
# Digital Sufficiency in Data Centers: Studying the Impact of User Behaviors

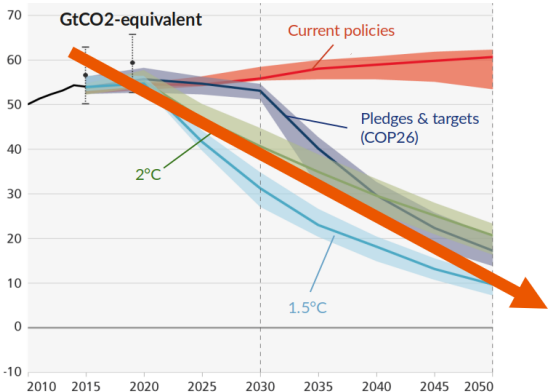
**Maël Madon**'s PhD defense

Thesis supervised by **Georges Da Costa** and **Jean-Marc Pierson**
Auditorium Jacques Herbrand, IRIT, Toulouse, France

# Growing greenhouse gases emissions

IPCC emission scenarios:

(**I**ntergovernmental **P**anel on **C**limate **C**hange)

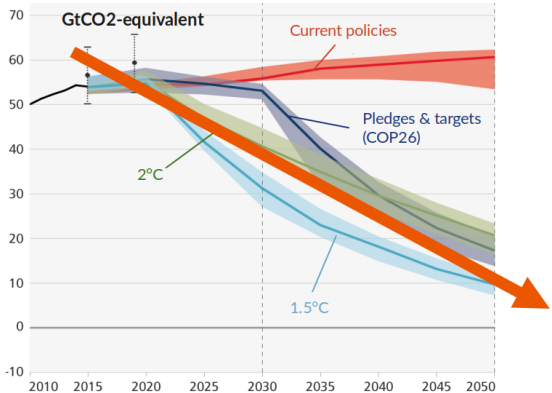



*Source: Valérie Masson-Delmotte, data from IPCC*

# Growing greenhouse gases emissions

## IPCC emission scenarios:

(**I**ntergovernmental **P**anel on **C**limate **C**hange)

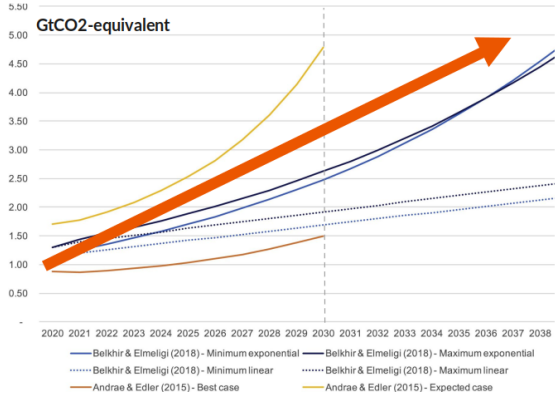



*Source: Valérie Masson-Delmotte, data from IPCC*

## IT industry =

(**I**nformation **T**echnology)

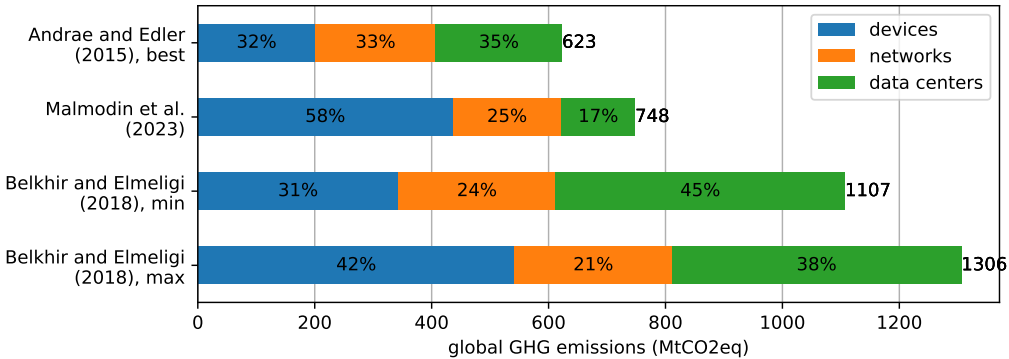
**2–4% global emissions**




*Source: Freitag et al. 2021 [1]*

# Where do IT emissions come from?

Global greenhouse gas emission estimates for IT industry in 2020:




*Source: Figures from Malmodin et al. original paper [2], and Freitag et al. supplementary material [1] for Andrae and Edler [3] and Belkhir and Elmeligi [4].*

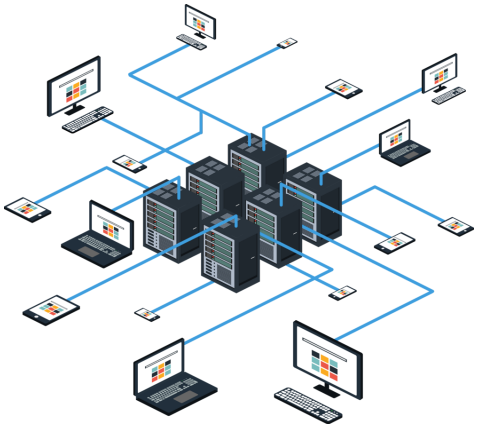
# Data centers

**Data center**

Room or building housing a group of computers connected to the network

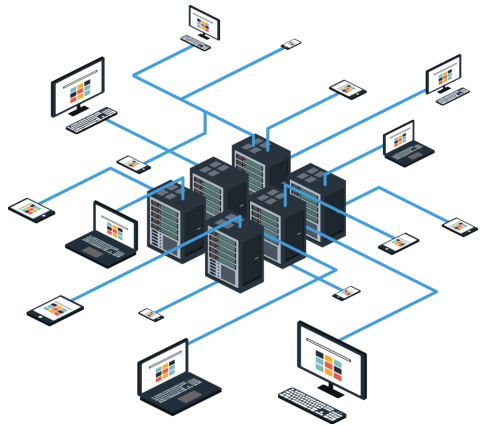
→ broad definition, from small server room to hyperscale infrastructure

- 240–340 TWh in 2022, or **1–1.3% global electricity demand** (IEA [5])



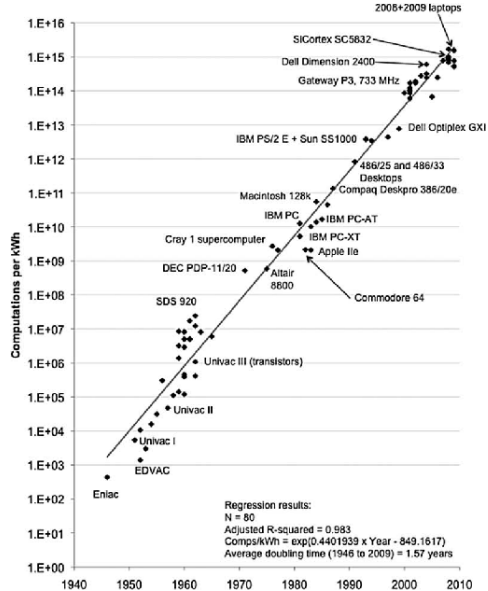
# Traditional footprint reduction techniques

- **Thermal and cooling management** [6]
  - Waste heat utilization
  - Free cooling
  - Thermal-aware scheduling
- **Energy-aware resource management** [7]
  - Server shutdown
  - Dynamic Voltage and Frequency Scaling
  - Leveraging platform heterogeneity
- **Use of renewable energies** [8]
  - Workload adaptation to power envelope
  - Geographic load shifting
  - Use of electricity storage systems



# Energy efficiency

- **Koomey's law** [9]: doubling the number of computations / kWh every 1.57 years




*Source: Koomey et al. 2011 [9]*

# Rebound effect

**Global trends in digital and energy indicators, 2015-2022**

| | 2015 | 2022 | Change |
|---|---|---|---|
| Internet users | 3 billion | 5.3 billion | +78% |
| Internet traffic | 0.6 ZB | 4.4 ZB | +600% |
| Data centre workloads | 180 million | 800 million | +340% |
| Data centre energy use (excluding crypto) | 200 TWh | 240-340 TWh | +20-70% |
| Crypto mining energy use | 4 TWh | 100-150 TWh | +2300-3500% |
| Data transmission network energy use | 220 TWh | 260-360 TWh | +18-64% |

*Source: IEA [5]*

# Rebound effect

**Global trends in digital and energy indicators, 2015-2022**

| | 2015 | 2022 | Change |
|---|---|---|---|
| Internet users | 3 billion | 5.3 billion | +78% |
| Internet traffic | 0.6 ZB | 4.4 ZB | +600% |
| Data centre workloads | 180 million | 800 million | +340% |
| Data centre energy use (excluding crypto) | 200 TWh | 240-340 TWh | +20-70% |
| Crypto mining energy use | 4 TWh | 100-150 TWh | +2300-3500% |
| Data transmission network energy use | 220 TWh | 260-360 TWh | +18-64% |

*Source: IEA [5]*

**Rebound effect**

Mechanism that reduce the potential energy savings from improved energy efficiency.

# Sufficiency

## Sufficiency policies (IPCC, 2022) [10]

A set of measures and daily practices that **avoid demand** for energy, materials, land and water **while delivering human well-being** for all within planetary boundaries.

# Sufficiency

## Sufficiency policies (IPCC, 2022) [10]

A set of measures and daily practices that **avoid demand** for energy, materials, land and water **while delivering human well-being** for all within planetary boundaries.

## Digital sufficiency (Santarius et al., 2022) [11]

Any strategy aimed at directly or indirectly **decreasing the absolute level of resource and energy demand from the production or application of IT**.

# Sufficiency

## Sufficiency policies (IPCC, 2022) [10]

A set of measures and daily practices that **avoid demand** for energy, materials, land and water **while delivering human well-being** for all within planetary boundaries.

## Digital sufficiency (Santarius et al., 2022) [11]

Any strategy aimed at directly or indirectly **decreasing the absolute level of resource and energy demand from the production or application of IT**.

Four dimensions of digital sufficiency:

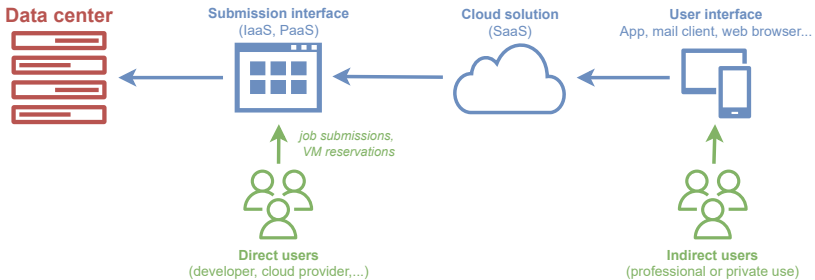
- **User sufficiency:** frugal use, IT for sufficiency-oriented lifestyles
- **Hardware sufficiency:** longevity, repairability
- **Software sufficiency:** long-term functionality, minimum data and utilization
- **Economic sufficiency:** IT for common good rather than economic growth

# Research questions

- **Which are the levers of user sufficiency in data centers?**

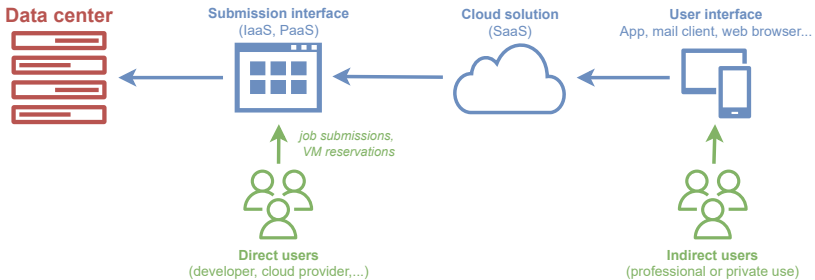
# Research questions

- **Which are the levers of user sufficiency in data centers?**



# Research questions

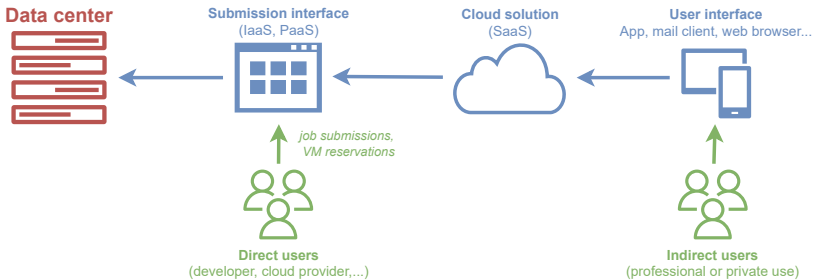
- **Which are the levers of user sufficiency in data centers?**




1. How to accurately model the interaction between **direct users** and the data center?

# Research questions

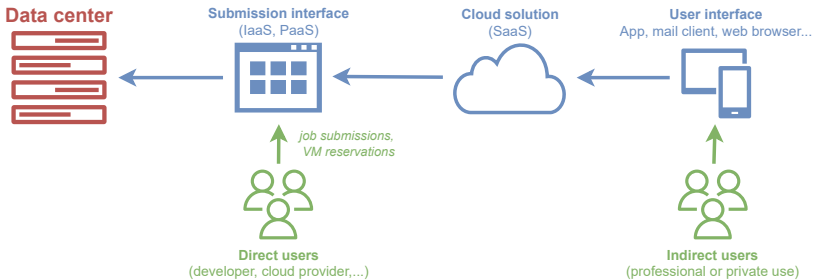
- **Which are the levers of user sufficiency in data centers?**




1. How to accurately model the interaction between **direct users** and the data center?
2. Which “sufficiency behaviors” can be adopted by **direct users**, and how does user effort translate into footprint reduction?

# Research questions

- **Which are the levers of user sufficiency in data centers?**




1. How to accurately model the interaction between **direct users** and the data center?
2. Which “sufficiency behaviors” can be adopted by **direct users**, and how does user effort translate into footprint reduction?
3. What are the opportunities for digital sufficiency for **indirect users**?

# Contents


1. Context and research problem
2. Sufficiency for direct data center users
   - Five “sufficiency behaviors”
   - Experimental characterization
   - Results
3. Sufficiency for indirect data center users
   - Study design
   - Findings
4. Open challenges
   - Using recorded workloads in simulations
   - Quantify user interactions
5. Conclusion

# Contents

1 Context and research problem

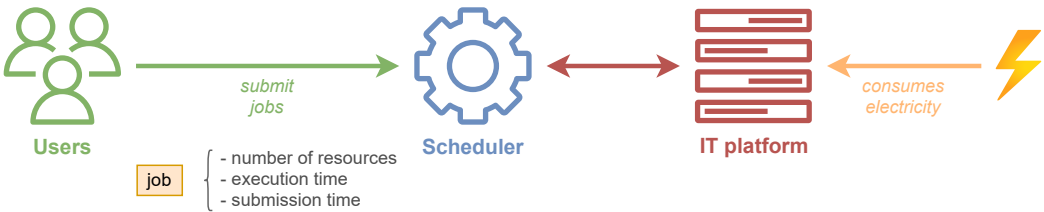
2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

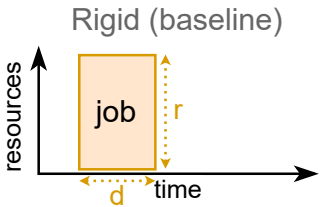
4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

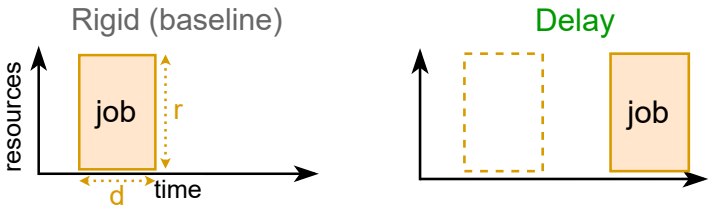
# Direct data center users



# “Sufficiency behaviors” for direct users

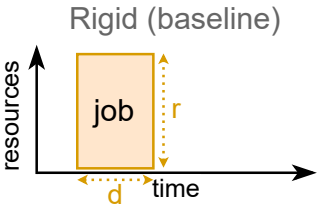


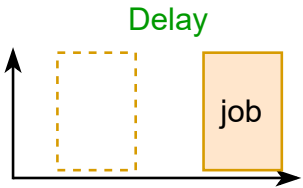
# “Sufficiency behaviors” for direct users



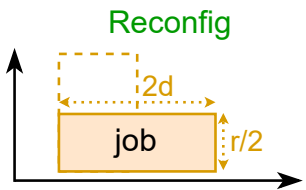

***ex:*** *submit tomorrow*

# “Sufficiency behaviors” for direct users



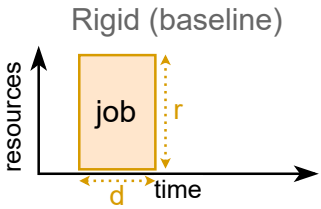



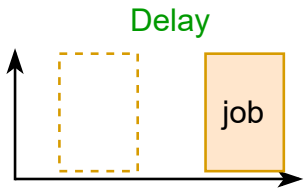

***ex:*** *submit tomorrow*



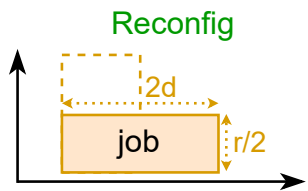

***ex:*** *fewer nodes for image processing*

# “Sufficiency behaviors” for direct users

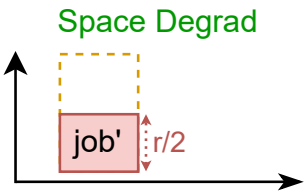



***ex:*** *submit tomorrow*

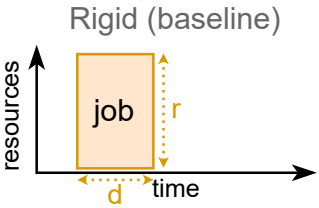


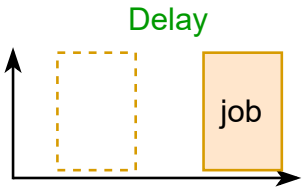

***ex:*** *fewer nodes for image processing*



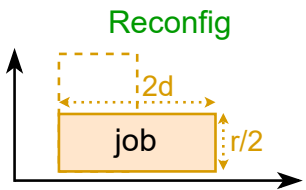

***ex:*** *only 5 outputs instead of 10*

# “Sufficiency behaviors” for direct users

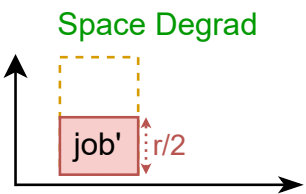


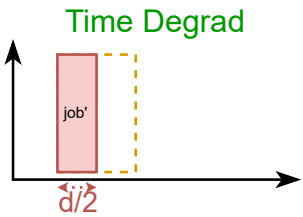

***ex:*** *submit tomorrow*




***ex:*** *fewer nodes for image processing*

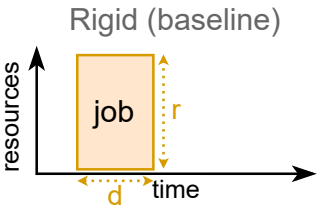


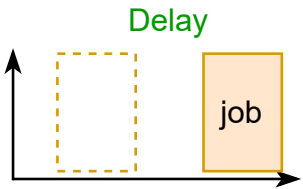

***ex:*** *only 5 outputs instead of 10*



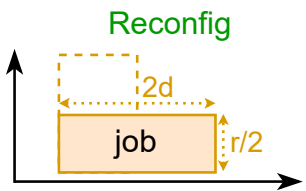

***ex:*** *lower accuracy in a linear solver*

# “Sufficiency behaviors” for direct users

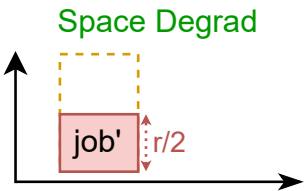


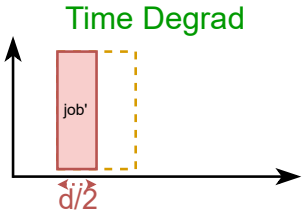

***ex:*** *submit tomorrow*



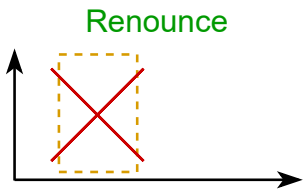

***ex:*** *fewer nodes for image processing*




***ex:*** *only 5 outputs instead of 10*

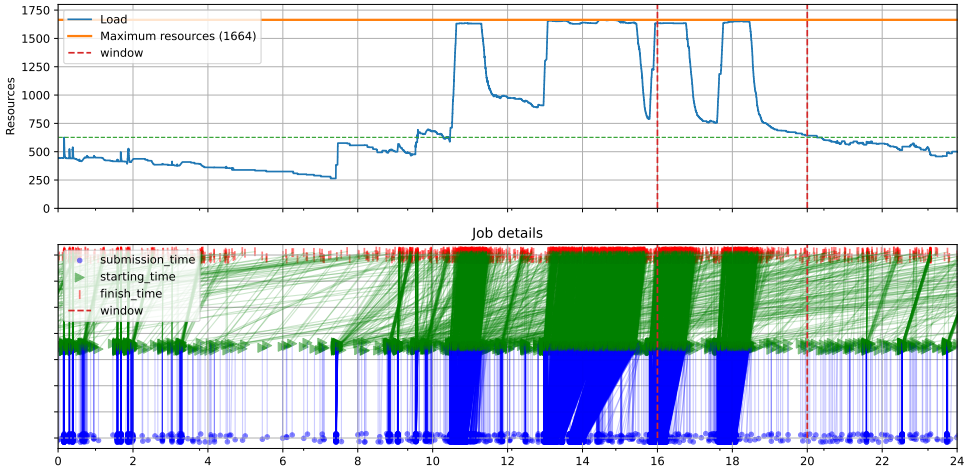



***ex:*** *lower accuracy in a linear solver*



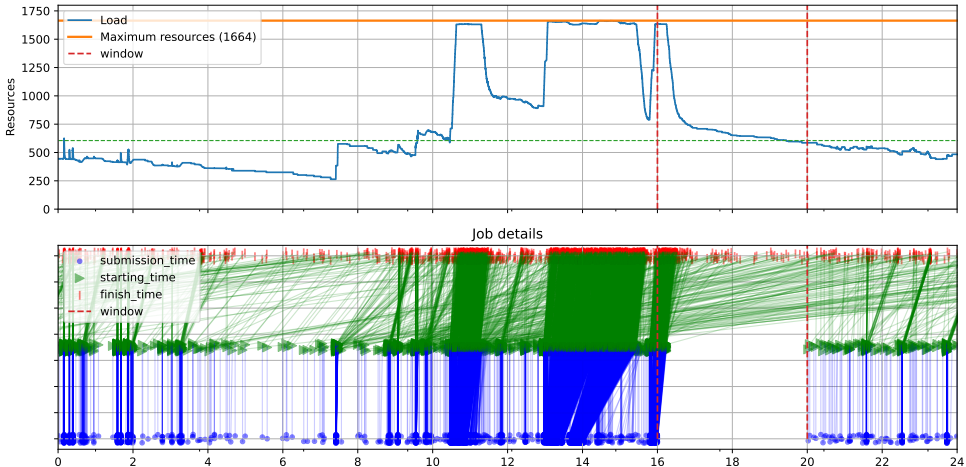
# Behaviors in practice

Replay of workload Metacentrum (Parallel Workloads), *Thursday June 5 2014*
→ behavior during window: **100% Rigid**



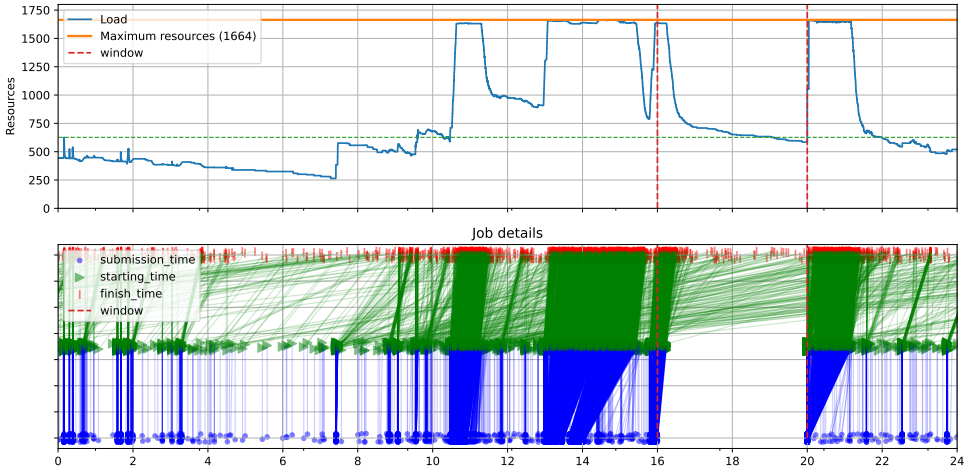
# Behaviors in practice

Replay of workload Metacentrum (Parallel Workloads), *Thursday June 5 2014*
→ behavior during window: **100% Renounce**



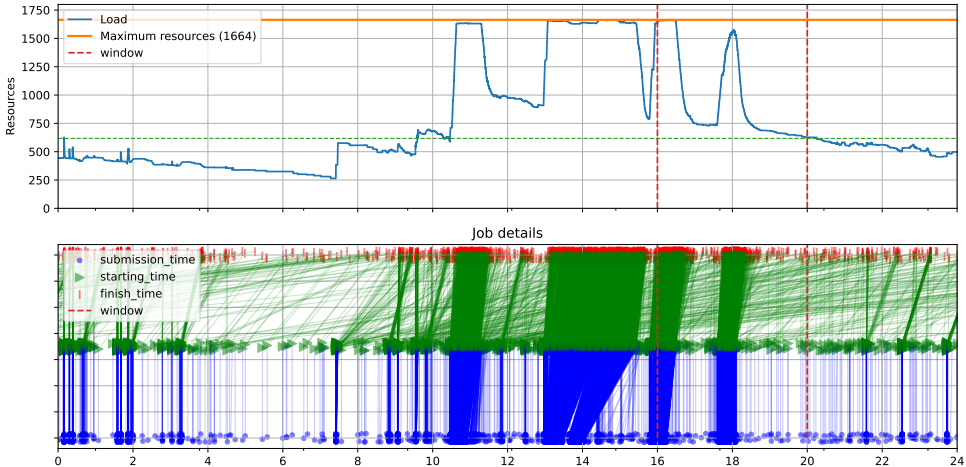
# Behaviors in practice

Replay of workload Metacentrum (Parallel Workloads), *Thursday June 5 2014*
→ behavior during window: **100% Delay**



# Behaviors in practice

Replay of workload Metacentrum (Parallel Workloads), *Thursday June 5 2014*
→ behavior during window: **100% Space Degrad**



# Contents

1 Context and research problem

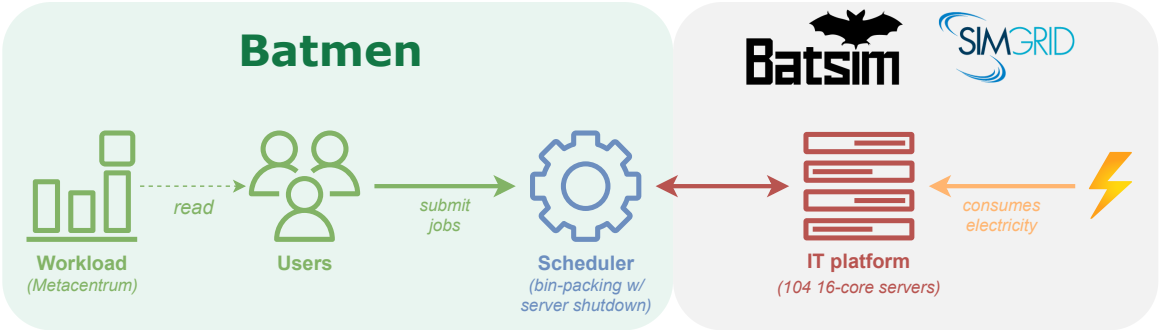
2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

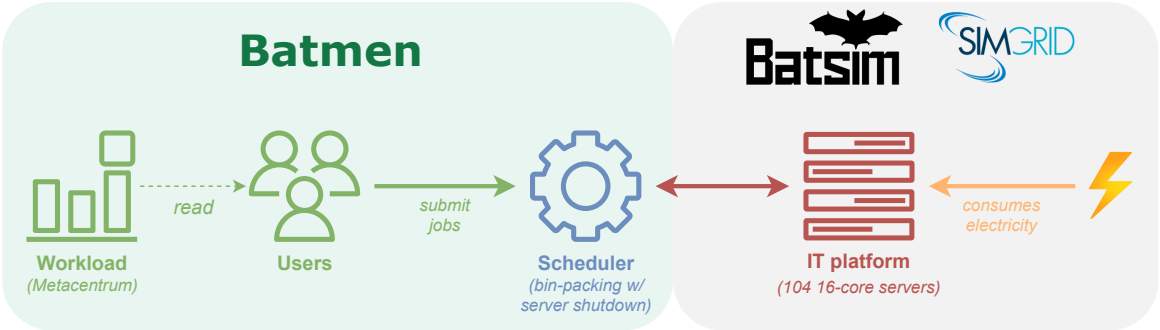
4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

# Simulation environment



# Simulation environment




- **open source** (LGPL) & **tested**
  - Batmen repository: https://gitlab.irit.fr/sepia-pub/mael/batmen
- using **trusted simulators**
- **reproducible** experimental campaign
  - https://gitlab.irit.fr/sepia-pub/open-science/demand-response-user

# Experimental campaign[1]

1. Cleaning the original workload Metacentrum2 (Parallel Workloads)
   - keeping only requested cores < 16 and execution time < 1 day

[1] M. Madon, G. Da Costa, and J.-M. Pierson, *Characterization of Different User Behaviors for Demand Response in Data Centers*, in Euro-Par 2022, 10.1007/978-3-031-12597-3_4

# Experimental campaign[1]

1. Cleaning the original workload Metacentrum2 (Parallel Workloads)
   - keeping only requested cores < 16 and execution time < 1 day
2. Sizing the platform to fit the workload
   - 104 16-core servers

[1] M. Madon, G. Da Costa, and J.-M. Pierson, *Characterization of Different User Behaviors for Demand Response in Data Centers*, in Euro-Par 2022, 10.1007/978-3-031-12597-3_4

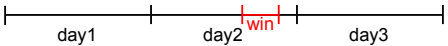
# Experimental campaign[1]

1. Cleaning the original workload Metacentrum2 (Parallel Workloads)
   - keeping only requested cores < 16 and execution time < 1 day
2. Sizing the platform to fit the workload
   - 104 16-core servers
3. Selecting 105 days to run the experiments
   - all weekdays between Jun 1, 2014 and Oct 23, 2014

[1] M. Madon, G. Da Costa, and J.-M. Pierson, *Characterization of Different User Behaviors for Demand Response in Data Centers*, in Euro-Par 2022, 10.1007/978-3-031-12597-3_4

# Experimental campaign[1]

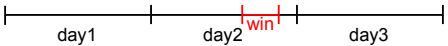
1. Cleaning the original workload Metacentrum2 (Parallel Workloads)
   - keeping only requested cores < 16 and execution time < 1 day
2. Sizing the platform to fit the workload
   - 104 16-core servers
3. Selecting 105 days to run the experiments
   - all weekdays between Jun 1, 2014 and Oct 23, 2014
4. Simulating the six behaviors independently on each day
   - to all jobs arriving in the window between 16:00 and 20:00




[1] M. Madon, G. Da Costa, and J.-M. Pierson, *Characterization of Different User Behaviors for Demand Response in Data Centers*, in Euro-Par 2022, 10.1007/978-3-031-12597-3_4

# Experimental campaign[1]

1. Cleaning the original workload Metacentrum2 (Parallel Workloads)
   - keeping only requested cores < 16 and execution time < 1 day
2. Sizing the platform to fit the workload
   - 104 16-core servers
3. Selecting 105 days to run the experiments
   - all weekdays between Jun 1, 2014 and Oct 23, 2014
4. Simulating the six behaviors independently on each day
   - to all jobs arriving in the window between 16:00 and 20:00




→ Total: (105 days) x (6 behaviors) = **630 3-day simulations**
*(the whole simulation campaign runs in <2h on a general purpose machine)*

[1] M. Madon, G. Da Costa, and J.-M. Pierson, *Characterization of Different User Behaviors for Demand Response in Data Centers*, in Euro-Par 2022, 10.1007/978-3-031-12597-3_4

# Contents

1 Context and research problem

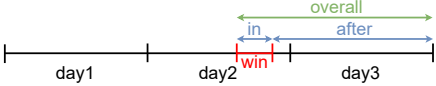
2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

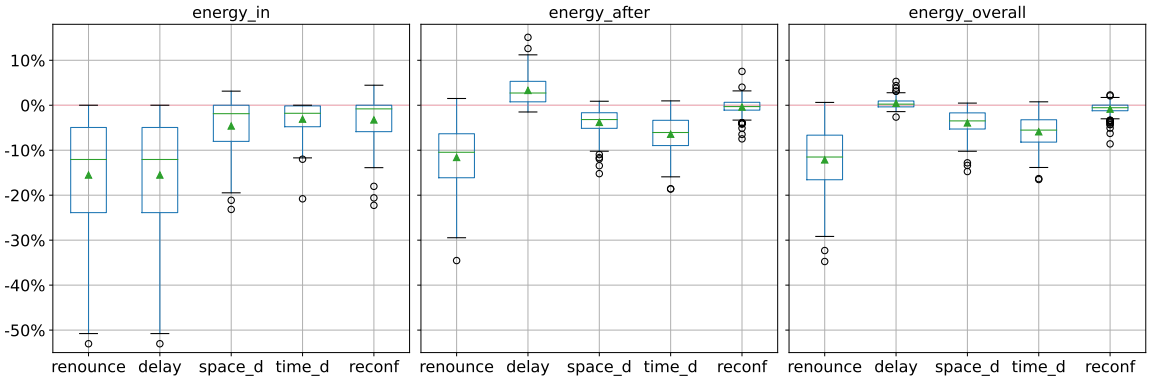
4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

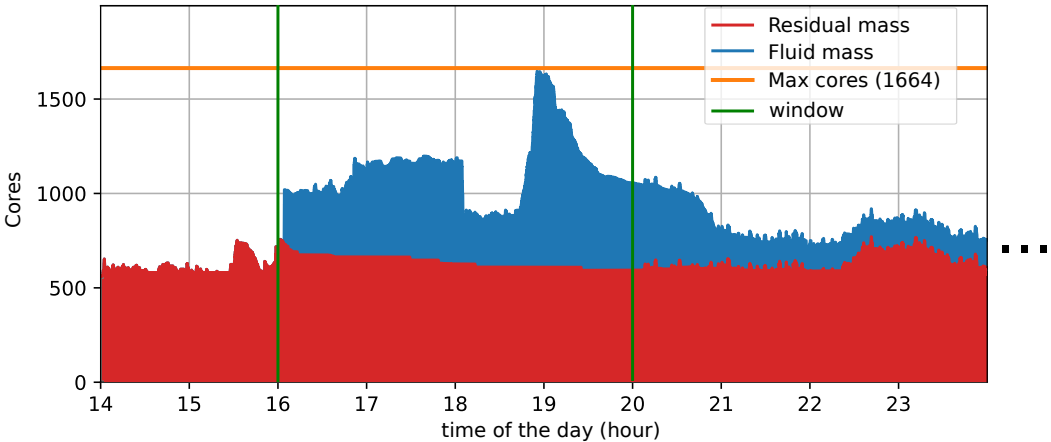
# Results: energy



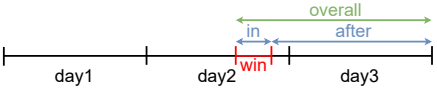

- **In % difference from the baseline** (Rigid behavior)



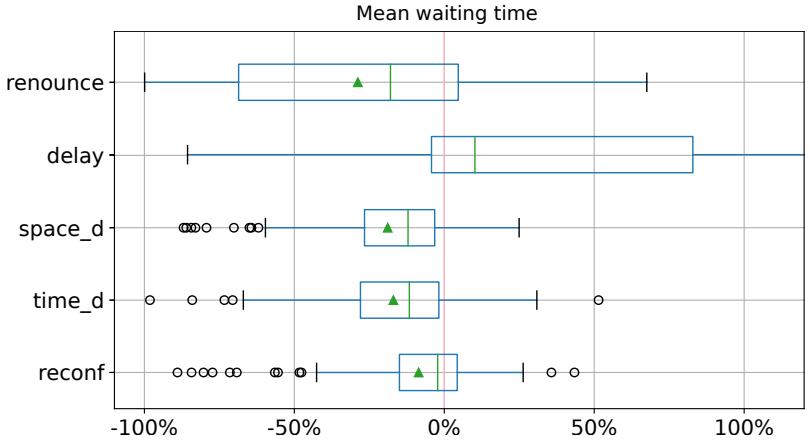
# Explanation of the energy results



# Results: scheduling metrics




- **In % difference from the baseline** (Rigid behavior)




(similar results with the scheduling metric *slowdown*)

# Pros and cons of each behavior

| behavior | energy in | energy overall | sched. metrics | "acceptability" |
|---|---|---|---|---|
| **Renounce** | 1st | 1st | 1st* | 5th |
| **Delay** | 1st | 5th | 5th | 2nd |
| **Space Degrad** | 3rd | 3rd | 2nd | 3rd |
| **Reconfig** | 3rd | 4th | 4th | 1st |
| **Time Degrad** | 5th | 2nd | 2nd | 3rd |

* scheduling metrics are not defined for renounced jobs

# In a renewable context[2]

- **Approach:**
  - windows = periods of low renewable energy production
  - users adopt a mix of sufficiency behaviors
  - 0–25–50–75–100% jobs modified in windows

[2] J. Gatt, M. Madon, and G. Da Costa, *Digital sufficiency behaviors to deal with intermittent energy sources in a data center*, accepted to ICT4S 2024.

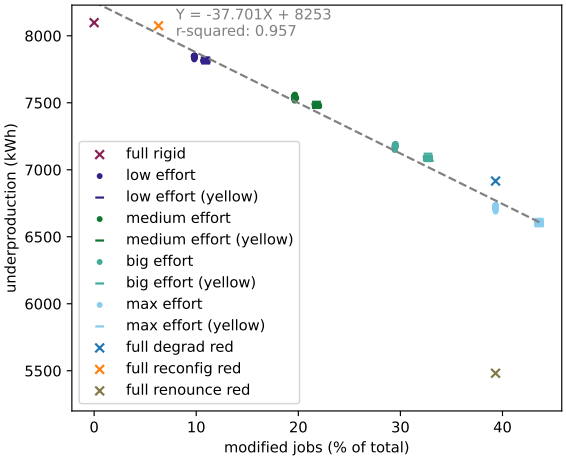
# In a renewable context[2]

- **Approach:**
  - windows = periods of low renewable energy production
  - users adopt a mix of sufficiency behaviors
  - 0–25–50–75–100% jobs modified in windows
- **Objective:** minimize underproduction (aka “brown energy”)

[2] J. Gatt, M. Madon, and G. Da Costa, *Digital sufficiency behaviors to deal with intermittent energy sources in a data center*, accepted to ICT4S 2024.

# In a renewable context[2]

- **Approach:**
  - windows = periods of low renewable energy production
  - users adopt a mix of sufficiency behaviors
  - 0–25–50–75–100% jobs modified in windows
- **Objective:** minimize underproduction (aka “brown energy”)
- **Results:**
  - energy savings linear with the size of the effort
  - maximum effort saves 18.4% underproduction compared to no effort




[2] J. Gatt, M. Madon, and G. Da Costa, *Digital sufficiency behaviors to deal with intermittent energy sources in a data center*, accepted to ICT4S 2024.

# Contents

1 Context and research problem

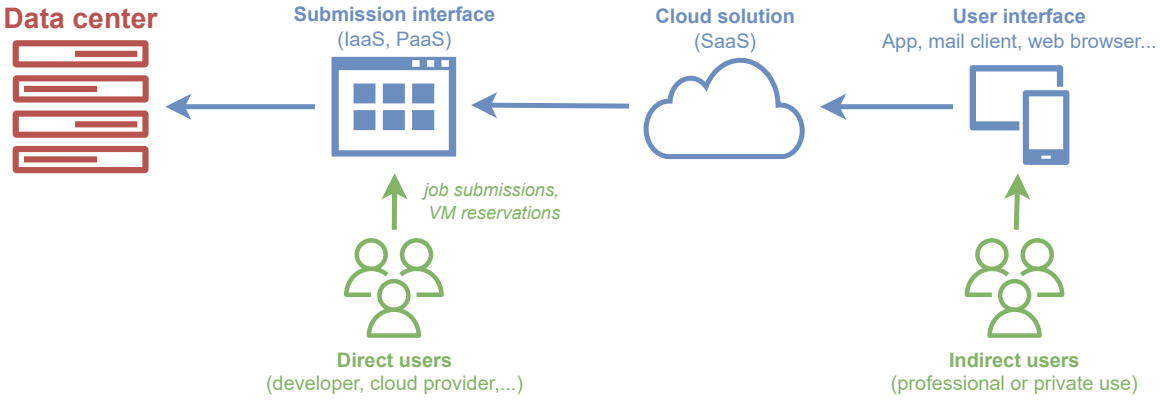
2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

# Data center user?




(cloud context)

# Digital Sufficiency in Flexible Work[3]

- **Focus:** professional cloud usage in a context of flexible work
- **Question:** which cloud usage is *necessary* and which one is *superfluous* according to the practitioners?
- **Method:** focus groups and thematic analysis
- Study carried out as part of a two-month research visit at the VU Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM


[3] M. Madon and P. Lago, *"We Are Always on, Is That Really Necessary?" Exploring the Path to Digital Sufficiency in Flexible Work*, in ICT4S 2023, 10.1109/ICT4S58814.2023.00012

# Qualitative research method

- **Data collection:**
  - 3 focus groups (Netherlands)
  - 2 companies (small IT company / large consulting firm)
  - 11 participants (consultants, developers, HR, ...)


- **Data analysis:**

**Video-recording** (5h)

↓

**Transcripts** (67 pages)

↓

**Open coding** (248 codes)

↓

**Thematic analysis** (12 main themes)

↓

**Codebook**

# Contents

1 Context and research problem

2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

# List of digital usages for work

| Task | # |
|---|---|
| email | 9 |
| messaging | 6 |
| planning | 5 |
| online meeting | 5 |
| phone | 4 |
| reviewing | 4 |
| project management | 4 |
| data analysis | 4 |
| preparing presentation | 3 |
| giving presentation | 3 |
| gathering information (internal) | 3 |
| gathering information (external) | 3 |
| writing time | 2 |
| writing documents | 2 |
| watching video | 1 |
| taking notes | 1 |
| online training | 1 |
| online presence | 1 |
| creating visuals | 1 |
| brainstorming | 1 |
| attending digital event | 1 |

**#**: number of mentions of the task in the focus group discussions

*“Everything is cloud-based in the work that we do”*

Emily, HR Manager, 46-55 yo

- Categories of cloud-based tasks:

interactive | offline with regular synchronization | off-cloud

# Necessary/superfluous: online meetings

*“There’s a bunch of meetings about meetings and pre calls for the meeting and then different meetings to evaluate the meetings. It’s a lot of... yeah... meetings”*

John, HR recruiter, 18-25 yo

| Necessary | Superfluous |
|---|---|
| Save travel time | Too many meetings |
| To keep human contact | Too long meetings |
| Easier to arrange | Recurring meetings |
| For team work | Duplication of channels |
| Camera for facial expressions | ... |
| ... | |

# Tactics towards sufficiency

- **48 tactics towards sufficiency** extracted from the discussions

## Examples


Turning the video off in an online meeting


Setting a lower value for the default meeting duration


Cancelling the next session of a recurring meeting when it has no purpose


Decreasing the frequency of a recurring meeting

# Tactics towards sufficiency

- **48 tactics towards sufficiency** extracted from the discussions

| Examples | Suff. behavior? |
|---|---|
|  Turning the video off in an online meeting | → Space Degrad |
|  Setting a lower value for the default meeting duration | → Time Degrad |
|  Cancelling the next session of a recurring meeting when it has no purpose | → Renounce |
|  Decreasing the frequency of a recurring meeting | → Delay? Degrad? |

# Tactics towards sufficiency

- **48 tactics towards sufficiency** extracted from the discussions

| Examples | Suff. behavior? |
|---|---|
|  Turning the video off in an online meeting | → Space Degrad |
|  Setting a lower value for the default meeting duration | → Time Degrad |
|  Cancelling the next session of a recurring meeting when it has no purpose | → Renounce |
|  Decreasing the frequency of a recurring meeting | → Delay? Degrad? |

- Categorization of the tactics into
  - 38 human-oriented / 5 system-oriented / 6 context-oriented

  → echoes digital sufficiency dimensions

# Contents

1 Context and research problem

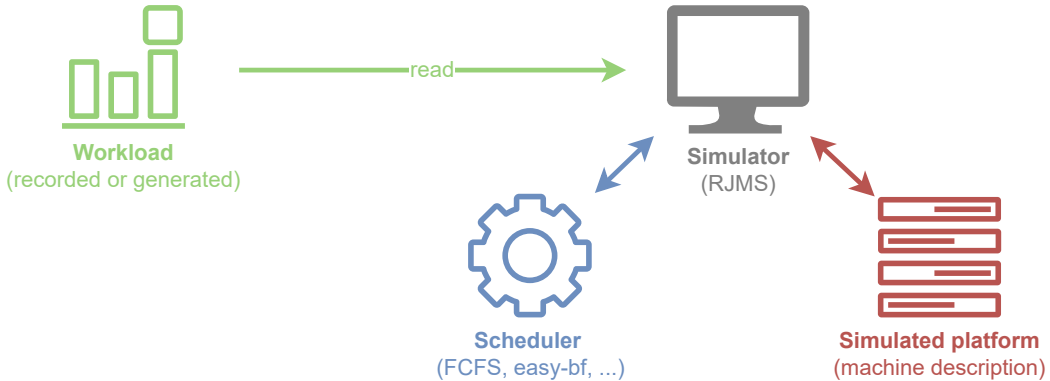
2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

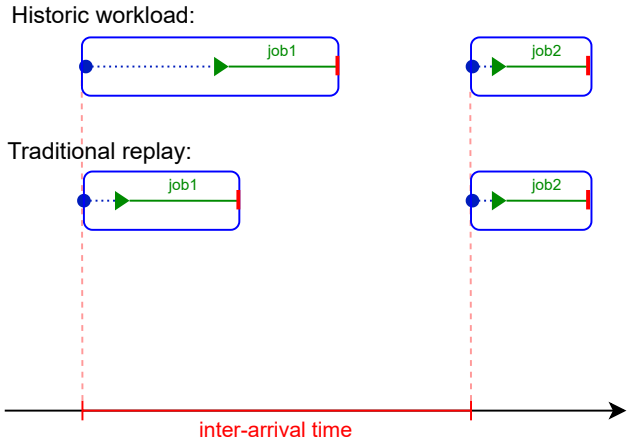
4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

# Traditional replay: principle

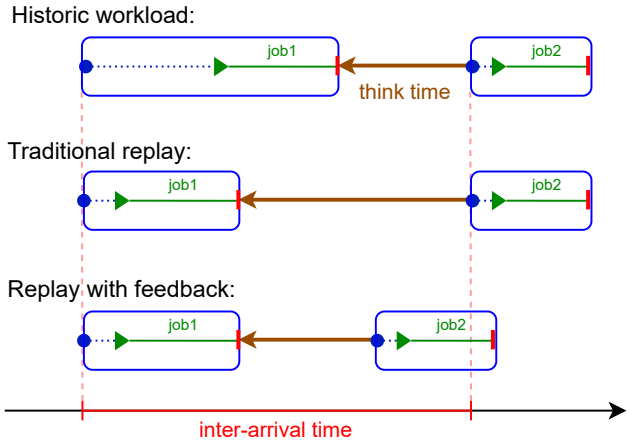


# Traditional replay: shortcomings



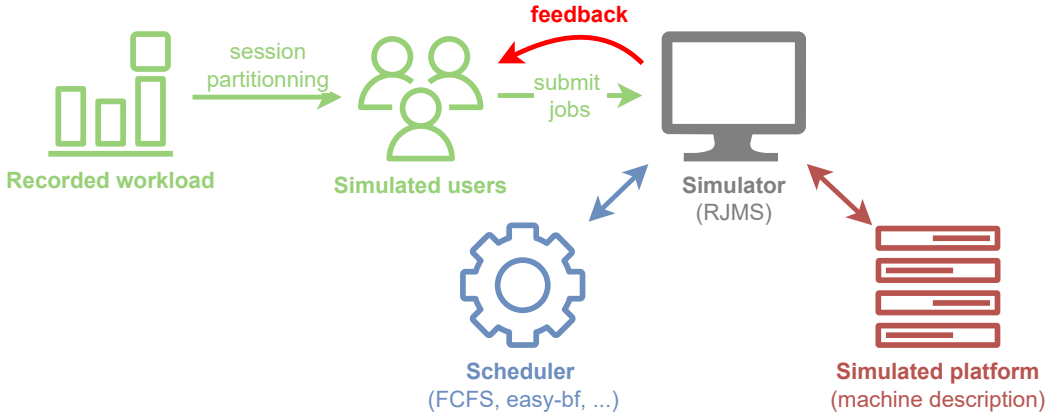

(original work from Zakay and Feitelson 2015 [12])

# Traditional replay: shortcomings



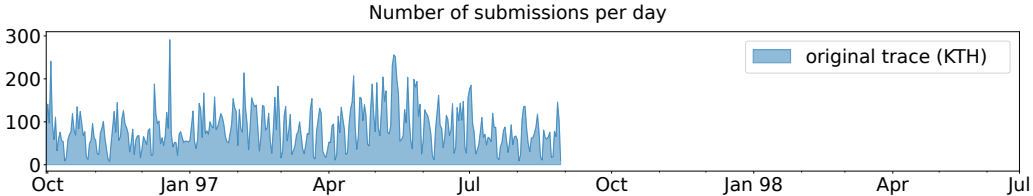

(original work from Zakay and Feitelson 2015 [12])

# Replay with feedback



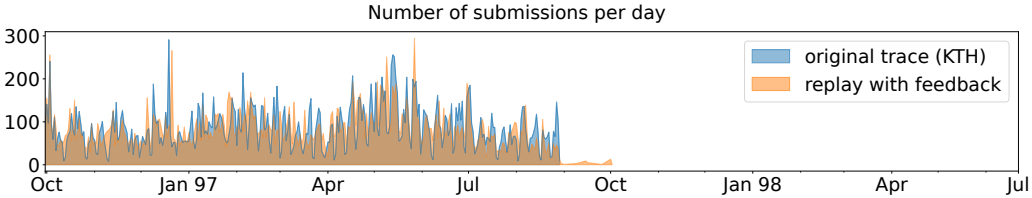

- Implementation available in **Batmen**: https://gitlab.irit.fr/sepia-pub/mael/batmen
- Reproducible experimental campaign:
  https://gitlab.irit.fr/sepia-pub/open-science/expe-replay-feedback

# Distribution of jobs' submission times[4]



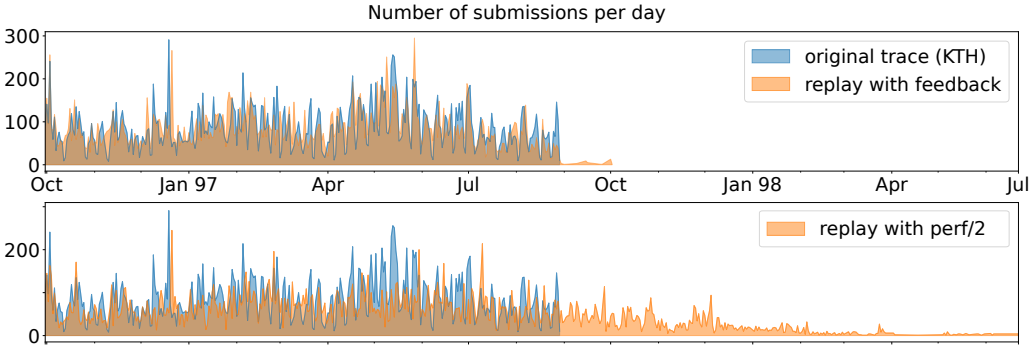

[4] M. Madon, G. Da Costa, and J.-M. Pierson, *Replay with Feedback: How Does the Performance of HPC System Impact User Submission Behavior?*, in Future Generation Computer Systems 2024, 10.1016/j.future.2024.01.024

# Distribution of jobs' submission times[4]



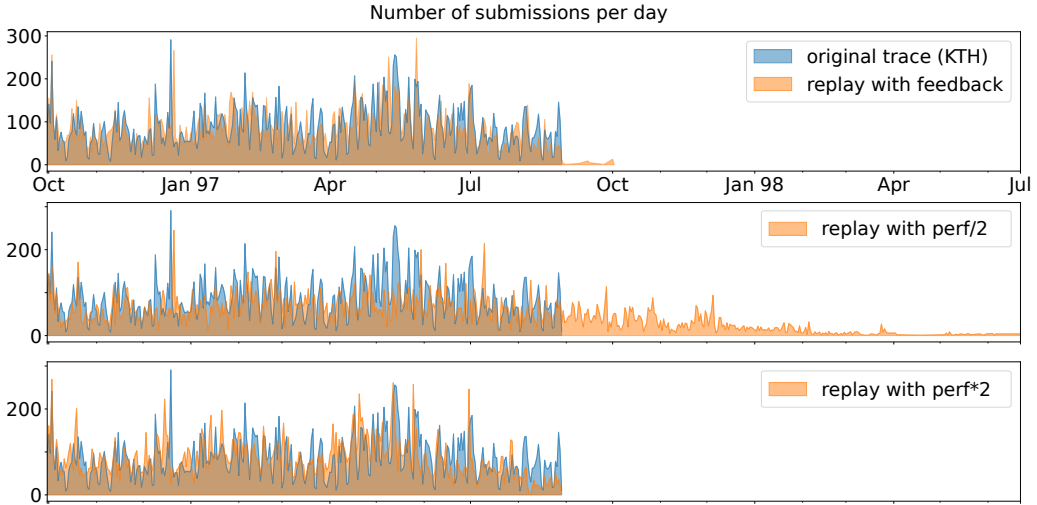

[4] M. Madon, G. Da Costa, and J.-M. Pierson, *Replay with Feedback: How Does the Performance of HPC System Impact User Submission Behavior?*, in Future Generation Computer Systems 2024, 10.1016/j.future.2024.01.024

# Distribution of jobs' submission times[4]




[4] M. Madon, G. Da Costa, and J.-M. Pierson, *Replay with Feedback: How Does the Performance of HPC System Impact User Submission Behavior?*, in Future Generation Computer Systems 2024, 10.1016/j.future.2024.01.024

# Distribution of jobs' submission times[4]




[4] M. Madon, G. Da Costa, and J.-M. Pierson, *Replay with Feedback: How Does the Performance of HPC System Impact User Submission Behavior?*, in Future Generation Computer Systems 2024, 10.1016/j.future.2024.01.024

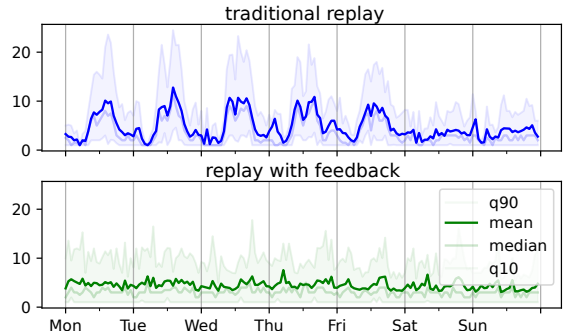
# Towards more realistic simulations

What remains?

# Towards more realistic simulations

What remains?

- account for day/night, weekday/weekend variability

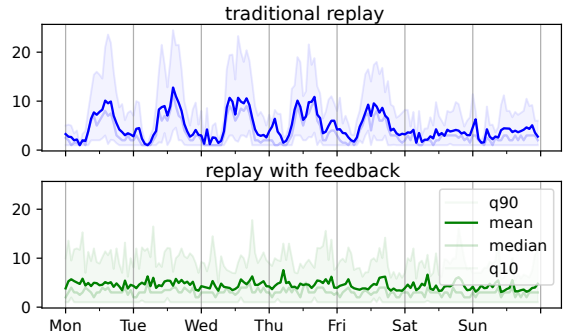



*Number of submissions per hour of the week, KTH log*

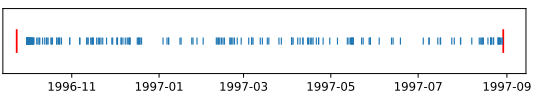
# Towards more realistic simulations

What remains?

- account for day/night, weekday/weekend variability
- consider arrival/departure of users




*Number of submissions per hour of the week, KTH log*

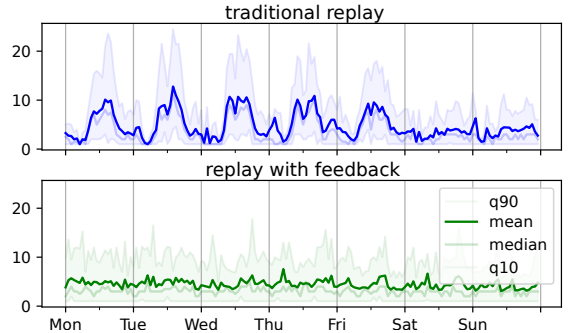



*User arrival in the platform, KTH log*

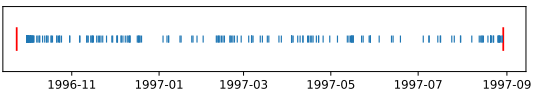
# Towards more realistic simulations

What remains?

- account for day/night, weekday/weekend variability
- consider arrival/departure of users
- study digital sufficiency behaviors with replay with feedback
- . . .




*Number of submissions per hour of the week, KTH log*



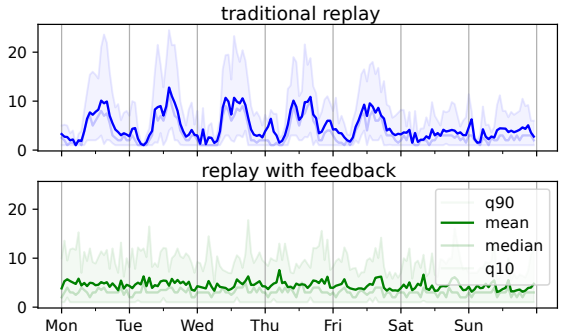

*User arrival in the platform, KTH log*

# Towards more realistic simulations

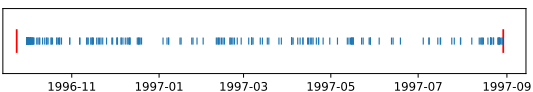
What remains?

- account for day/night, weekday/weekend variability
- consider arrival/departure of users
- study digital sufficiency behaviors with replay with feedback
- . . .

First and foremost: **scientific validation of the replay method**




*Number of submissions per hour of the week, KTH log*




*User arrival in the platform, KTH log*

# Contents

1 Context and research problem

2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

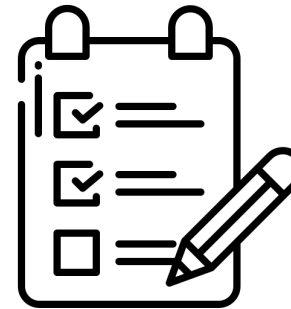
5 Conclusion

# Quantitative user studies

- Lack of quantitative data to evaluate:
  - the proportion of users willing to adopt the **sufficiency behaviors**
  - the impact of the **tactics towards sufficiency**
  - the model of **replay with feedback**

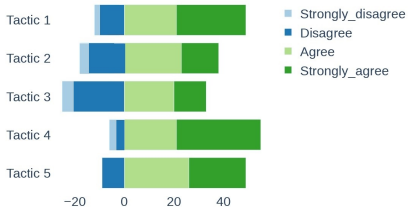
# Quantitative user studies

- Lack of quantitative data to evaluate:
  - the proportion of users willing to adopt the **sufficiency behaviors**
  - the impact of the **tactics towards sufficiency**
  - the model of **replay with feedback**
- How to obtain this data?
  - analyze existing data (infrastructure traces, macroeconomic data, ...)
  - roll out questionnaire surveys on a large sample of data center users

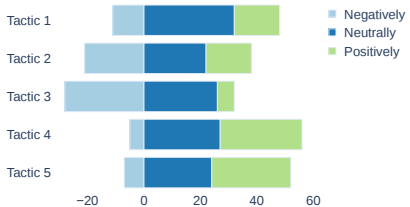


# Example: preference on applying tactics and effect on work productivity[5]

- Questionnaire survey with 61 responses:
  - **Tactic 1:** using off-cloud version of an application if the work doesn't need to be shared
  - **Tactic 2:** performing a task locally with regular cloud synchronization
  - **Tactic 3**: turning off camera or lowering video quality in meetings
  - **Tactic 4**: enclosing email attachment as a link
  - **Tactic 5**: closing an application, window or tab when it is not needed anymore




"I would consider applying this tactic in my daily work life"




"How would this tactic affect your work productivity?"

[5] M. Nawshin Rahman, M. Madon and P. Lago, *Sufficient Use of the Cloud for Work: Practitioners' Perception and Potential for Energy Saving*, accepted to ICT4S 2024

# Contents

1 Context and research problem

2 Sufficiency for direct data center users
- Five “sufficiency behaviors”
- Experimental characterization
- Results

3 Sufficiency for indirect data center users
- Study design
- Findings

4 Open challenges
- Using recorded workloads in simulations
- Quantify user interactions

5 Conclusion

# Conclusion

## Going beyond *efficiency*, investigating *sufficiency* for data centers

# Conclusion

**Going beyond *efficiency*, investigating *sufficiency* for data centers**

1. How to accurately model the interaction between **direct users** and the data center?
   - ✓ Batmen: open-source user simulator for Batsim
   - ✓ Replay with feedback

# Conclusion

**Going beyond *efficiency*, investigating *sufficiency* for data centers**

1. How to accurately model the interaction between **direct users** and the data center?
   - ✓ Batmen: open-source user simulator for Batsim
   - ✓ Replay with feedback
2. Which “sufficiency behaviors” can be adopted by **direct users**, and how does user effort translate into footprint reduction?
   - ✓ Five “sufficiency behaviors”: Renounce, Delay, Space Degrad, Time Degrad, Reconfig
   - ✓ A characterization of their energy saving potential and impact on scheduling metrics
   - ✓ A study of their usefulness in a renewable energy context

# Conclusion

**Going beyond *efficiency*, investigating *sufficiency* for data centers**

1. How to accurately model the interaction between **direct users** and the data center?
   - ✓ Batmen: open-source user simulator for Batsim
   - ✓ Replay with feedback
2. Which “sufficiency behaviors” can be adopted by **direct users**, and how does user effort translate into footprint reduction?
   - ✓ Five “sufficiency behaviors”: Renounce, Delay, Space Degrad, Time Degrad, Reconfig
   - ✓ A characterization of their energy saving potential and impact on scheduling metrics
   - ✓ A study of their usefulness in a renewable energy context
3. What are the opportunities for digital sufficiency for **indirect users**?
   - ✓ Qualitative study of digital needs for professional cloud users

# Publications

- Journal:
  - FGCS 2024: *Replay with Feedback: How Does the Performance of HPC System Impact User Submission Behavior?*, **M. Madon**, G. Da Costa and J.-M. Pierson
- Conferences:
  - Euro-Par 2022: *Characterization of Different User Behaviors for Demand Response in Data Centers*, **M. Madon**, G. Da Costa, and J.-M. Pierson
  - LIMITS 2022: *The Dark Side of Cloud and Edge Computing: An Exploratory Study*, K. Toczé, **M. Madon**, M. Garcia and P. Lago
  - ICT4S 2023: *"We Are Always on, Is That Really Necessary?" Exploring the Path to Digital Sufficiency in Flexible Work*, **M. Madon** and P. Lago
  - ICT4S 2024: *Digital sufficiency behaviors to deal with intermittent energy sources in a data center*, J. Gatt, **M. Madon**, and G. Da Costa
  - ICT4S 2024: *Sufficient Use of the Cloud for Work: Practitioners' Perception and Potential for Energy Saving*, M. Nawshin Rahman, **M. Madon** and P. Lago

# Bibliography I


[1] Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair, and Adrian Friday. “The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations”. In: *Patterns* 2.9 (Sept. 2021). ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100340.

[2] Jens Malmodin, Nina Lövehagen, Pernilla Bergmark, and Dag Lundén. “ICT Sector Electricity Consumption and Greenhouse Gas Emissions – 2020 Outcome”. In: *Telecommunications Policy* (Dec. 2023), p. 102701. ISSN: 0308-5961. DOI: 10.1016/j.telpol.2023.102701.

[3] Anders Andrae and Tomas Edler. “On Global Electricity Usage of Communication Technology: Trends to 2030”. In: *Challenges* 6.1 (Apr. 2015), pp. 117–157. ISSN: 2078-1547. DOI: 10.3390/challe6010117.

[4] Lotfi Belkhir and Ahmed Elmeligi. “Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations”. In: *Journal of Cleaner Production* 177 (Mar. 2018), pp. 448–463. ISSN: 09596526. DOI: 10.1016/j.jclepro.2017.12.239.

[5] IEA. *Tracking Data Centres and Data Transmission Networks*. Tech. rep. Paris, France: International Energy Agency (IEA), July 2023. URL: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks (visited on 02/26/2024).

# Bibliography II


[6] Ali Habibi Khalaj and Saman K. Halgamuge. “A Review on Efficient Thermal Management of Air- and Liquid-Cooled Data Centers: From Chip to the Cooling System”. In: *Applied Energy* 205 (Nov. 2017), pp. 1165–1188. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2017.08.037.

[7] Salil Bharany, Sandeep Sharma, Osamah Ibrahim Khalaf, Ghaida Muttashar Abdulsahib, Abeer S. Al Humaimeedy, Theyazn H. H. Aldhyani, Mashael Maashi, and Hasan Alkahtani. “A Systematic Survey on Energy-Efficient Techniques in Sustainable Cloud Computing”. In: *Sustainability* 14.10 (Jan. 2022), p. 6256. ISSN: 2071-1050. DOI: 10.3390/su14106256.

[8] G. Rostirolla et al. “A Survey of Challenges and Solutions for the Integration of Renewable Energy in Datacenters”. In: *Renewable and Sustainable Energy Reviews* 155 (Mar. 2022), p. 111787. ISSN: 1364-0321. DOI: 10.1016/j.rser.2021.111787.

[9] Jonathan Koomey, Stephen Berard, Marla Sanchez, and Henry Wong. “Implications of Historical Trends in the Electrical Efficiency of Computing”. In: *IEEE Annals of the History of Computing* 33.3 (Mar. 2011), pp. 46–54. ISSN: 1058-6180. DOI: 10.1109/MAHC.2010.28.

[10] IPCC. *Climate Change 2022: Mitigation of Climate Change. Summary for Policymakers.* Tech. rep. 6. IPCC, 2022. URL: https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_SPM.pdf (visited on 07/15/2022).

# Bibliography III


[11] Tilman Santarius, Jan C. T. Bieser, Vivian Frick, Mattias Höjer, Maike Gossen, Lorenz M. Hilty, Eva Kern, Johanna Pohl, Friederike Rohde, and Steffen Lange. “Digital Sufficiency: Conceptual Considerations for ICTs on a Finite Planet”. In: *Annals of Telecommunications* (May 2022). ISSN: 0003-4347, 1958-9395. DOI: 10.1007/s12243-022-00914-x.

[12] Netanel Zakay and Dror G. Feitelson. “Preserving User Behavior Characteristics in Trace-Based Simulation of Parallel Job Scheduling”. In: *Proceedings of the 8th ACM International Systems and Storage Conference*. Haifa Israel: ACM, May 2015, pp. 1–1. ISBN: 978-1-4503-3607-9. DOI: 10.1145/2757667.2778191.

[13] A. Orgerie, L. Lefèvre, and J. Gelas. “Save Watts in Your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems”. In: *2008 14th IEEE International Conference on Parallel and Distributed Systems*. Dec. 2008, pp. 171–178. DOI: 10.1109/ICPADS.2008.97.

[14] Cinzia Cappiello, Paco Melià, Barbara Pernici, Pierluigi Plebani, and Monica Vitali. “Sustainable Choices for Cloud Applications: A Focus on CO2 Emissions”. In: *ICT for Sustainability 2014 (ICT4S-14)*. Atlantis Press, Aug. 2014, pp. 352–358. ISBN: 978-94-6252-022-6. DOI: 10.2991/ict4s-14.2014.43.

[15] David Guyon, Anne-Cécile Orgerie, Christine Morin, and Deb Agarwal. “Involving Users in Energy Conservation: A Case Study in Scientific Clouds”. In: *International Journal of Grid and Utility Computing* 10.3 (Jan. 2019), pp. 272–282. ISSN: 1741-847X. DOI: 10.1504/IJGUC.2019.099667.

# Bibliography IV


[16] David Guyon, Anne-Cécile Orgerie, and Christine Morin. “Energy - Efficient IaaS-PaaS Co-Design for Flexible Cloud Deployment of Scientific Applications”. In: *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. Sept. 2018, pp. 69–76. DOI: 10.1109/CAHPC.2018.8645888.

[17] Robert Basmadjian, Juan Felipe Botero, Giovanni Giuliani, Xavier Hesselbach, Sonja Klingert, and Hermann De Meer. “Making Data Centers Fit for Demand Response: Introducing GreenSDA and GreenSLA Contracts”. In: *IEEE Transactions on Smart Grid* 9.4 (July 2018), pp. 3453–3464. ISSN: 1949-3061. DOI: 10.1109/TSG.2016.2632526.

[18] Arun Vishwanath, Fatemeh Jalali, Kerry Hinton, Tansu Alpcan, Robert W. A. Ayre, and Rodney S. Tucker. “Energy Consumption Comparison of Interactive Cloud-Based and Local Applications”. In: *IEEE Journal on Selected Areas in Communications* 33.4 (Apr. 2015), pp. 616–626. ISSN: 0733-8716. DOI: 10.1109/JSAC.2015.2393431.

[19] Leonhard Wattenbach, Basel Aslan, Matteo Maria Fiore, Henley Ding, Roberto Verdecchia, and Ivano Malavolta. “Do You Have the Energy for This Meeting?: An Empirical Study on the Energy Consumption of the Google Meet and Zoom Android Apps”. In: *IEEE/ACM International Conference on Mobile Software Engineering and Systems*. Pittsburgh Pennsylvania, May 2022. DOI: 10.1145/3524613.3527812.

# Credits

- slide 4: https://www.zagtech.com/solutions/data-center/
- slide 26: www.pexels.com and https://cdn.biblemoneymatters.com/wp-content/uploads/2008/07/needs-versus-wants.jpg
- icons slides 29, 31, 39: www.flaticon.com

# Related works sufficiency behaviors

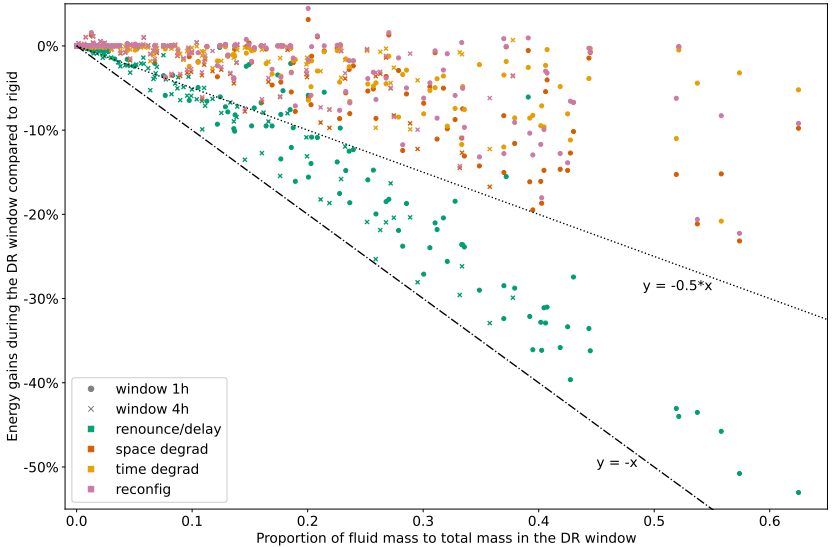
| | RE?* | Delay | Reconfig | Degrad | **Renounce** |
|---|---|---|---|---|---|
| Orgerie 2008 [13] | | ✓ | | | |
| Cappiello 2014 [14] | ✓ | ✓ | | | |
| Guyon 2019 [15] | | | ✓ | | |
| Guyon 2018 [16] | | ✓ | ✓ | | |
| All4Green 2018 [17] | ✓ | ✓ | | ✓ | |
| **This thesis** | ✓ | ✓ | ✓ | ✓ | ✓ |

*is the work in the context of Renewable Energy integration?

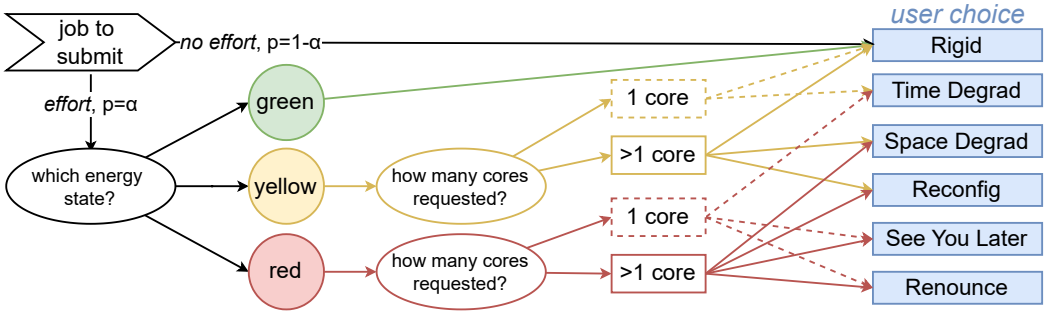
# Reproducible experimental environment

- input data from Parallel Workload Archive
- home-made tools for parsing (under GNU GPLv3)
  - batmen-tools
  - swf2userSession
- software version management
  - Git and Gitlab
  - declarative package manager Nix
- experiments and data analysis as jupyter notebooks
  - https://gitlab.irit.fr/sepia-pub/open-science/demand-response-user
  - https://gitlab.irit.fr/sepia-pub/open-science/sufficient-behaviors-with-renewables
  - https://gitlab.irit.fr/sepia-pub/open-science/expe-replay-feedback


# Energy gain in function of the fluid-residual ratio

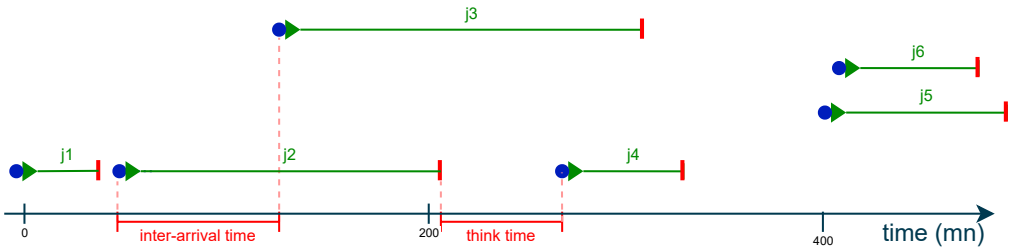


# 3-color eco-feedback energy model



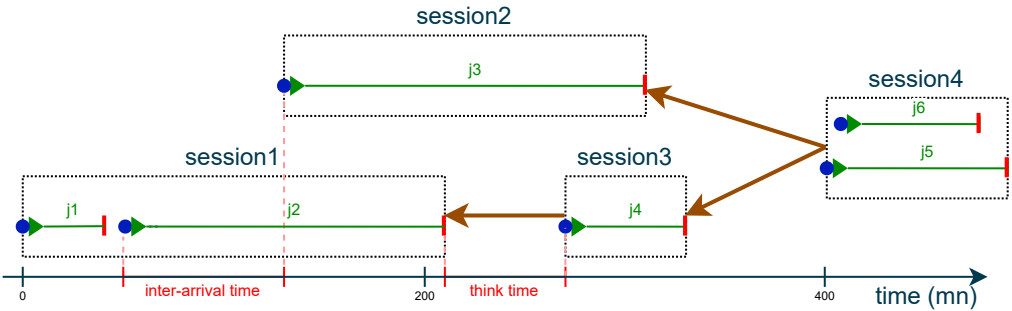
# Session partitioning

- based on threshold on inter-arrival time
- we tried two values: 0 minute and 60 minutes

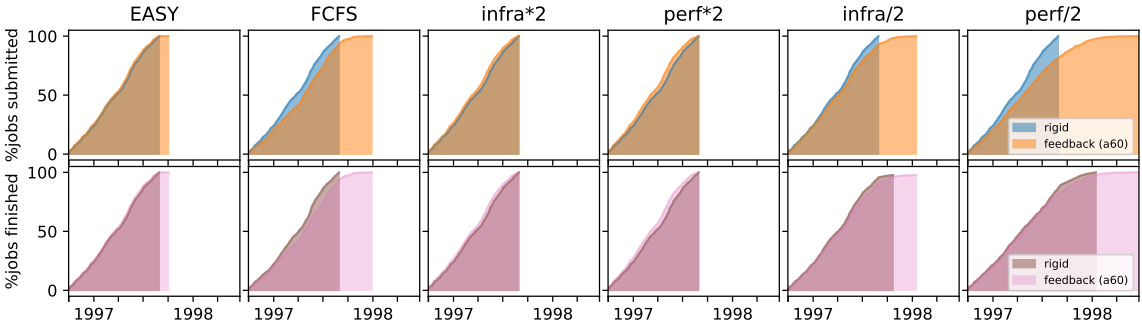


# Session partitioning

- based on threshold on inter-arrival time
- we tried two values: 0 minute and 60 minutes



# Cumulative number of jobs submitted (top) and finished (bottom)




*KTH log*

# New metrics for analysis

Traditional metrics (makespsan, waiting time, slowdown) do not make sense anymore. They are rather preserved in a system with feedback.

## Lateness

The **lateness** $\ell(i)$ of job $j_i$ is the difference between its submission time in the replay and in the original record: $\ell(i) = \hat{a_i} - a_i$.
We can define the **additional lateness** $\delta_{i+1}$ between job $j_i$ and the next job $j_{i+1}$: $\delta_{i+1} = \ell(i+1) - \ell(i)$.

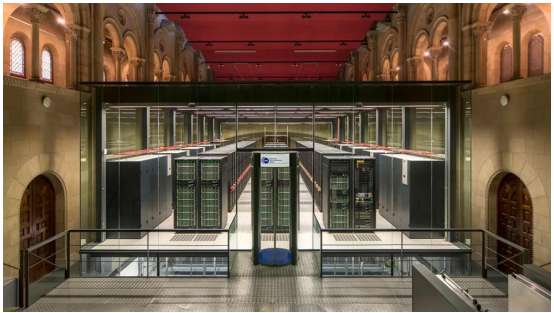
$$\text{mean lateness} = \bar{\ell} = \frac{1}{n}\sum_{i=0}^{n-1} \ell(i) = \frac{1}{n}\sum_{i=0}^{n-1} (\hat{a_i} - a_i) \tag{1}$$

$$\text{relative lateness} = 1 + \frac{\bar{\ell}}{a_{n-1} - a_0} \tag{2}$$

$$\text{additional lateness} = \delta = \frac{2\bar{\ell}}{n-1} \tag{3}$$

# Submission with feedback and malleability

- collaboration with Sergio Iserte, Barcelona Supercomputing Center (BSC)
  - previous work on malleable applications
  - access to Marenostrum (largest Spanish supercomputer)
- **objective**: submitter to Marenostrum (Slurm), accounting for feedback on the termination of malleable applications



# Energy savings of tactics: literature review

- **Tactic 1:** using off-cloud version of an application if the work doesn't need to be shared
- **Tactic 2:** performing a task locally with regular cloud synchronization
- **Tactic 3**: turning off camera or lowering video quality in meetings
- **Tactic 4**: enclosing email attachment as a link
- **Tactic 5**: closing an application, window or tab when it is not needed anymore

| | Energy saved | | Saved | Ref. |
|---|---|---|---|---|
| | Client-side | Server-side | data traffic | |
| **T.1** | 0.3-1W* | 0.25W** | all | [18] |
| **T.2** | 0.3-1W* | ∅ | 2-3 orders of magnitude*** | [18] |
| **T.3** | 4W† | ∅ | ∅ | [19] |
| **T.4** | ∼0 | ? | ∼0 | rough estimate |
| **T.5** | no statistical evidence | ∅ | 1 order of magnitude | our measurments |

∅ indicates "out of scope"
* word processing in google drive, difference between power consumption of netbook in offline and online edition (Table VI [18])
** assumption used in the article (Section IV.D.6 [18])
*** slopes in Figure 4 [18]
† mean values in Table 7 [19], converted to Watts