Latency, Energy and Carbon Aware Collaborative Resource Allocation with Consolidation and QoS Degradation Strategies in Edge Computing

Authors
Wedan Emmanuel GNIBGA
Anne BLAVETTE
Anne-Cécile ORGERIE

GreenDays - Toulouse - 2024

Introduction

- Edge computing goal : reduce latency
- 100 ms of latency in Amazon AWS service response resulted in a 1% decline in sales [1]
- Video traffic represented nearly four-fifths of global mobile data traffic in 2022[2]
- Video streaming as an important share of energy consumption and carbon impact of Edge computing

Our goals

- Edge infrastructure powered by renewable energy
- Increase the collective self consumption

[1] Zakarya, M. (2017). Energy and performance aware resource management in heterogeneous cloud datacenters. University of Surrey (UK)
[2] C. M. Vni, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022, Cisco, San Jose, CA, USA, Feb. 2019.

Introduction

- Definition:

  Collective self consumption: total energy consumed by the distributed data centers based from on-site generated energy

- Two ways to improve the collective self consumption in Edge computing

  1. Use the geographical distribution of PVs and migration to consume more on-site energy overall

  Usual practice: Load balancing with VMs migration

  Limitation: Migration is costly and energy consuming

  Our goal: Exchange load between sites at the resource allocation phase

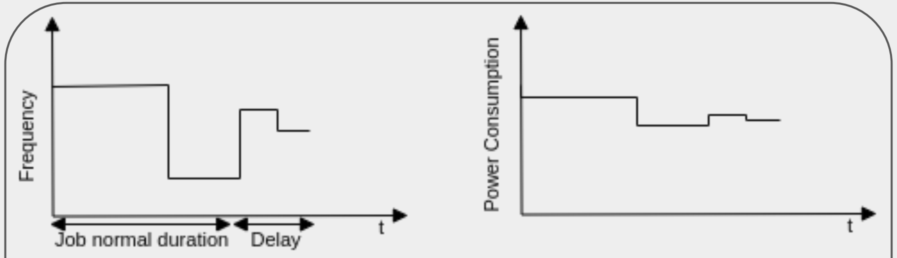
Introduction

❑ Two ways to improve the collective self consumption in Edge computing

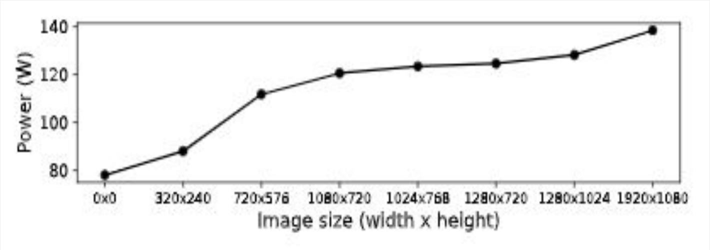
2. Reduce the power consumption of the Edge-Data centers (Edge-DCs)

Usual practice: Hardware solutions (Power capping/DVFS)

Limitation: Affects application runtime ⇒ risk of QoS violation



DVFS representation



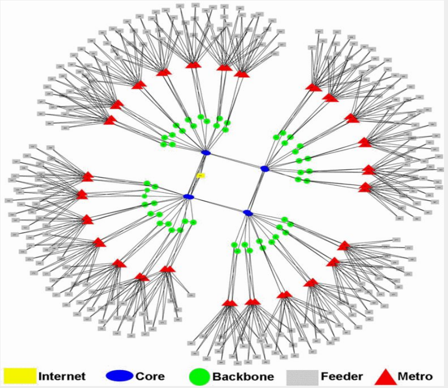
Observation : Nova (Grid'5000) node power consumption for several streaming resolution

Our goal:

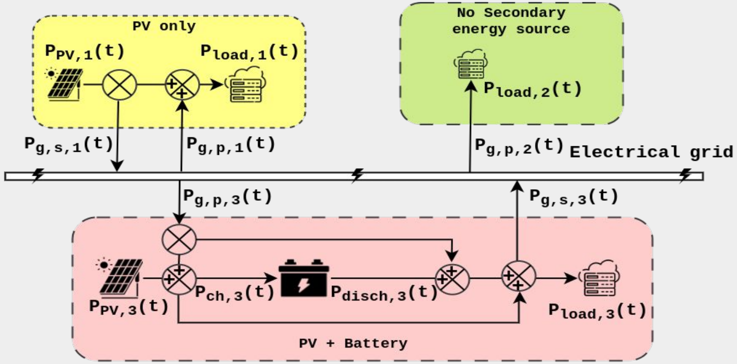
❑ Dynamically adjust the images resolution (application-level power reduction)

❑ Advantage: No impact on the duration.

Infrastructure design



Data centers network[3]: Each leaf contains an Edge Data center (Edge-DC)



Electrical infrastructure

- ❑ Fully decentralized resource management
- ❑ 260 DCs: 45 servers and 13 switches per Edge-DC

- ❑ Edge-DCs connected to the electrical grid.
- ❑ Three categories of Edge-DCs with:
  - ➔ Photovoltaic power plant
  - ➔ Photovoltaic plant + battery storage
  - ➔ no on-site power source

[3] Chiaraviglio, L., Mellia, M., & Neri, F. (2009, June). Energy-aware backbone networks: a case study. In 2009 IEEE International Conference on Communications Workshops (pp. 1-5). IEEE.

Algorithm

- Step 1: Fully distributed **resource allocation:** each controller uses Best Fit to allocate resources to the jobs submitted by nearby users.

- Step 2: Inter-DC **negotiation**: relocate the arriving job on the Edge-DCs site with the most available on-site energy

- Step 3: Application-level **performance degradation**: reasonably reduce the quality of the images on the Edge-DCs with energy on-site deficit to reduce containers power consumption

- Step 4: **Consolidation:** migrate the running containers in all the Edge-DCs on the minimum set of servers and switch off the idle ones.

Experimental conditions

- Hand made jobs
  - ➔ Arrival : Poisson distribution
  - ➔ Duration : Exponential distribution
  - ➔ RAM, CPU request, CPU Usage : Normal distribution

- Solar generation traces from Pecan street[4]

- Simulation on SimGrid

- Simulation on 4 weeks, results focussed on the last week

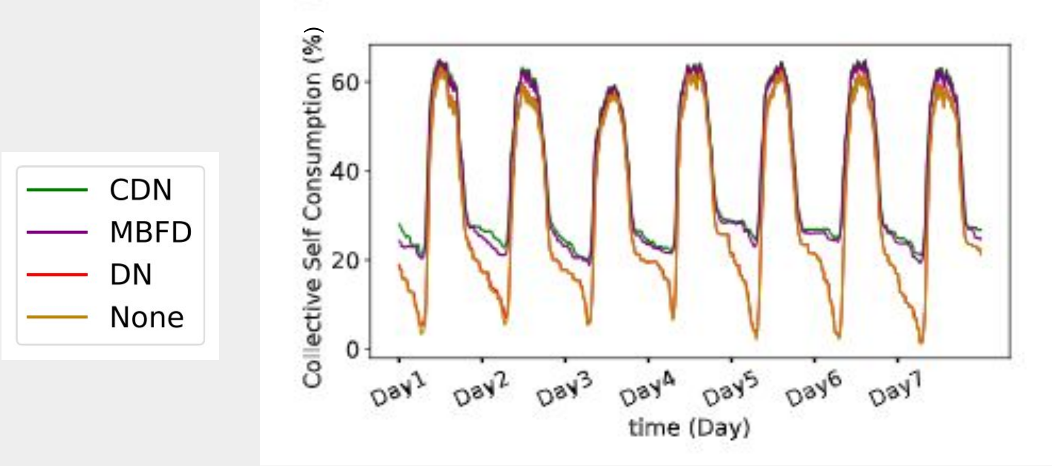
[4] Pecan street dataport : https://www.pecanstreet.org/dataport/, 2018

Results

Collective self-consumption improvement

- Collective self-consumption = (sum of on-site power consumption)/(total power consumption)

- Goal of the scenarios: show the impact of each part of our algorithm



CDN : Consolidation + Degradation + Negotiation

MBFD : Modified Best Fit Degrading (Best Fit + Consolidation)

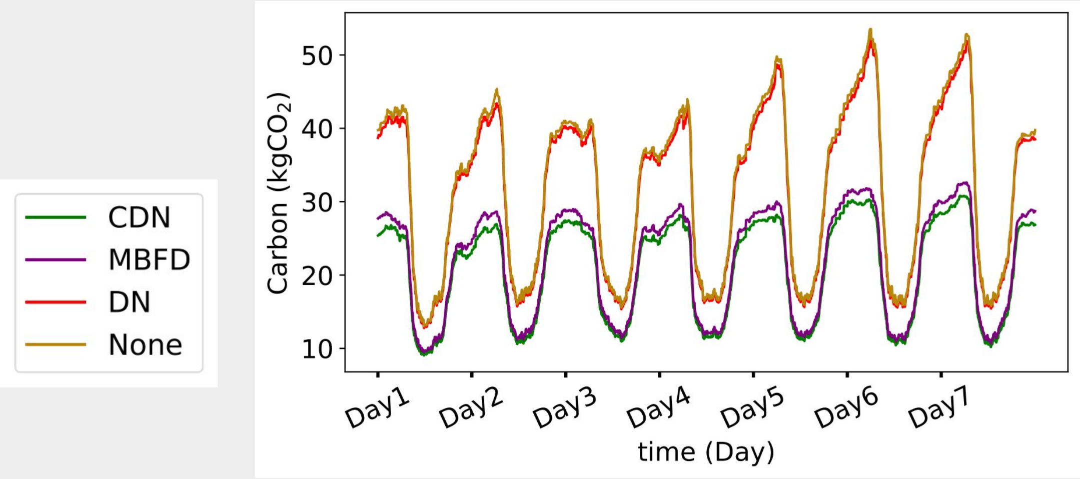
DN : Degradation + Negotiation

None : No optimization strategy

Results

Carbon emission savings

- The algorithm reduces the carbon consumption impact of the Edge-DCs (scope 2)



CDN : Consolidation + Degradation + Negotiation

MBFD : Modified Best Fit Degrading (Best Fit + Consolidation)

DN : Degradation + Negotiation

None : No optimization strategy

Conclusions

Summary

- ❑ Energy aware collaborative resources allocations
- ❑ QoS aware power load reduction (limited effect)
- ❑ Consolidation: key tool to considerably improve Collective Self-consumption and reduce carbon

Future work

- ❑ Investigate the sizing of the battery and the PV plants

Paper reference :

W. E. Gnibga, A. Blavette and A.-C. Orgerie, "Latency, Energy and Carbon Aware Collaborative Resource Allocation with Consolidation and QoS Degradation Strategies in Edge Computing", in IEEE International Conference on Parallel and Distributed Systems (ICPADS 2023).

Link hal

Contacts wedan-emmanuel.gnibga@irisa.fr
http://people.irisa.fr/Wedan.Gnibga

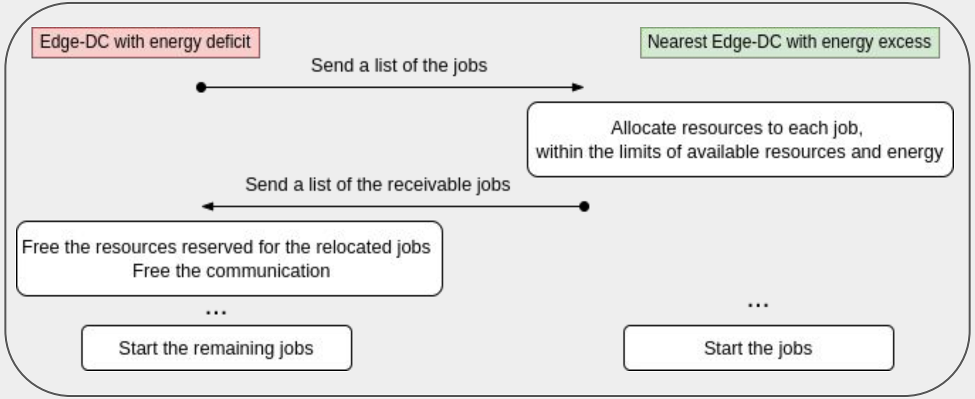
Methodology

Resource allocation and Negotiation

Objective : Minimize the exchanges of energy with the main grid (brown energy)
How : Deploy the jobs in the Edge-DCs with the highest production on-site

- ❑ Step 1: Fully distributed **resource allocation** using the **Best Fit** policy

- ❑ Step 2: Inter-DC **negotiation**: relocate the arriving job on the greenest Edge-DCs



<u>Destinations Edge-DCs hierarchy</u>

1. Edge-DCs with PV only
2. Edge-DCs with PV and battery
3. Edge-DCs with no on-site source (only in case of saturation)

❖ The P2P negotiations are limited to the metropolitan area to guarantee low-latency

Methodology

Application-level performance degradation

Objective : Reduce the power consumption of Edge-DCs due to computation
How : Stream images with a low resolution
When : power generation deficit (including battery low)

- The green-SLA defines the minimum resolution (px_{min}) acceptable by a given end-user

Optimization problem

- Find a trade-off between
  - the images quality reduction
  - the battery usage (when it is getting low)
  - grid electricity importations/carbon impact
- Respect the green-SLA

Methodology

Application-level performance degradation

Objective : Reduce the power consumption of Edge-DCs due to computation
How : Stream images with a low resolution
When : power generation deficit (including battery low)

Algorithm

- ❏ Determine the amount of power to shed as a function of
  - ❏ On-site power generation
  - ❏ Battery state of charge
  - ❏ Grid carbon intensity
- ❏ Determine the theoretical resolution of images streamed by all the jobs
- ❏ Choose the highest standard resolution (px_0) close to the theoretical value
- ❏ Finally, apply the highest resolution between px_0 and px_{min}

Methodology
Consolidation

Objective : Reduce the static power consumption of Edge-DCs
How : Relocate the containers of the less loaded nodes to the most loaded ones
When : Periodically

- ❑ Step 1: Try to reserve resources to each container of the less loaded node, on the other nodes using the Best Fit policy
- ❑ Step 2: Migrate the containers if all of them are assigned to destination nodes
- ❑ Step 3: Switch off the freed node

Thank you for your attention

Any question ?

wedan-emmanuel.gnibga@irisa.fr
http://people.irisa.fr/Wedan.Gnibga