical key isolation, requiring zero outbound internet connectivity or telemetry phone-home.

100%
Air-Gapped Isolation

2U / 4U
Form Factors

BSI C5
Audit Compliant

Appliance Technical Specifications

Specification	2U Sovereign Edge	4U Enterprise Core
GPU Hardware	2x NVIDIA L40S (96GB VRAM)	4x NVIDIA H100 SXM5 (320GB)
Throughput	1,850 tok/s (Concurrency: 64)	8,400 tok/s (Concurrency: 256)
Active Weight Pool	7B to 32B Sovereign Weights	70B Dense / MoE 8x22B Weights
Security Hardware	TPM 2.0 + SED NVMe	FIPS 140-3 Level 3 PCIe HSM
Telemetry Requirement	Strictly ZERO Egress	Strictly ZERO Egress
Deployment Time	<4 Hours (Plug-and-Play)	<1 Day (Data Center Rack)

High-Availability Architecture & 99.99% SLA

AI inference failures immediately halt developer IDEs and customer-facing workflows. Mastrion is architected with platform engineering rigor, backed by contractual 99.99% availability service level agreements.

Multi-region mesh topologies, active GPU health checks, automatic memory defragmentation, and queue-shed circuit breakers prevent cascading failures during peak load. If an on-premises GPU node degrades, queries automatically failover to secondary nodes within 12 milliseconds.

99.99%	<12ms	24/7/365
Contractual SLA	Failover Reroute	German SRE On-Call

SRE High-Availability Policy & Circuit Breaker

```yaml
// Automated SRE Resilience & Circuit Breaker Spec
apiVersion: resilience.mastrion.io/v1alpha1
kind: SRECircuitBreakerPolicy
spec:
  vramThresholdTrigger: 92%
  queueShedding:
    action: DropNonInteractiveBackgroundEmbeddings
    preserveInteractivePriority: true
  nodeDegradationFailover:
    healthProbeInterval: 250ms
    heartbeatTimeout: 1000ms
    failoverTarget: cluster://secondary-az-pool
  slaGuarantees:
    monthlyUptimeCommitment: "99.99%"
    p99LatencyCeiling: "120ms"
    financialPenaltyCredit: "100% OpEx credit on breach"
```

5-Day Production PoC Sprint (€7,500 Full Rebate)

We do not sell software through abstract slide decks or lengthy procurement cycles. We deploy live in your real environment and prove concrete metrics in five business days.

For a flat fee of €7,500, Mastrion engineers deploy into your AWS/Azure VPC or on-premises cluster. In five days, we verify zero data egress, in-line PII redaction, and >60% token reclaim on real enterprise developer traffic. If you proceed with any annual production license, 100% of the €7,500 fee is credited toward your bill.

5 Days
To Production Proof

100%
Rebated on Contract

€7,500
All-Inclusive Fixed Sprint

5-Day PoC Milestone Schedule

Day	Milestone & Concrete Deliverable
Day 1	VPC / On-Prem Deployment & Waypoint Socket Setup
Day 2	Corporate SSO Integration & Pilot Dev Team Onboarding
Day 3	Live Model Memory Firewall & In-Line PII Redaction Audit
Day 4	Radix-Tree Prefix Cache Benchmarking (>60% Reclaim Verification)
Day 5	Executive FinOps Audit Report & 38-Day Payback Roadmap

✓ Guaranteed Outcome: If metrics are not proven on Day 5, sprint fee is fully refunded.

Built for Mandatory European AI Governance

European regulatory mandates (EU AI Act Article 12, BaFin BAIT circulars, NIS2, and GDPR) make unmonitored frontier API calls a high-liability corporate practice.

Article 12 mandates automated tamper-evident logging of all AI event transactions. Enterprises cannot fulfill this obligation through disparate application code. Mastrion's ambient waypoint proxies act as the mandatory control plane, enforcing policy at the transport layer and generating immutable compliance receipts.

Art. 12
EU AI Act Compliant

BaFin
Audit Ready

36 Mo.
CVE Backport Assurance

Regulatory Defensibility & Moat Architecture

```
// Compliance & Audit Architecture
standardsAssurance:
  - framework: "EU AI Act Article 12"
    mechanism: HMAC-SHA256 Cryptographic Ingress Ledger
    status: "Audit Safe · Tamper-Evident"

  - framework: "BaFin BAIT Cloud Circulars"
    mechanism: Strict Airgap Routing for Regulated Core
Data
    status: "Zero Unsanctioned Cloud Egress"

  - framework: "BSI C5 Assurance"
    mechanism: Local Hardware Root-of-Trust & HSM Key
Isolation
    status: "Government Grade Isolation"
```

Partner with Mastrion GmbH

Headquartered in Germany with engineering presence across Munich and Berlin, Mastrion partners with leading European financial institutions, defense prime contractors, and scaled SaaS enterprises.

Schedule an executive technical discovery call or book your 5-Day Production PoC Sprint. Our engineering leadership is available for technical deep-dives, architectural reviews, and press inquiries.

30 Min
Discovery Call

Direct
Founder Access

Munich / Berlin
German Headquarters

Executive Contact & Booking Information

```
// Executive Contacts & Booking
corporateEntity: "Mastrion GmbH"
webPortal: "https://mastrion.eu"
executiveContact: "mofesola@mastrion.eu"
partnerInquiries: "contact@mastrion.eu"

// Online Meeting Booking (Google Calendar)
meetingSchedule: "https://mastrion.eu/book"
schedulingType: 30-Min Executive Technical Briefing
format: Google Meet / In-Person (Munich / Berlin)

// Press Kit & Security Fingerprints
pressKitArchive: "https://mastrion.eu/docs/press-kit.zip"
gpgSecurityKey: "F649 B4A8 2883 9161 02CD 78A2 4910"
```

 All Rights Reserved

Take Sovereign Control of Your Model Runtime

Generative AI is not an opaque third-party black box; it is an active enterprise runtime requiring deterministic boundaries, in-line compliance receipts, and intelligent compute arbitration. Mastrion provides European enterprises with an immutable foundation that proves zero cloud egress while reclaiming over 60% of GPU compute capacity.

01 // ENGAGE

30-Min Discovery Briefing

Direct architectural consultation on Google Meet or in-person in Munich/Berlin.

Schedule at mastrion.eu/book

02 // VALIDATE

5-Day PoC Sprint (€7,500)

Deployed live in your VPC or cluster. 100% credited toward production licensing upon sign-off.

Inquire at contact@mastrion.eu

03 // DEPLOY

Turnkey 2U/4U Appliances

Pre-loaded sovereign hardware appliances for air-gapped core banking and defense installations.

Explore at mastrion.eu/pricing