VECA: Reliable and Confidential Resource Clustering for Volunteer Edge-Cloud Computing[1]

Prasad Calyam*, Saptarshi Debroy**, University of Missouri-Columbia*, City University of New York**

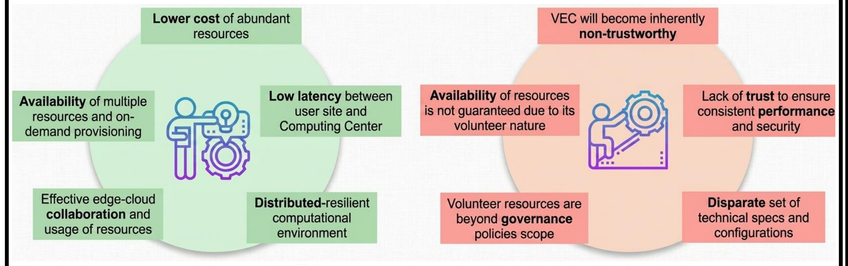
Problem & Motivation

Volunteer Edge-Cloud (VEC) computing paradigm can offer decentralized compute power for ML/DL scientific workflows, but faces three core challenges:

- **Volatility:** VEC nodes have unpredictable, intermittent availability, leading to workflow execution failures
- **Heterogeneity:** Diverse capacity profiles (CPU, RAM, storage) complicate resource matching
- **Privacy risk:** Model and data must remain confidential from VEC resource providers

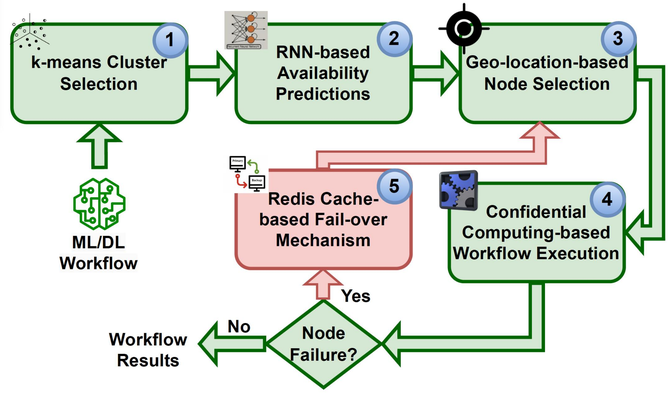
Existing solutions (VECFlex[2], VELA[3], CLARA[4]) suffer from inefficient node search, random cluster selection, and lack mechanisms for confidentiality and intermittency handling

Need: a reliable + confidential resource orchestration framework for ML/DL scientific workflows on VEC nodes



Mechanisms are needed to **rank** VEC resources based on their historical **behavior** in terms of resource availability, usage policies, and performance consistency, and that ranking can then be used as a **trustworthy** metric for resource provisioning and placement.

Approach

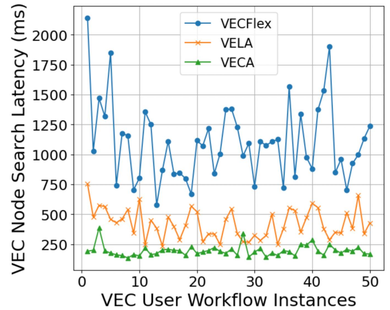


① **k-means Cluster Selection** Group VEC nodes by CPU, RAM, and storage similarity. Match incoming workflow to the nearest capacity-aligned cluster.
② **RNN-Based Availability Prediction** Forecast each node's online probability from historical time-series patterns (weekday, hour, node ID).
③ **Geo-Location-Based Node Selection** Filter for nodes with predicted availability ≥ 0.8; select the one closest to the user.
④ **Confidential Computing Execution** Run workflow inside an AWS Nitro Enclave (TEE) with cryptographic attestation, isolating model and data from the resource provider.
⑤ **Redis Cache–Based Fail-over** On failure, retrieve cached node ordering and re-dispatch instantly, i.e., no re-clustering, no re-inference.

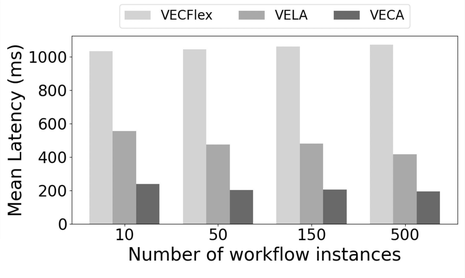
Evaluation Results

- **Testbed**: OpenFaaS + MicroK8s + Docker on AWS; 50 VEC nodes across 4 clusters
- **Workflows**: G2P-Deep (bioinformatics), PAS-ML (health informatics)
- **VEC Node Search Latency**: ~2× reduction vs. VELA (next best); large gap vs. VECFlex
- **Scalability**: VECA maintains advantage across 10, 50, 150, 500 workflow instances
- **Productivity Rate** (post-failure recovery efficiency):
  - VECA: **86.9%**
  - VELA: 66.7%
  - VECFlex: 65.7%

- VECA delivers >20% improvement in productivity rate following execution failures



Results on VEC node search latency across 50 workflow instances.



Performance of the different approaches over a varying number of workflow instances

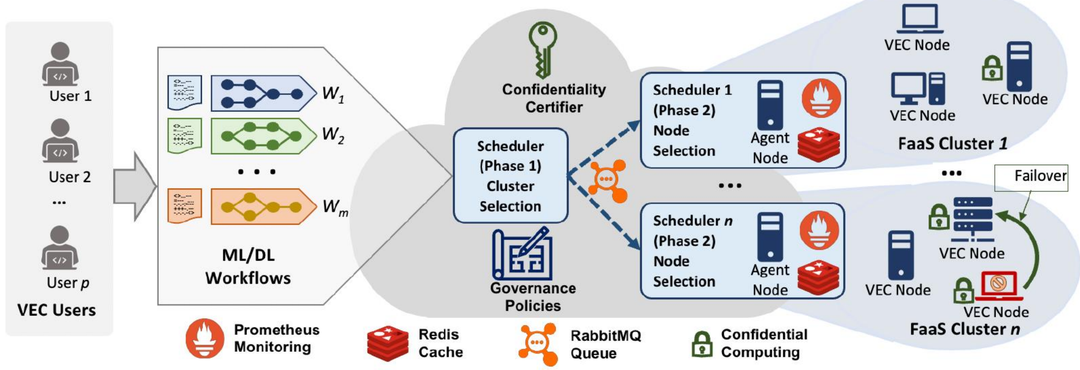
Solution & Novelty

VECA — a three-fold framework:

1. **Capacity-Based k-means Clustering**
   1. Groups VEC nodes by CPU, RAM, and storage similarity
   2. Optimal cluster count determined via Elbow method (k = 4 for 50 nodes)
   3. Re-clustering triggered when node pool grows by 10%

2. **Two-Phase Distributed Scheduling**
   1. *Phase 1 (Cloud Hub)*: Cluster selection matched to workflow capacity needs
   2. *Phase 2 (Cluster Agent)*: RNN-based time-series forecasting predicts node availability; geo-location filtering selects nearest node with availability ≥ 0.8
   3. Redis-cache-based fail-over enables rapid recovery without re-running RNN inference
   4. RabbitMQ for async inter-scheduler communication

3. **Confidential Computing Integration**
   1. AWS Nitro Enclaves provide TEE-based execution
   2. Four-step lifecycle: build, run, validate (cryptographic attestation), terminate
   3. Ensures model/data privacy from VEC resource providers



The VECA solution architecture illustrates users submitting ML/DL-based workflows to a Cloud Hub. Here, volunteer resources are clustered using the k-means algorithm and secured through a confidential computing framework. Two-phase distributed scheduling mechanism selects the most suitable cluster and the optimal VEC node within the selected cluster to execute the submitted workflow and meet user performance and security requirements

Salient Findings

- Intelligent capacity-aware clustering beats both exhaustive sampling (VECFlex) and random cluster selection (VELA) i.e., clustering granularity matters more than search exhaustiveness
- RNN-based time-series forecasting is effective for predicting binary node availability from temporal patterns (weekday/hour/node ID)
- Redis caching of pre-computed node orderings eliminates redundant RNN inference during fail-over, drastically reducing recovery time
- Latency convergence between VECA and VELA occurs only when the pool of free nodes shrinks thus confirming clustering benefits are realized when sufficient resources are available
- First work to apply time-series forecasting for VEC node availability prediction

Future Work

- Integrate federated machine learning to construct cluster capacities while also exploring richer node attributes (network bandwidth, energy availability) for clustering.
- Extend VECA to workflows with unique performance/privacy/security needs (e.g., medical imaging)
- Explore open-source confidential computing paradigm for workflow security on edge devices
- Evaluate VECA on real-world VEC node availability traces beyond synthetic datasets

References

[1] Yeddulapalli, Hemanth Sai, Mauro Lemus Alarcon, Upasana Roy, Roshan Lal Neupane, Durbek Gafurov, Motahare Mounesan, Saptarshi Debroy, and Prasad Calyam. "Veca: Reliable and confidential resource clustering for volunteer edge-cloud computing." In *2024 IEEE International Conference on Cloud Engineering (IC2E)*, pp. 152-159. IEEE, 2024.
[2] M. L. Alarcon, M. Nguyen, A. Pandey, S. Debroyand, and P. Calyam, "Vecflex: Reconfigurability and scalability for trustworthy volunteer edge-cloud supporting data-intensive scientific computing," in 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC), 2022, pp. 151–156.
[3] T. Pusztai, S. Nastic, P. Raith, S. Dustdar, D. Vij, and Y. Xiong, "Vela: A 3-phase distributed scheduler for the edge-cloud continuum," in 2023 IEEE International Conference on Cloud Engineering (IC2E). Los Alamitos, CA, USA: IEEE Computer Society, sep 2023, pp. 161–172.
[4] S. Gonzalo, J. M. Marqu`es, A. Garc´ıa-Villoria, J. Panadero, and L. Calvet, "Clara: A novel clustering-based resource allocation mechanism for exploiting low-availability complementarities of voluntarily contributed nodes," Future Generation Computer Systems, vol. 128, pp. 248–264, 2022.

This material is based upon work supported by the National Science Foundation under Award No. OAC- 2232889. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Electrical Engineering & Computer Science
University of Missouri