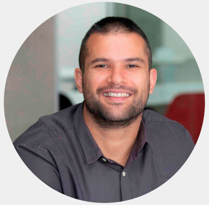
# Panel: Performance Modeling For The Computing Continuum



spec RG Cloud

# Panelists

**Matthijs Jansen**
PostDoc
VU Amsterdam

**Padma Apparao**
Perform. Architect
Intel

**José Castillo Lema**
Software Engineer
Red Hat

**Tommaso Cucinotta**
Assoc. Prof.
Scuola Superiore
Sant'Anna

# Model, Compare, and Predict in the Cloud Continuum

Use performance modeling to:

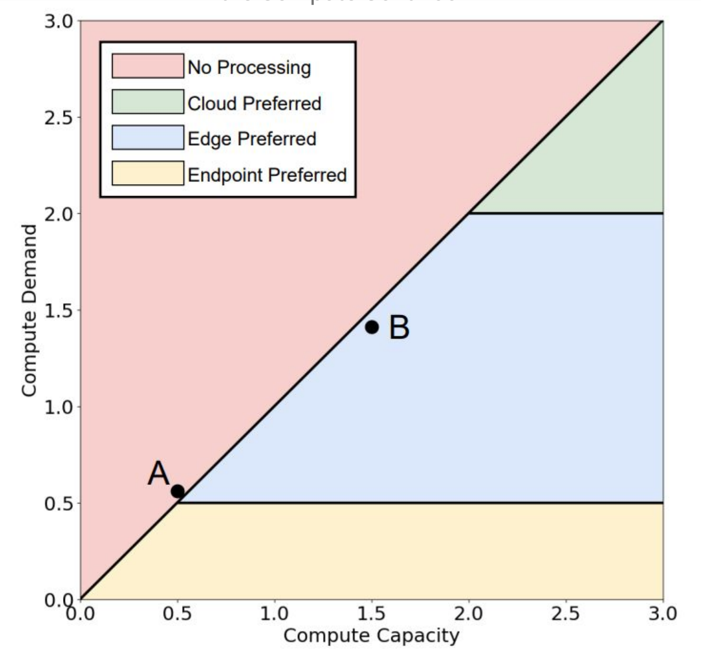
1. Make task offloading decisions
2. Tune system configurations
3. Predict application performance

Modeling requires real-world data

Challenge: Lack of traces and performance data

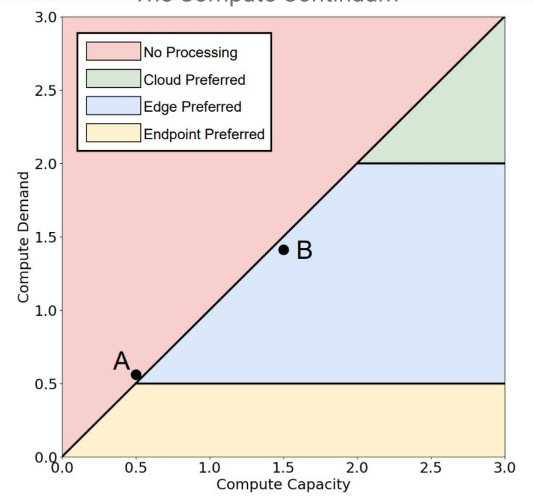
- Limited data for individual systems (cloud)
- No public data across the continuum

The SPEC-RG Reference Architecture for the Compute Continuum



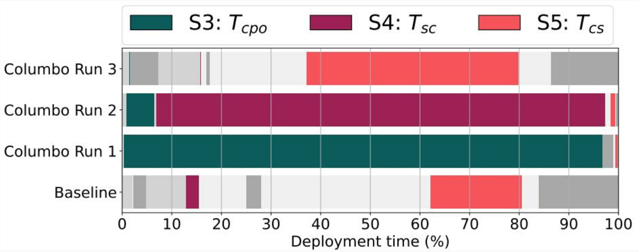
# Model, Compare, and Predict in the Cloud Continuum

The SPEC-RG Reference Architecture for The Compute Continuum



Compare task offloading scenarios

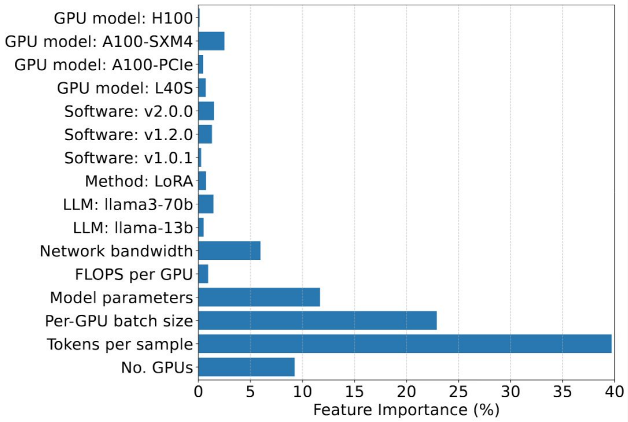
Columbo: A Reasoning Framework for Kubernetes' Configuration Space



$$\begin{aligned} T = \quad & (T_{S1} + T_{S2}) \times \lceil J/U_c \rceil \\ & + (T_{S3} + T_{S4}) \times \lceil (J \times P)/U_c \rceil \\ & + (T_{S5} + T_{S6}) \times \lceil (J \times P)/(N \times U_w) \rceil \end{aligned}$$

Predict best-case system performance

Optimizing ML Job Scheduling with Configuration Knowledge



Model performance of ML configurations

# Performance Modeling in the Compute Continuum
# Predicting the Unpredictable

About me

- Strong advocate for consistent performance methodologies
- Designs and architects RAG-based solutions, leading the evaluation of OPEA-built solutions and driving compliance standards

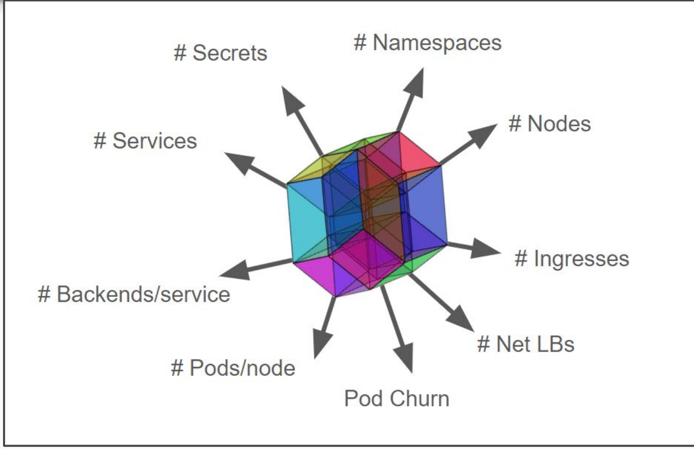
Performance is a dynamic equilibrium of compute, memory, and data movement tradeoffs

Across cloud, edge, and device, no universal model fits precision demands adaptability and real-time resilience

# Performance Modeling for the Computing Continuum

Red Hat engages in performance modeling across the computing continuum through:

- Engineering practices
- Open-source tooling
- Collaborative research initiatives



Active Red Hat **research** initiatives

- **CODECO**: A smart, and cross-layer orchestration between the decentralised data flow, computation, and networking services, to address Edge-Cloud challenges
- **AC3**: Employs AI/ML algorithms to predict resource usage and availability in cloud-edge infrastructures.

# Modeling Performance in the Cloud Continuum

Goal

- Predictable performance across compute, storage, and networking
- Critical for end-to-end application reliability

Key Challenges

- Resource Heterogeneity (Cloud vs Edge)
- Variable resource allocation flexibility (e.g., multi-tenancy)

Research Need

- Energy-aware, predictable interfaces
- Fine-grained control over latency-impacting features