Green-Aware Dynamic Workload Distribution and Edge Resource Control for AI Inference

Toyota, F5 Networks

Goal

Maximize renewable energy utilization for distributed AI inference

- Renewable availability varies by edge sites and time

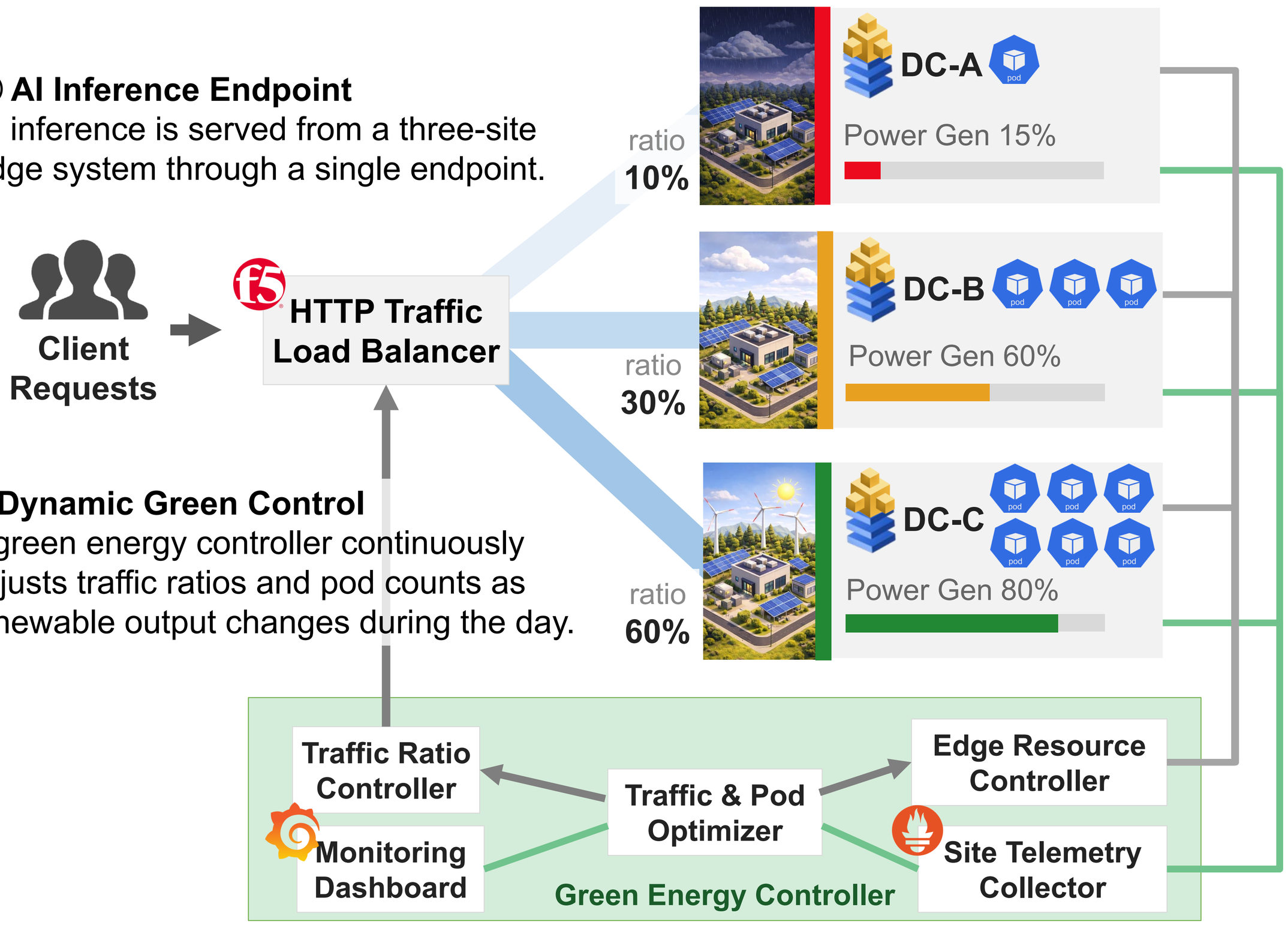
→ **Dynamically adjusting both traffic routing ratios and GPU pod scaling across geo-distributed edge sites**

Demo Overview

① Heterogeneous Renewable Profiles
Each site has a different renewable energy generation profile.

② AI Inference Endpoint
AI inference is served from a three-site edge system through a single endpoint.

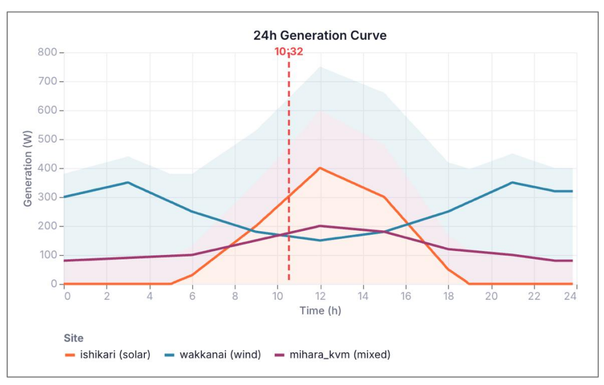
③Dynamic Green Control
A green energy controller continuously adjusts traffic ratios and pod counts as renewable output changes during the day.



Enabling Technology

Site Telemetry

Predefined generation scenarios



KEPLER

Power telemetry collection

Green Energy Controller

Traffic & Pod Optimizer
Computes the traffic ratios and pod allocation across sites to minimize brown energy use.

Traffic Ratio Controller
Uses the F5 API to update LB ratios.

Edge Resource Controller
Uses the K8s API to adjust pods.

Network

F5XC HTTP Load Balancer

F5XC APP Stack
- Managed K8s
- Network Mesh

AI Inference

llama.cpp

Qwen 3.5 4B