ConRes
IT SOLUTIONS

# Accelerating infrastructure operations with AI-powered retrieval

A leading optical networking technology provider faced challenges retrieving infrastructure component information quickly and accurately across its data center environment. With assets spanning switches, routers, firewalls, servers and cables, the process was time-consuming and complex. ConRes partnered with the provider to design and implement an AI-powered natural language processing (NLP) solution, enabling real-time, precise access to asset information that streamlined operations and improved visibility.

## Challenge

With a strong technical infrastructure in place, the client recognized an opportunity to expand into AI, but needed support in defining practical use cases. ConRes provided targeted education, guided experimentation and a strategic framework to help uncover operational AI opportunities that could enhance infrastructure management.

## Approach

ConRes initiated a discovery phase to introduce stakeholders to generative AI and large language models (LLM), providing hands-on experiences through demos and chat-based interfaces. As part of this process, we evaluated multiple technology solutions within our portfolio and determined that Red Hat OpenShift was the most suitable platform for the client's needs. To ensure scalability and security, we then implemented a production-grade AI platform on OpenShift, tailored for enterprise deployment.

## Key use case

ConRes implemented AI-powered natural language processing (NLP) for infrastructure component information retrieval, enabling real-time, accurate access to data center asset information about switches, routers, firewalls, servers and cables. This approach reduced coding efforts, minimized manual errors, accelerated troubleshooting and supported predictive maintenance, ultimately improving operational efficiency and data accuracy.

# Solution—Red Hat OpenShift AI Infrastructure

Built on Red Hat OpenShift Container Platform, the environment integrates components to support secure, scalable AI workloads.

### Model serving and orchestration

- KServe with vLLM engine for serverless inference
- NVIDIA GPU Operator for multi-model GPU support

### Security and networking

- Istio Service Mesh for secure traffic control
- Authorino for API security
- Red Hat Single Sign-On (Keycloak) for identity management

### Data management and storage

- OpenShift Data Foundation for S3-compatible, high-availability storage
- Open Data Hub (ODH) for model downloads and persistent data

### User experience

- Containerized UIs (AnythingLLM, Open-WebUI) with Model Context Protocol (MCP)
- MCP Proxy Server integrating with OpenAI APIs and NetBox for real-time insights

### Model deployment

- IBM Granite 3.3 (OpenShift Lightspeed) for localized inference
- Testing with LLaMa 3, LLaMa 4 and Mistral models

## Outcome

By integrating its infrastructure with ConRes's AI Professional Services Solutions on Red Hat OpenShift, the client substantially enhanced operational efficiency. Implementing NLP-driven retrieval reduced manual coding efforts and manual errors, accelerating troubleshooting and enabling predictive maintenance. This AI-driven approach laid the foundation for future automation, improved data accuracy and supported their digital transformation in next-generation optical interconnect solutions.