

Guido Maria Serra

Principal Infrastructure Engineer | AI/ML Systems & GPU Platforms

Berlin, Deutschland

+49 173 7349 795 | guido@whitehat.berlin | linkedin.com/in/zeph1ro | github.com/zeph

Profil

Principal Infrastructure Engineer mit ~30 Jahren Erfahrung im Aufbau und Betrieb großskaliger verteilter Systeme — von früher Linux-Infrastruktur bis zu modernen GPU-basierten Kubernetes-Plattformen für AI/ML-Workloads. Spezialisiert auf Stabilisierung kritischer Infrastruktur, Kostenoptimierung (z.B. \$300K/Jahr bei OLX) und Coaching von SRE/DevOps-Teams.

Highlights: Infrastruktur mit 55k Hosts (12 RZ); Teams durch produktionskritische Vorfälle geführt.

Ausgewählte Projekte

PlayStation Now — Globaler Infrastrukturbetrieb

Betrieb einer Multi-Datacenter-Plattform (55k Hosts, 12 RZ); Aufbau von Ceph Storage, PKI-basiertem SSH. Entwicklung einer eigenen Metrics-DSL (TextX) mit Grafana-Integration. Gentoo Maintainer für Ganeti/KVM.
Ergebnis: Plattform unter globalem Skalierungsdruck stabilisiert; PlayStation Now Launch ermöglicht.

HelloFresh — Security & Infrastruktur-Skalierung

Security-Team 1→5 FTEs in 2 Wochen. Vuln Scanning, Elasticsearch SIEM, Incident Response.
Ergebnis: SIEM operativ; AWS/Cloudflare Edge-Monitoring etabliert.

OLX Group — Infrastrukturkosten & Zuverlässigkeit

APM-Strategie über 4 Hubs geleitet. CDN/SSL/DNS-Anbietersauswahl. NewRelic Vendor Manager (\$2M Budget). Load Testing und Performance-Validierung (eigene Tools).
Ergebnis: \$300K/Jahr Einsparungen durch FinOps; verbesserte Plattformzuverlässigkeit.

Berufserfahrung

Security & DevOps Consultant | DFND Security / Forward Earth / 3T Labs 2025
Remote — Mehrere Kunden

- Fastly (Signal Sciences) WAF und Edge-Security-Deployments für EMEA-Kunden.
- ISO 27001 Zertifizierungsprojekt; Vanta Auditor Training und Compliance-Automatisierung.

Software Engineer | FinTech (stealth) 2025
Tokenisierte Finanzsysteme (Python, TypeScript, Microservices).

- Backend-Komponenten für tokenisierte Finanzsysteme entwickelt.
- AI-gestützte Coding-Workflows evaluiert und angewendet zur Beschleunigung der Iterationszyklen.
- Load Testing und Reverse Engineering von Tools/Libraries zur Validierung des Systemverhaltens.

Head of SRE | Octostar 2024–2025
GPU Kubernetes-Plattform für ML-Workloads; ClickHouse/MinIO Datenplattform.

- GPU-basierte Kubernetes-Cluster für ML Training/Inference konzipiert.
- Kosten und Performance für ML-Workloads optimiert (gemischte On-Prem/Cloud GPU-Nodes).
- Workload-Portabilität zwischen lokalem Kubernetes (kind/minikube) und Cloud (Hetzner) validiert.

DevOps Lead (Advisory) | Contensi 2024–heute
CDKTF Infrastruktur-Automatisierung; Mentoring von Junior Engineers.

Staff SRE (Fractional) | Prima Assicurazioni 2022–2023
12-köpfiges SRE-Team aus operativem Stillstand geocoacht; AWS Multi-Region-Rearchitektur (London, Barcelona, Mailand).

Cloud Architect | ClovrLabs 2021–2023
 5-köpfiges Security Ops Team aufgebaut; Kubernetes/GitOps-Plattform (ArgoCD, Traefik, Datadog).

Site Reliability & Security Engineer | Max Planck Institute 2022–2024
 Hybrid OpenStack/Kubernetes; 1.5PB Storage; föderiertes IAM (LDAP/SAML/OAuth2).

Security / DevOps Engineer | HelloFresh 2020–2021

DevOps Engineer | Mobivia / ATU 2018–2022
 4-jährige Modernisierung: Cloudflare/Traefik-Migrationen, Keycloak SSO, ArgoCD GitOps.

DevOps Team Lead | Urban Mobility Int. (WeShare/VW) 2019
 1-wöchiges Pre-Launch-Engagement; GCP Functions und Kubernetes.

Start-up CTO | SiWeGO (Trento, Italien) 2017–2019
 Lösungsdesign für niedrige OPEX; algorithmisches Design von Graph-Überlappungen; Lieferantenmanagement.

Senior Performance Engineer | OLX Group 2016–2018

Site Reliability Engineer | Sony PlayStation (Gaikai → PlayStation Now) 2013–2016

Frühere Karriere (2008–2013)

Rocket Internet — Chief Troubleshooter & Senior Data Architect (2013) | Berlin
txtr (3M) — Backend QA Manager (2011–2012) | Berlin
ProfitBricks (→ IONOS) — Senior System Engineer (2011) | Berlin
Nokia Maps (Gate5 → HERE) — Senior System Engineer (2010–2011) | Berlin
Vodafone Group — QA Architect (2008–2010) | Düsseldorf

Gründungsphase (1996–2008)

Solution Architect für KMUs in Italien. LDAP/VoIP/Firewall-Lösungen konzipiert; Patches für GOsa², PHP Seagull, GNU TLA veröffentlicht. Kunden waren ISPs, Banken (BPU Banca) und Telcos — darunter die digitale Bibliothek Stendhal für Biblioteca Sormani (noch online: digitami.it/stendhal). Praktikant bei Alessandro Rubini für Embedded Linux R&D. Lab Assistant am Politecnico di Milano (C/Linux Socket-Programmierung, 2002). Aktiv bei FOSDEM seit 2007.

Technische Kernkompetenzen

Verteilte Systeme & Infrastruktur	Large-scale Linux, Ceph, Distributed Storage, Fleet Management
Cloud & Platform Engineering	AWS, GCP, Kubernetes, Terraform, GitOps
Observability & Reliability	Prometheus, Grafana, ELK, OpenTelemetry, SLO/SLI
Security & Compliance	WAF, IAM/SSO, ISO 27001, SIEM
AI/ML Workloads	GPU-Umgebungen, Data Pipelines, Local/Cloud-Portabilität
Programmierung	Go, TypeScript, PHP, SQL; hands-on, Stärken in Debugging und Systemanalyse
Low-Level Systems	C, Embedded Linux, Kernel-Level Performance

Ausbildung & Zertifikate

Bachelor of Engineering in Informatik (letztes Jahr)
 — Politecnico di Milano | 2000–2008, wiederaufgenommen 2025
Schulungen: Red Hat Ceph Storage (CEPH125), NYC 2016; ISO 27001 / Vanta Auditor, 2025

Zusätzliches

Sprachen: Italienisch (Muttersprache), Englisch (Fließend), Deutsch (Gut)
Security: DEFCON CTF Finalist; iCTF World Champion; SOC GOON (2025)
Open Source: FSFE Fellow; Mitgründer POuL (Politecnico Open *nix Labs); wicd Maintainer
Führerscheine & Interessen: Klasse A, B; Hochseesegeln (Regatta, Seefunk); Skifahren; Wandern