

Michele Cipriano

AI Solution Architect · Cloud & Generative AI

Ingegnere informatico specializzato in AI generativa, sistemi multi-agente e architetture cloud native. Porto soluzioni GenAI dalla proof of concept alla produzione tenendo insieme costi, compliance e scalabilità. Speaker all'AWS Summit di Milano; contributi a case study e white paper Intel.

ciao@michelecipriano.it

michelecipriano.it

linkedin.com/in/michele-cipriano

Airola (BN), Italia

ESPERIENZA PROFESSIONALE

AI Solution Architect · AIOps Specialist · DevOps Engineer

08/2022 — oggi

Storm Reply, Roma — settore Energy & Utilities

- Progettazione di un **assistente virtuale multi-agente** (Strands, AWS AgentCore, Bedrock) che genera architetture tecniche da richieste in linguaggio naturale, standardizzando le soluzioni e riducendo i tempi di design.
- Sviluppo di un **troubleshooter GenAI basato su RAG** integrato nelle pipeline CI/CD Atlassian Bamboo: analizza i fallimenti di build e propone la remediation, riducendo l'MTTR.
- Migrazione dell'inferenza AI da on-premise (Kubernetes/GPU) a un'architettura **serverless su AWS Bedrock** (Lambda, Step Functions), con riduzione dei costi infrastrutturali e maggiore scalabilità.
- Realizzazione di un sistema **AIOps multi-agente** (CrewAI) per l'analisi dei log di piattaforma, il rilevamento delle anomalie e i suggerimenti di remediation.
- Automazione dei processi DevOps e provisioning AWS (CloudFormation, Lambda, EKS) per più team di sviluppo.

Technical Supervisor — tesi di ricerca in AI generativa

01/2025 — 03/2026

Storm Reply, Roma

- Supervisione di due tesi magistrali su sistemi multi-agente e GenAI applicata agli AIOps e ai contesti cloud native.
- Guida dell'evoluzione di un sistema AIOps multi-agente (CrewAI) per l'analisi e la diagnosi automatica delle anomalie nei log.
- Studio sull'**interoperabilità fra protocolli di agenti** (MCP, A2A, ACP) per definire un'architettura multi-agente scalabile, con valutazione del deployment su AWS AgentCore in sinergia con il team AWS.

Cloud & Generative AI Engineer

05/2024 — 12/2024

Storm Reply — settore Semiconduttori & Elettronica

- Sviluppo di un **chatbot RAG su AWS** (Bedrock, Lambda, Step Functions, DynamoDB) a supporto degli sviluppatori sulla documentazione tecnica interna.
- Implementazione di un modulo di reporting che analizza i pattern delle query per identificare vuoti di documentazione e colli di bottiglia nel processo di design dei chip.
- Co-progettazione di un **IDE web** (Theia) su AWS (ECS/EKS, Cognito), con microservizi e pipeline GitHub Actions per abilitare lo sviluppo remoto su hardware fisico.

AI DevOps Optimization Engineer

01/2024 — 04/2024

Storm Reply — settore Digital Services & Digital Identity

- Progettazione di una soluzione GenAI (RAG) per il troubleshooting DevOps integrata nelle pipeline CI/CD (Jenkins, Tekton), con riduzione dell'MTTR.
- Deploy e ottimizzazione di **Llama 3 on-premise** in pod Kubernetes su CPU Intel, garantendo la compliance sui dati ed eliminando i costi di inferenza cloud.
- Risultati economici tali da far includere il progetto nel **case study ufficiale Intel** sull'ottimizzazione di soluzioni AI su istanze CPU.

- PoC tecnica di **benchmarking dell'inferenza LLM su CPU Intel Xeon con AMX**, con validazione del rapporto costo/prestazioni rispetto alle soluzioni GPU tradizionali.
- Contributo di analisi e dati al **white paper ufficiale Intel** sul rapporto prezzo/prestazioni fra CPU e GPU.
- PoC comparativa AWS Bedrock contro Kubernetes on-premise su prestazioni, costi e sicurezza, a supporto della strategia di adozione AI del cliente.

Cloud & Performance Engineer

03/2022 — 07/2022

NTT Data, Napoli — settore Telecomunicazioni

- Ottimizzazione delle prestazioni di database PostgreSQL in ambienti multi-cloud (GCP) e on-premise, con riduzione della latenza delle query critiche.
- Sviluppo di strumenti di monitoraggio custom in **Go** per la validazione automatica dei sistemi.
- Progettazione di architetture Kubernetes per il deployment scalabile e resiliente di applicazioni containerizzate.

Systems Engineer

01/2021 — 02/2022

NTT Data, Napoli — settore Telecomunicazioni

- Gestione di un'infrastruttura di autenticazione enterprise (LDAP, Kerberos) su sistemi CentOS/RedHat mission-critical.
- Automazione di attività di amministrazione, monitoraggio e hardening di sicurezza tramite scripting Bash.

Web & Software Developer (freelance)

07/2020 — 04/2022

Settori Food & Retail, Media & Digital Events

- Progettazione e sviluppo di una piattaforma e-commerce completa (WordPress, PrestaShop), seguita end-to-end dai requisiti al rilascio.
- Sviluppo di un sistema di **control room virtuale** per la gestione di streaming multi-camera, con chat di regia in tempo reale e storage cloud condiviso.
- Realizzazione di una piattaforma di sale d'attesa virtuali con pannello di controllo in tempo reale per gli host.

FORMAZIONE

2023	Laurea Magistrale in Ingegneria Informatica 110 e lode Università degli Studi del Sannio · EQF 7 Tesi: <i>A Troubleshooter for DevOps based on Retrieval Augmented Generation and Large Language Models</i>	2017	iOS Foundation Program Università degli Studi del Sannio · Apple
		—	Lingue Italiano (madrelingua) · Inglese
2019	Laurea Triennale in Ingegneria Informatica Università degli Studi del Sannio		

CERTIFICAZIONI

2024	AWS Certified Solutions Architect – Associate Amazon Web Services	2022	CKA — Certified Kubernetes Administrator The Linux Foundation
2024	CKAD — Certified Kubernetes Application Developer The Linux Foundation		

COMPETENZE TECNICHE

AI generativa	RAG, agenti autonomi, sistemi multi-agente (CrewAI, Strands, AutoGen), AWS AgentCore, protocolli MCP / A2A / ACP, prompt engineering, benchmarking e valutazione dei modelli, database vettoriali
---------------	---

AWS	Bedrock, Lambda, Step Functions, DynamoDB, EC2, S3, ECS/EKS, Cognito, CloudFormation, CloudFront
Cloud & DevOps	Kubernetes (CKA, CKAD), Docker, GitHub Actions, Atlassian Bamboo, Jenkins, Tekton, GitLab CI, Terraform, Git, AIOps
Linguaggi & dati	Python, Go, Java, JavaScript, PHP, C++, Bash, PostgreSQL, MySQL, DynamoDB
Sistemi	Linux (CentOS, RedHat), LDAP, Kerberos, Intel AMX, ottimizzazione dell'inferenza su CPU
Trasversali	Public speaking, scrittura tecnica, knowledge management, mentoring e supervisione di tesisti, lavoro cross-funzionale, metodo agile

SPEAKING E PUBBLICAZIONI

05/2024	AWS Summit Milano — <i>Generative AI e CPU</i> · Speaker Storm Reply con Intel: piattaforma GenAI per DevOps su CPU, con confronto CPU/GPU su costi e prestazioni (Intel AMX)
06/2024	Università del Sannio — <i>Applicazioni avanzate e consolidate di RAG</i> · Speaker Analisi, benchmark e studi consolidati su RAG e loro applicazione in contesto industriale
12/2023	Università del Sannio — <i>Applicazioni innovative di RAG in contesto aziendale</i> · Speaker Dialogo fra ricerca accademica e industria sulle prime applicazioni di RAG
09/2023	Storm Reply Event, Milano Marittima — <i>RAG applicato al DevOps</i> · Speaker Studi preliminari sull'applicazione di Retrieval Augmented Generation al DevOps
2023— 24	White paper e case study Intel · Contributo Analisi e dati di benchmark sul rapporto prezzo/prestazioni dell'inferenza LLM su CPU

Autorizzo il trattamento dei miei dati personali contenuti nel presente curriculum ai sensi del Regolamento (UE) 2016/679 (GDPR) e del D.Lgs. 196/2003 come modificato dal D.Lgs. 101/2018. I progetti sono descritti in forma anonimizzata: nessun nome di cliente e nessuna informazione riservata è riportata in questo documento.