

Green Cloud Computing

Guest Lecture

Frankfurt University of Applied Sciences

envite 

Green Cloud Computing



Uwe Eisele
Sustainable Software
Architecture



Nadja Hagen
Sustainable Software
Architecture

Green Cloud Computing



Sustainability by IT

The climate crisis and global transformations pose new challenges for the use of new technologies and IT applications.

```
Map<Long, Double> threadsPower = computeTotalThreadsPowerConsumption();

// Now we allocate power for each method based on statistics
StringBuilder bufMeth = new StringBuilder();
for (Map.Entry<Long, Map<String, Integer>> entry : methodsStats.entrySet()) {

    for (Map.Entry<String, Integer> methEntry : entry.getValue().entrySet()) {
        String methName = methEntry.getKey();
        double methPower = saveEnergy(threadID) * (methEntry.getValue());
        if (methodsEnergy.containsKey(methEntry.getKey())) {
            // Add power (for 1 sec = energy) to total method energy
            double newMethEnergy = methodsEnergy.get(methName) + methPower;
            methodsEnergy.put(methName, newMethEnergy);
        } else {
            methodsEnergy.put(methName, methPower);
        }
        bufMeth.append(methName).append(" ").append(methPower).append(" ");
    }
}
```

Sustainability in IT

Software no longer has to be only in-time, in-function, in-budget and in-quality, but increasingly also in-climate.

Green Cloud Computing

How high do you estimate the energy consumption by the cloud?

What do you think are the biggest challenges in operating a software system in the cloud?

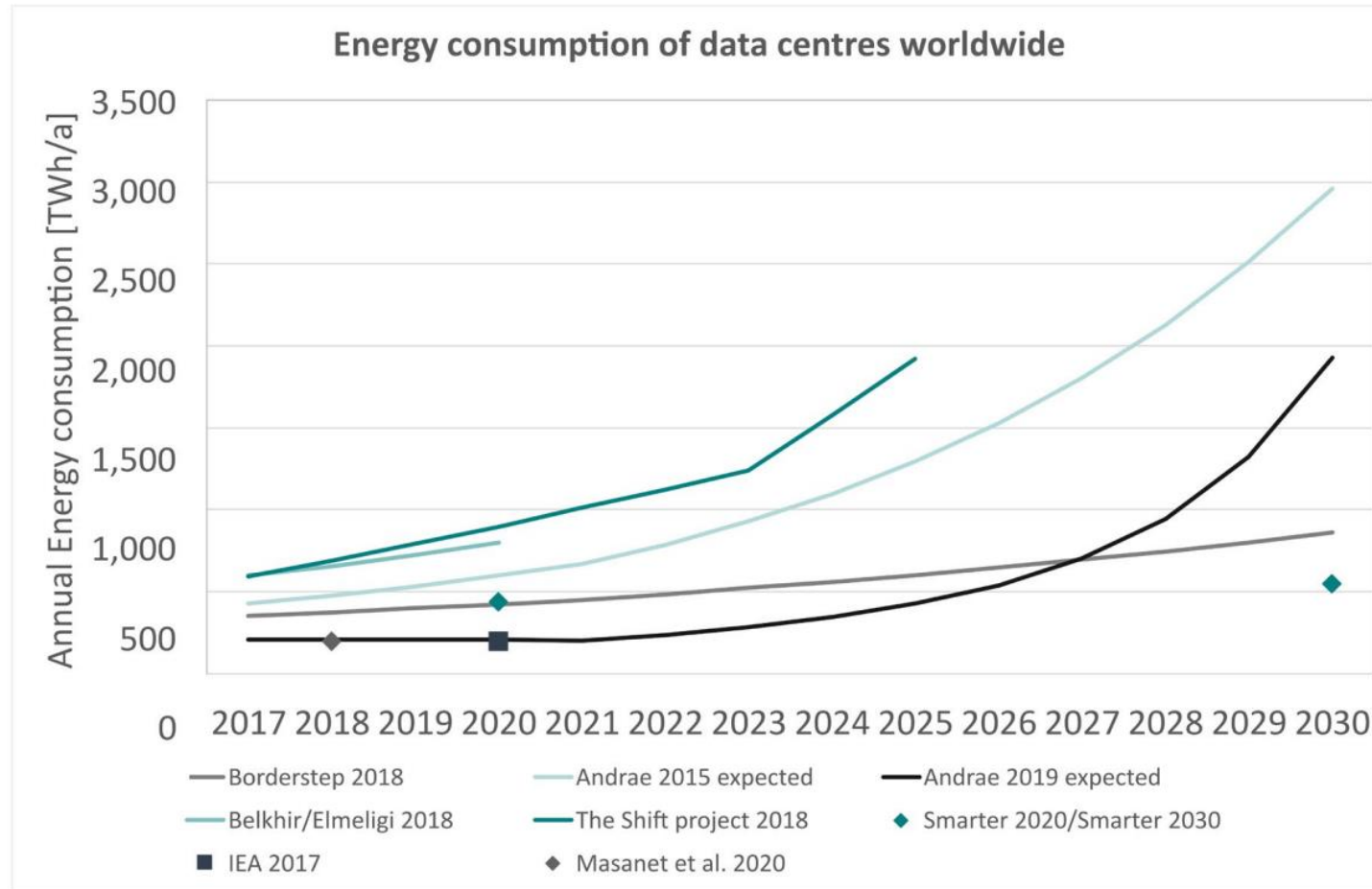
Green Cloud Computing

- Energy Consumption by Data Centers
- Metrics & Sustainability of Cloud Providers
- Resource Utilization in the Cloud
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Cloud Native Software Development
- Rebound Effects

Green Cloud Computing

- **Energy Consumption by Data Centers**
- Metrics & Sustainability of Cloud Providers
- Resource Utilization in the Cloud
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Cloud Native Software Development
- Rebound Effects

Energy Consumption by Data Centers

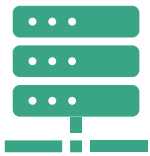


By 2030, data centers could represent **2.5% to 19% of annual global electricity consumption!**

Environment Agency Austria & Borderstep Institute: Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market (2021). European Commission.

Energy Efficiency in Cloud Computing

Why is Cloud Computing more energy efficient than operating On-Premises?



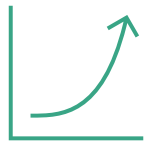
Dynamic Provisioning

- Traditional data centers are build for worst-case scenarios
- Cloud Computing can help to avoid long-term overprovisioning



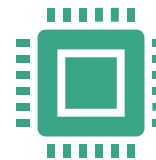
Multi-Tenancy

- Cloud providers serve multiple customers on the same infrastructure
- High number of customers flattens individual peaks



Server Usage

- On-Premises infrastructure has usually low utilization rates

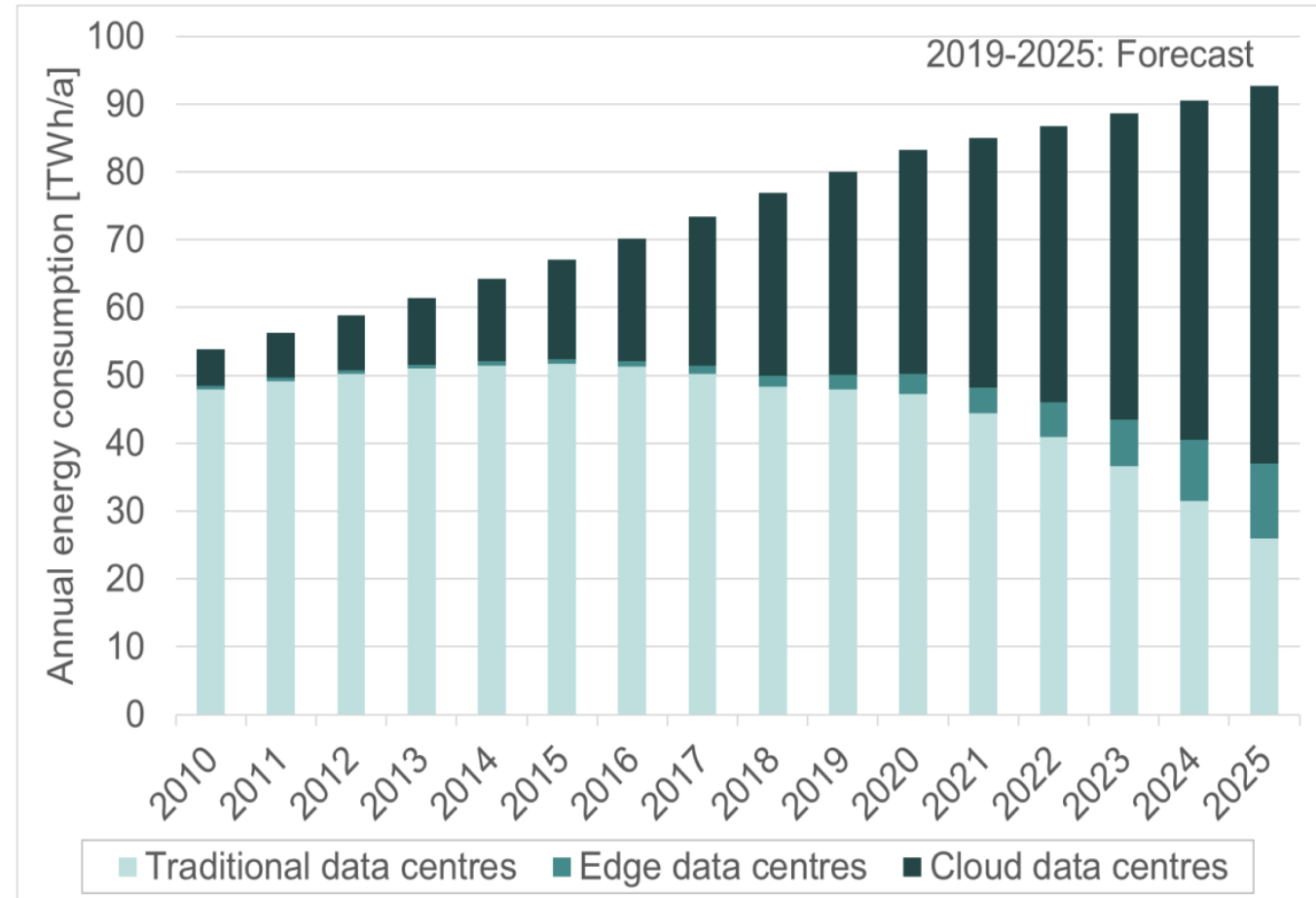


Hardware Efficiency

- Cloud data centers usually have a lower PUE value
- Use of modern technologies is more cost-effective

Energy Consumption by Data Centers – Global

- Share of cloud data centers is steadily increasing
- We already know:
 - In many cases, Cloud Computing is more energy-efficient than operating On-Premises.
 - Nevertheless, energy consumption by data centers is rising continuously
- **Resource consumption must also be reduced in the cloud!**



Environment Agency Austria & Borderstep Institute: Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market (2021). European Commission.

Energy Consumption by Data Centers – Germany

- Energy demand of data centers in Germany has increased in recent years
- Increase in energy demand coincides with increase in energy efficiency
- PUE fell from 1.98 to 1.63 from 2010 to 2020

More efficient data centers are not sufficient to counter the rising energy demand!

Energiebedarf von Rechenzentren in Deutschland

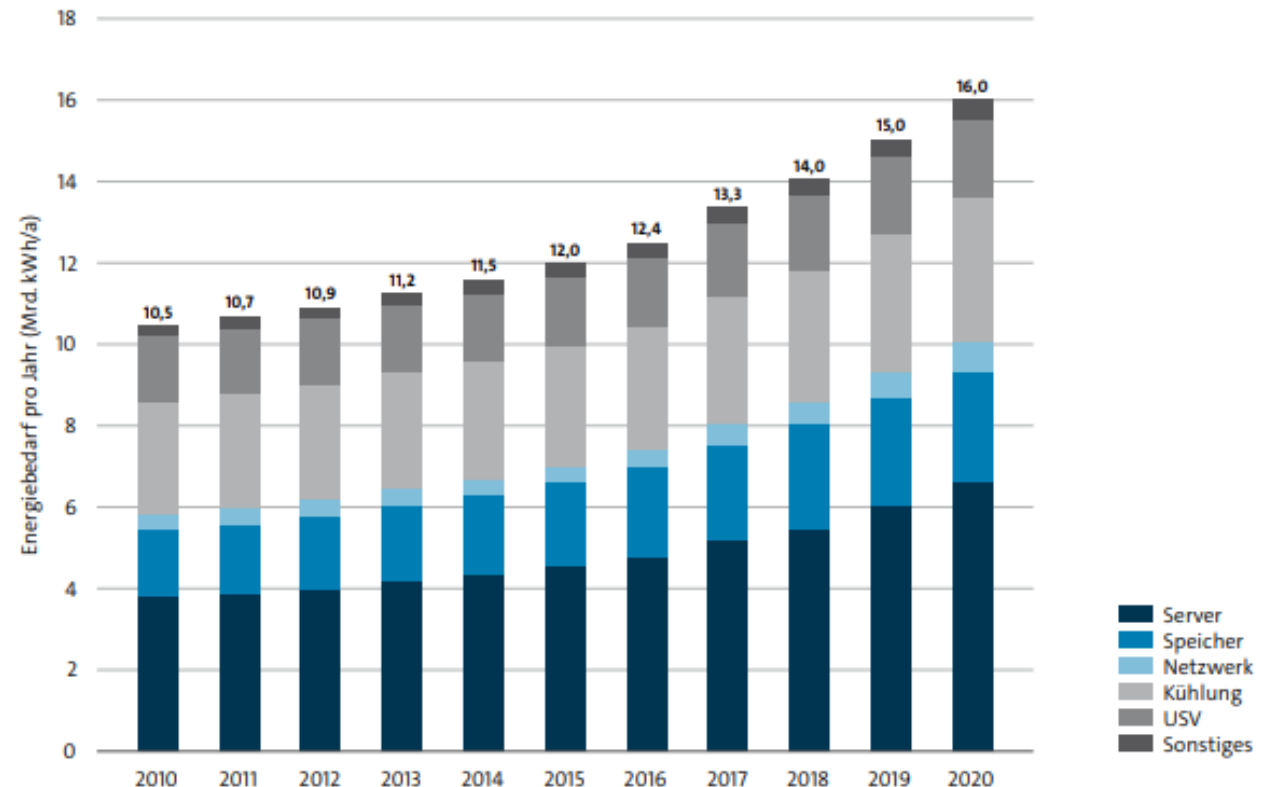


Abbildung 28 – Energiebedarf der Rechenzentren und kleineren IT-Installationen in Deutschland (in Mrd. kWh/a)

Quelle: Borderstep 2020

Energy Consumption by Data Centers – Germany

- 80% of greenhouse gas emissions from data centers are caused by **electricity demand**
- Greenhouse gas emissions have decreased due to improved electricity mix

In order to be able to cover the entire electricity demand completely by renewable energies, **savings are necessary!**

Treibhausgasemissionen der Rechenzentren in Deutschland

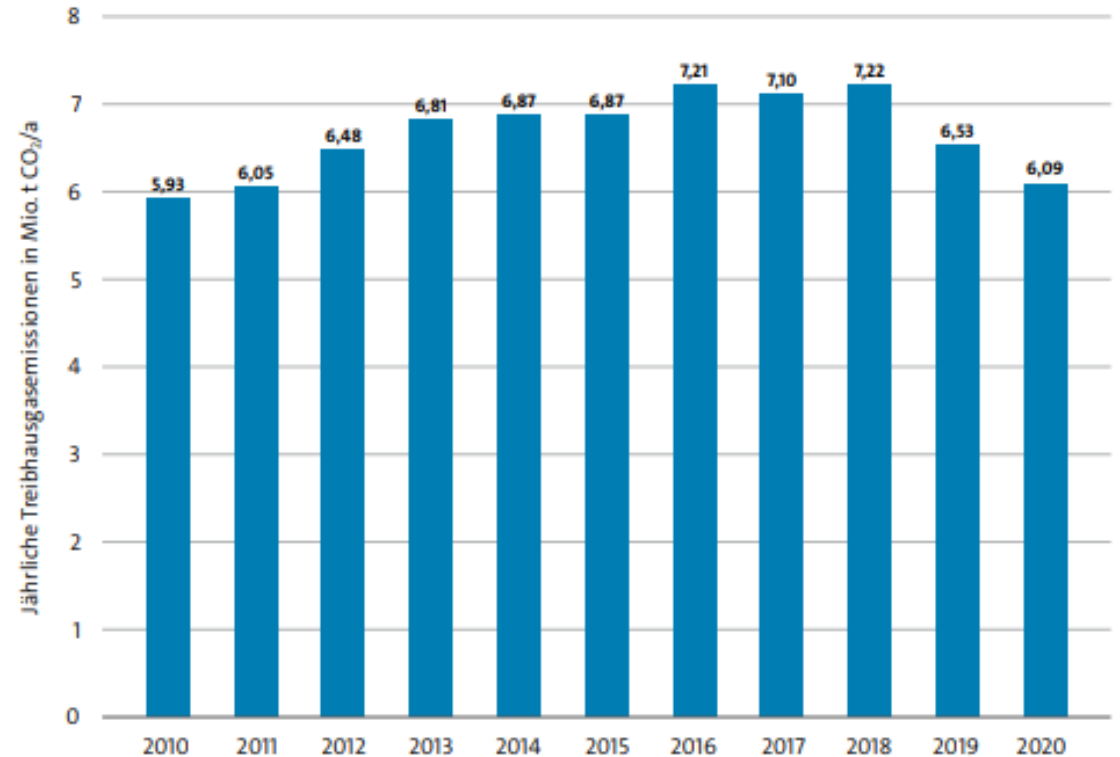


Abbildung 29 – Treibhausgasemissionen durch den Stromverbrauch der Rechenzentren und kleineren IT-Installationen in Deutschland (in Mio. t CO₂/a)

Green Cloud Computing

- Energy Consumption by Data Centers
- **Metrics & Sustainability of Cloud Providers**
- Resource Utilization in the Cloud
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Cloud Native Software Development
- Rebound Effects

Metrics – Carbon Proxies

How to measure carbon emissions of cloud operations or software applications?

- Direct measurement of carbon emissions is impossible in most cases
- Alternative metrics are needed to approximate the effectiveness of optimizations

Solution: Carbon Proxies

→ Use of metrics that correlate with carbon emissions



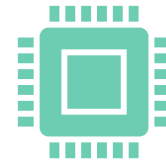
Electricity



Storage,
Memory



Costs



CPU
Load

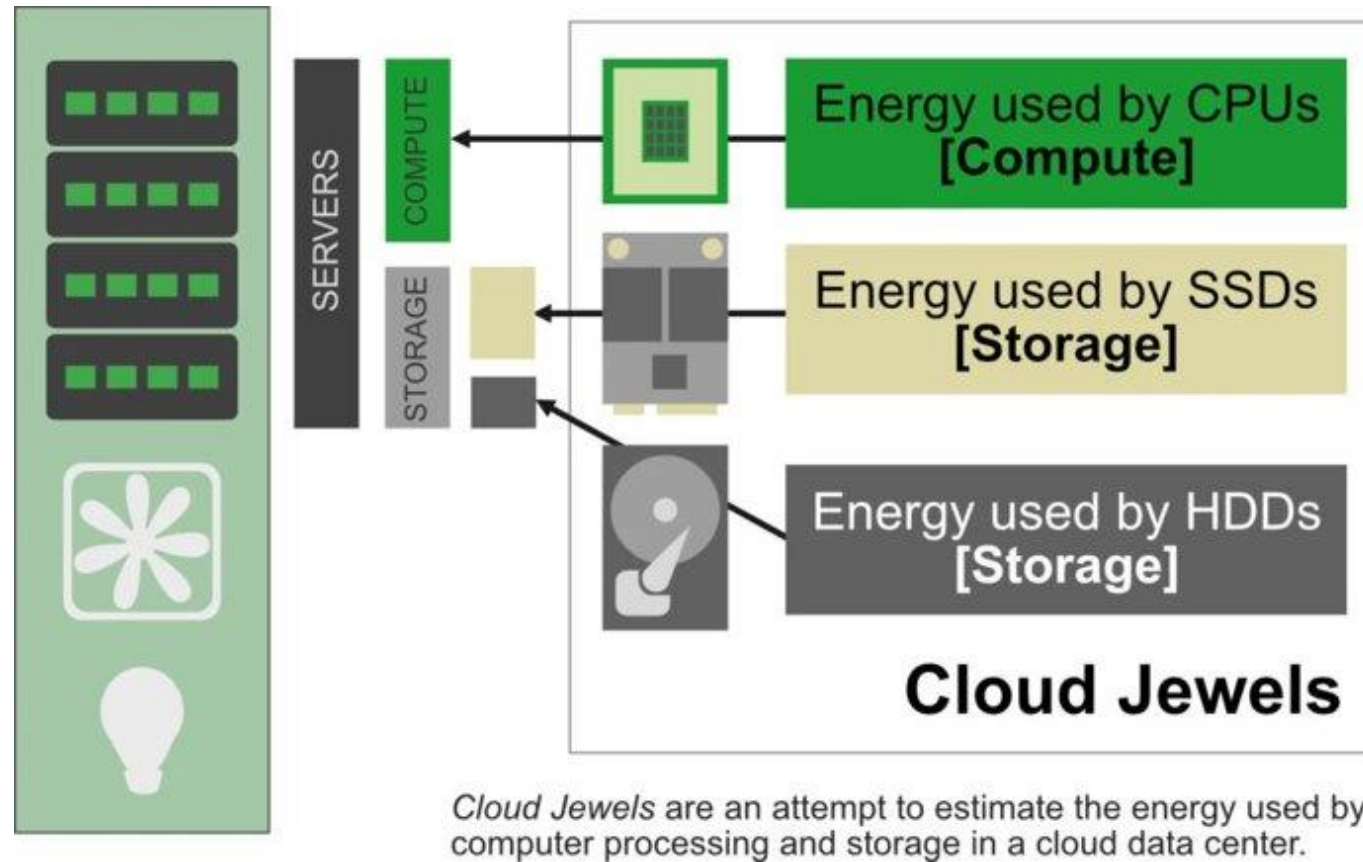


Network
Load

...

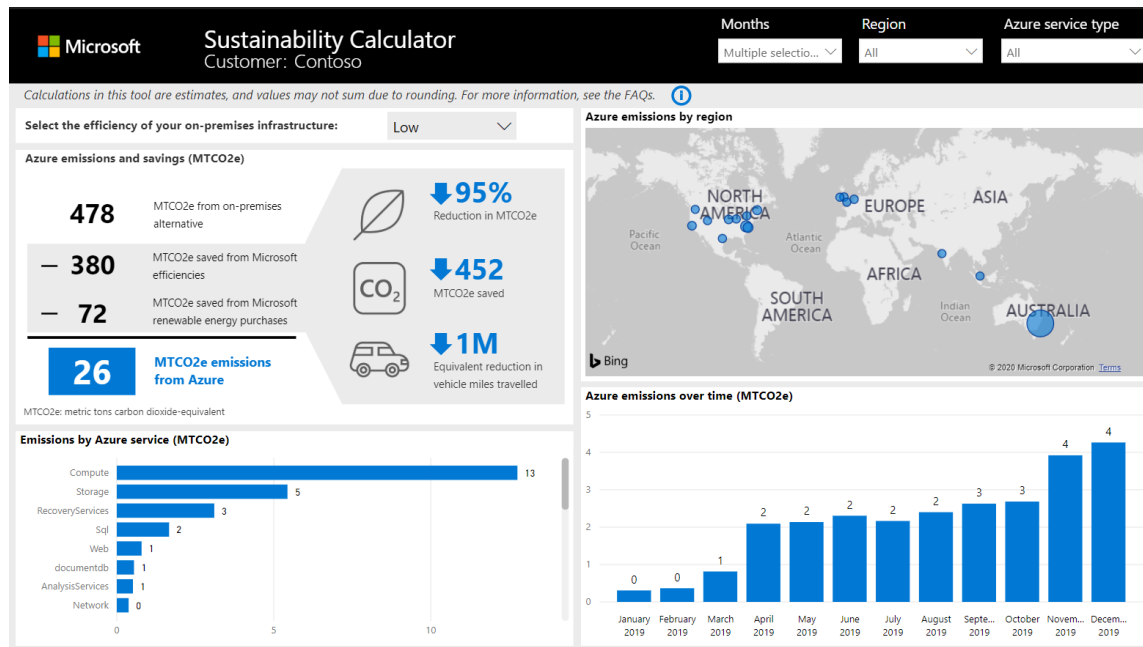
Metrics – Estimating Energy Usage in the Cloud

Etsy's Approximation Method – "Cloud Jewels"

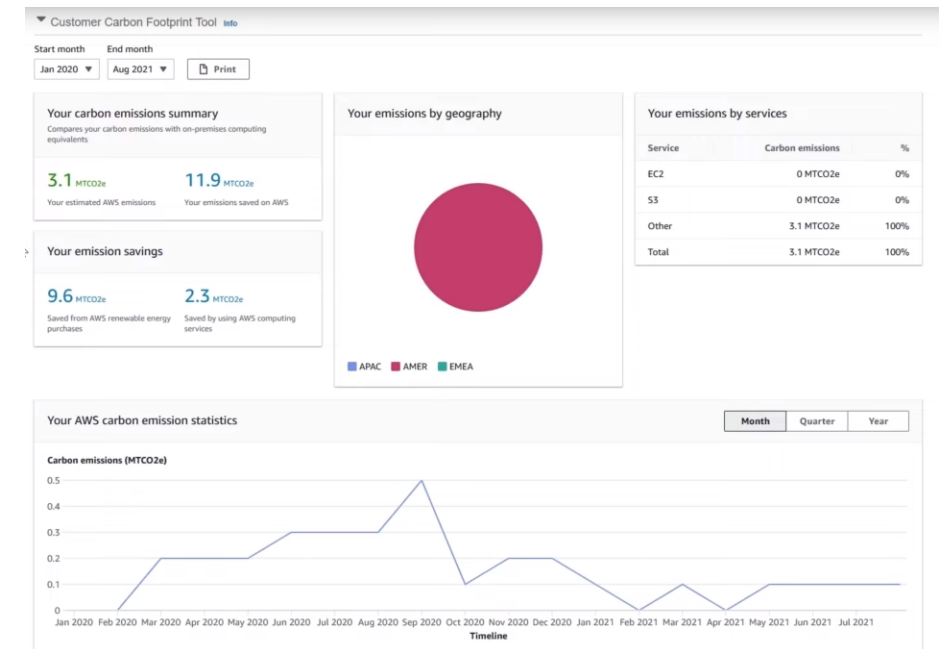


Metrics – Carbon Footprint Tools

- Some cloud providers offer tools to monitor carbon emissions
- Tools do not provide detailed data, only on service- and region-level
- Not suitable to identify components of high energy usage or for optimizations
→ **Carbon Proxies** are needed to provide more detailed data



Microsoft Sustainability Calculator



Amazon Customer Carbon Footprint Tool

Sustainability of Cloud Providers

Greenpeace Study 2017 – Clicking Clean



Categories used for the rating::

Transparency (20%); Renewable Energy Commitment & Siting Policy (20%);
Energy Efficiency & GHG Mitigation (10%); Renewable Energy Procurement (30%); Advocacy (20%)

Sustainability of Cloud Providers

How to choose the most sustainable cloud provider?

Comparison is difficult:

- No current data on greenhouse gases only emitted through cloud services
- Sustainability reports only include data on overall company

Idea: PUE = Power Usage Effectiveness

PUE = total energy usage of data center / energy usage by IT systems

Metrics – Energy Efficiency of Data Centers

- Close to 1,0 indicates a good data center efficiency
- Currently the **only international metric** to compare the efficiency of data centers

Critique:

- IT systems might be highly energy efficient, while other building components are not
→ results in high PUE

Average PUE metrics:

- **Microsoft Azure: 1.18** (first published in April 2022^[1])
- **Google Cloud: 1.10** (regular publication ^[2])
- **AWS: 1.135** (no publication, approximation by CCF^[3])

PUE	DCiE	Level of Efficiency
3.0	33%	Very Inefficient
2.5	40%	Inefficient
2.0	50%	Average
1.5	67%	Efficient
1.2	83%	Very Efficient

[1]: <https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/>



[2]: <https://www.google.com/about/datacenters/efficiency/>

[3]: <https://www.cloudcarbonfootprint.org/docs/methodology/#power-usage-effectiveness>

Green Cloud Computing

- Energy Consumption by Data Centers
- Metrics & Sustainability of Cloud Providers
- **Resource Utilization in the Cloud**
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Cloud Native Software Development
- Rebound Effects

Service Models – Overview

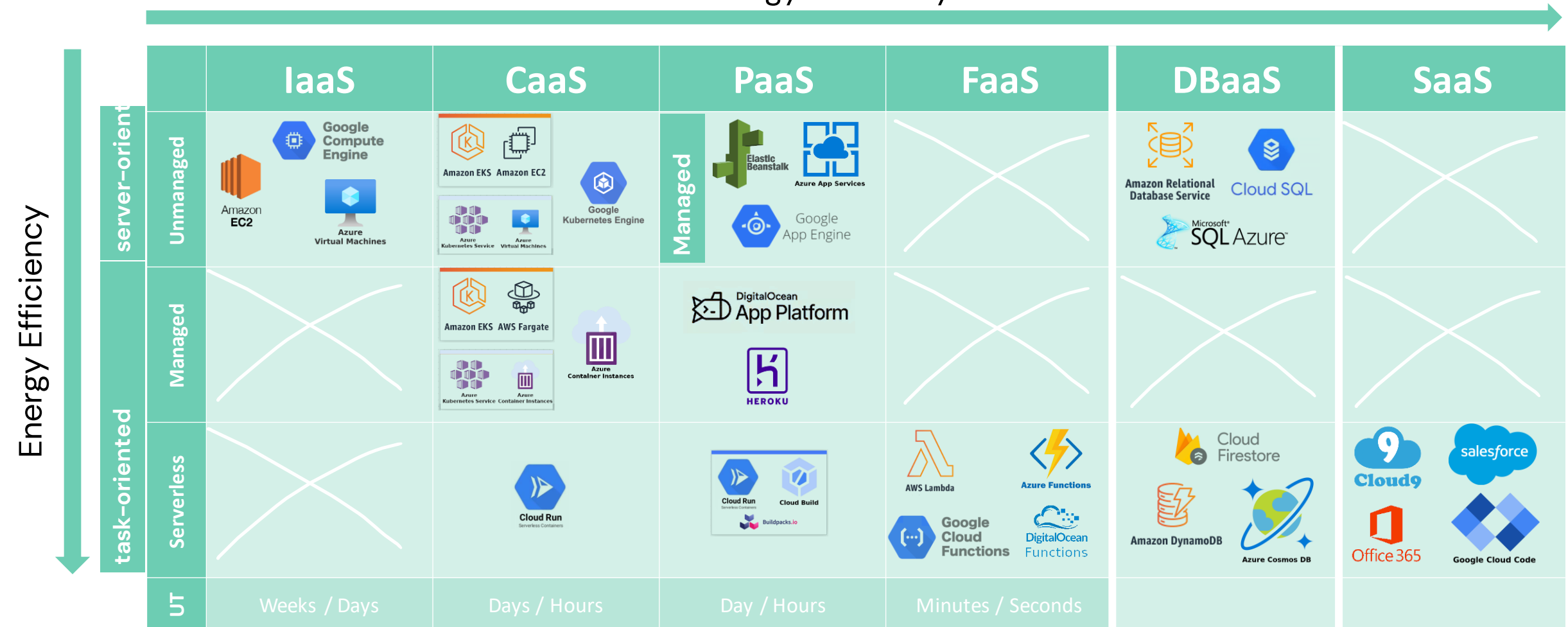
		Energy Efficiency 		
Energy Efficiency 		Unmanaged	Managed	Serverless
		Server-oriented		Task-oriented
	Single-Tenant	bare-metal / VM on dedicated host		dedicated resources
	Multi-Tenant	VM on shared host		shared resources

- **Server-oriented:** servers as basic unit
- **Task-oriented:** processes, containers, functions as basic unit

Multi-Tenancy enables the cloud provider to use resources more efficiently!

Service Models – Overview

Energy Efficiency



Service Models – Functions as a Service (FaaS)

- Uses the **serverless** operating model
- Serverless = management & scaling of the infrastructure are "hidden"
- Deployment of **functions** in the cloud, which are executed on demand
- Resources are dynamically allocated as demand increases and decreases

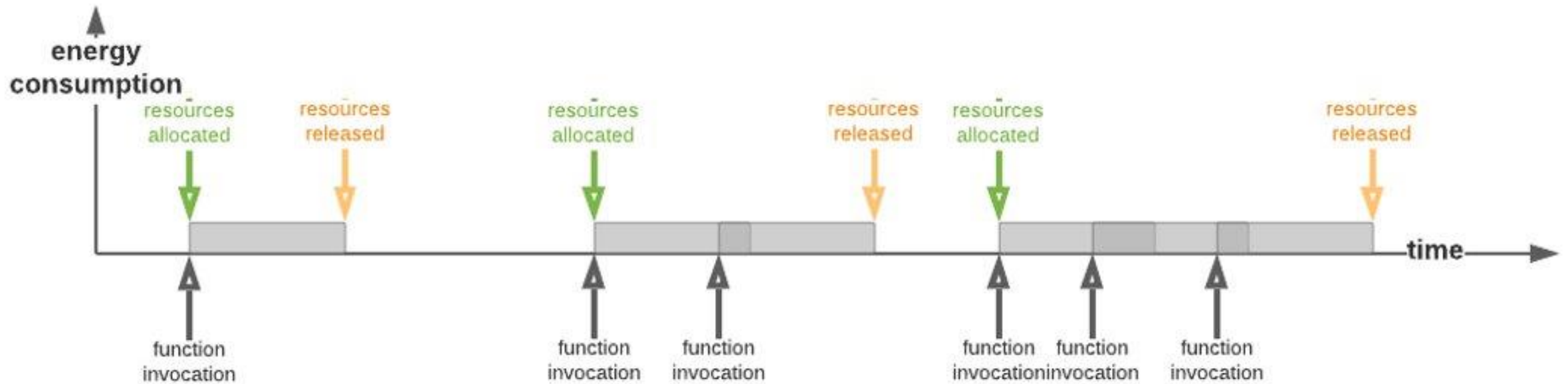
Examples:

- AWS Lambda, Google Functions, Azure Functions

Aspects of energy efficiency:

- Resources are used to fit; no over-provisioning or under-provisioning
- Different utilization of functions can be taken into account

Service Models – Functions as a Service (FaaS)

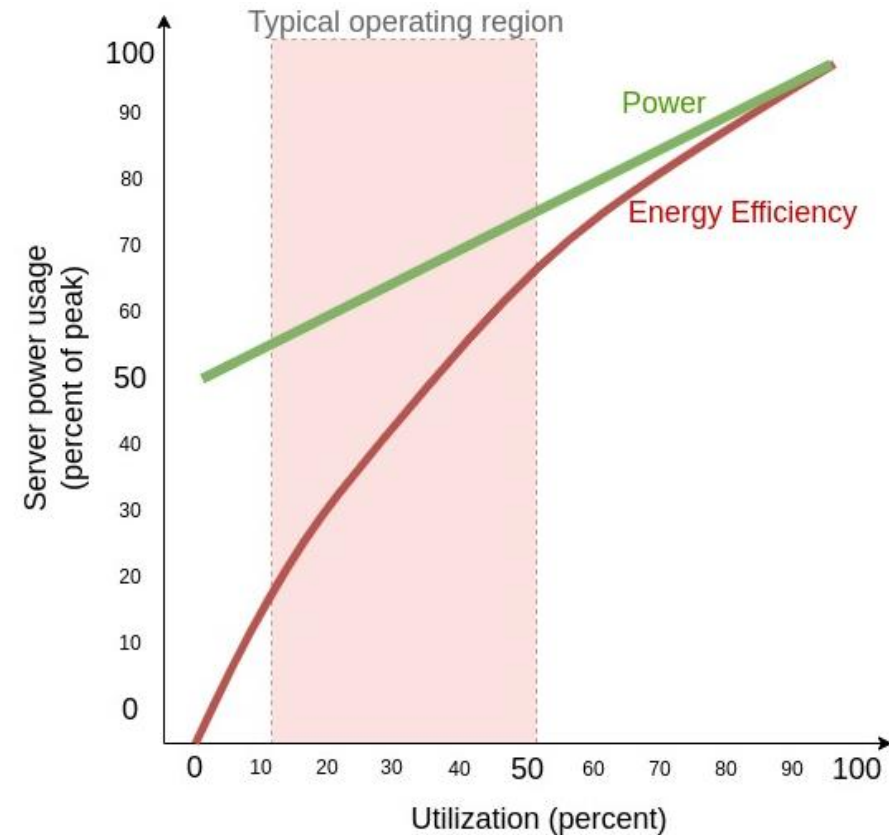


High Resource Utilizations

Why should servers be utilized as much as possible?

An unused server doesn't consume any electricity, does it?

- Depending on the server, already 50% of power are used without any workload
- Energy efficiency increases with increasing utilization of the server



L. A. Barroso and U. Hözl, "The Case for Energy-Proportional Computing," in *Computer*, vol. 40, no. 12, pp. 33–37, Dec. 2007, doi: 10.1109/MC.2007.443.

High Resource Utilizations

Why should virtual machines also be utilized as much as possible?

A virtual unit itself does not consume any power, does it?

- VMs consume very little power, depending on the size of the server and the hypervisor
- Resources can be reserved for potential VMs by the hypervisor
- Poor efficiency when few VMs are provisioned on a hypervisor

→ Cloud providers recommend **stopping unused VMs** to be able to use the resources on the same hypervisor for VMs of other customers

Service Models – Overview

FaaS, DBaaS, SaaS:

- For service models with a high degree of management, **optimizations are limited**
- Cloud providers need to ensure a high energy efficiency
- Energy efficiency is in the financial interest of the cloud provider

IaaS, CaaS, PaaS:

- More freedom, but **risk of non-optimal operation** is higher
- Own responsibility for energy-efficient operation

We will focus on IaaS & PaaS for the rest of this lecture!

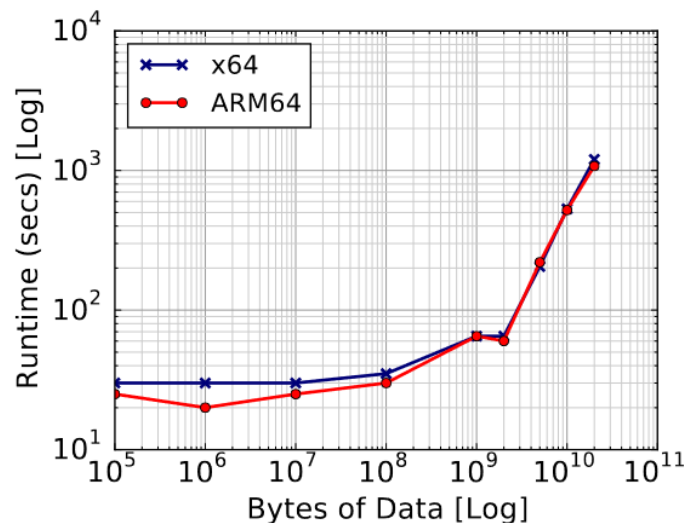
Green Cloud Computing

- Energy Consumption by Data Centers
- Metrics & Sustainability of Cloud Providers
- Resource Utilization in Cloud
- **Optimization measures for IaaS and PaaS**
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Rebound Effects

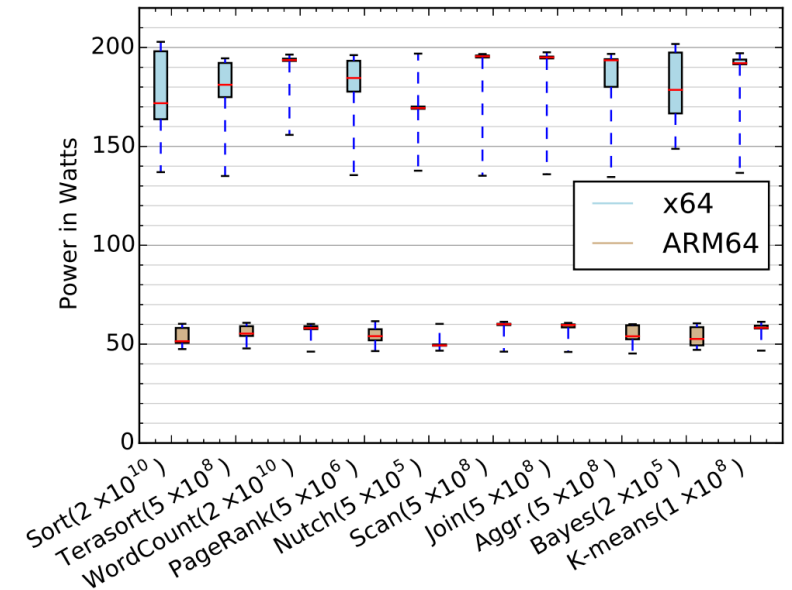
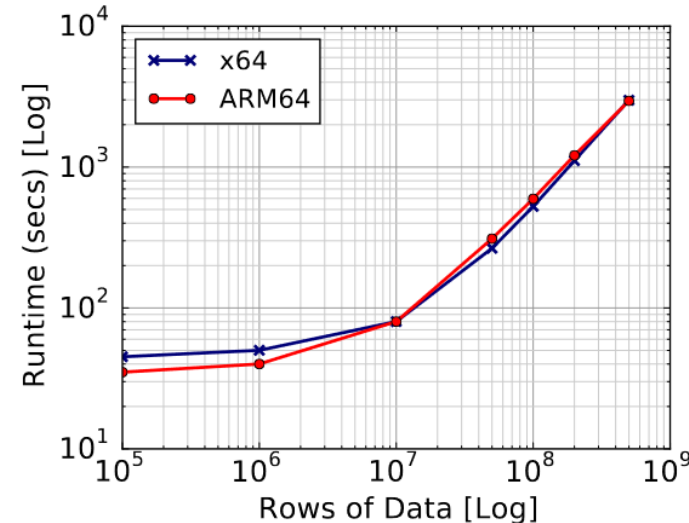
Resource Selection – CPU

Selecting a CPU:

- Architecture of the CPU is an important factor
- **ARM** based CPUs often much more energy efficient than **x86/64** alternatives
- Performance for many application areas comparable
- Recompilation might be necessary as many applications were developed on x64



Runtime of workloads x64 vs ARM64



Power consumption x64 vs. ARM64

Resource Selection – Locality

- Cloud regions vary significantly in terms of carbon emissions
- Google offers the **Region Picker** to take into account carbon footprint, price, and latency
- Region Picker does not take energy mix into account
 - Nuclear power plants are considered "low carbon"

Google Cloud Region Picker

This tool helps you pick a Google Cloud region considering carbon footprint, price and latency.

Optimize for

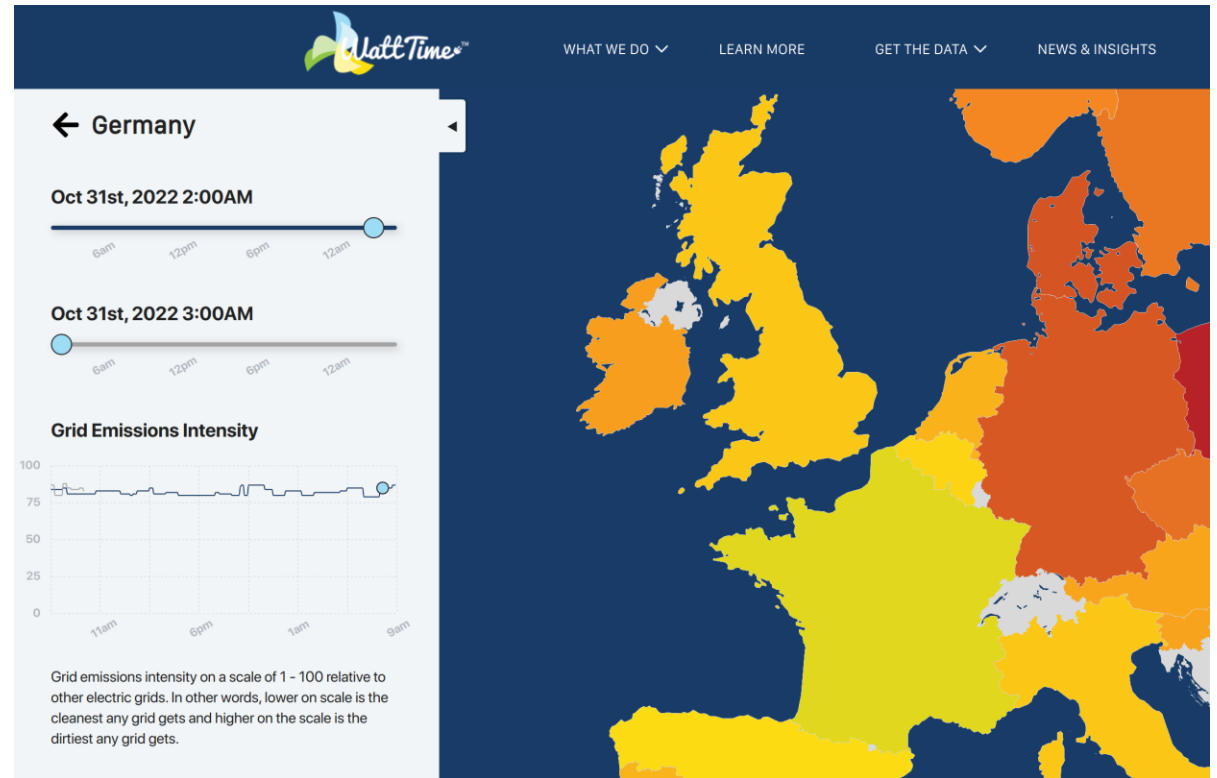
- 🌿 Lower carbon footprint ⓘ
Not important Important
- \$ Lower price ⓘ
Not important Important
- 🕒 Lower latency ⓘ
Not important Important

Where is your traffic coming from?

Your current location
Afghanistan
Albania
Algeria
American Samoa

Resource Selection – Locality

- **Watttime** as an alternative to Google Region Picker
- Watttime provides data on power plant emissions by using **measurements from space**
- Only useful if supplying power plant is known

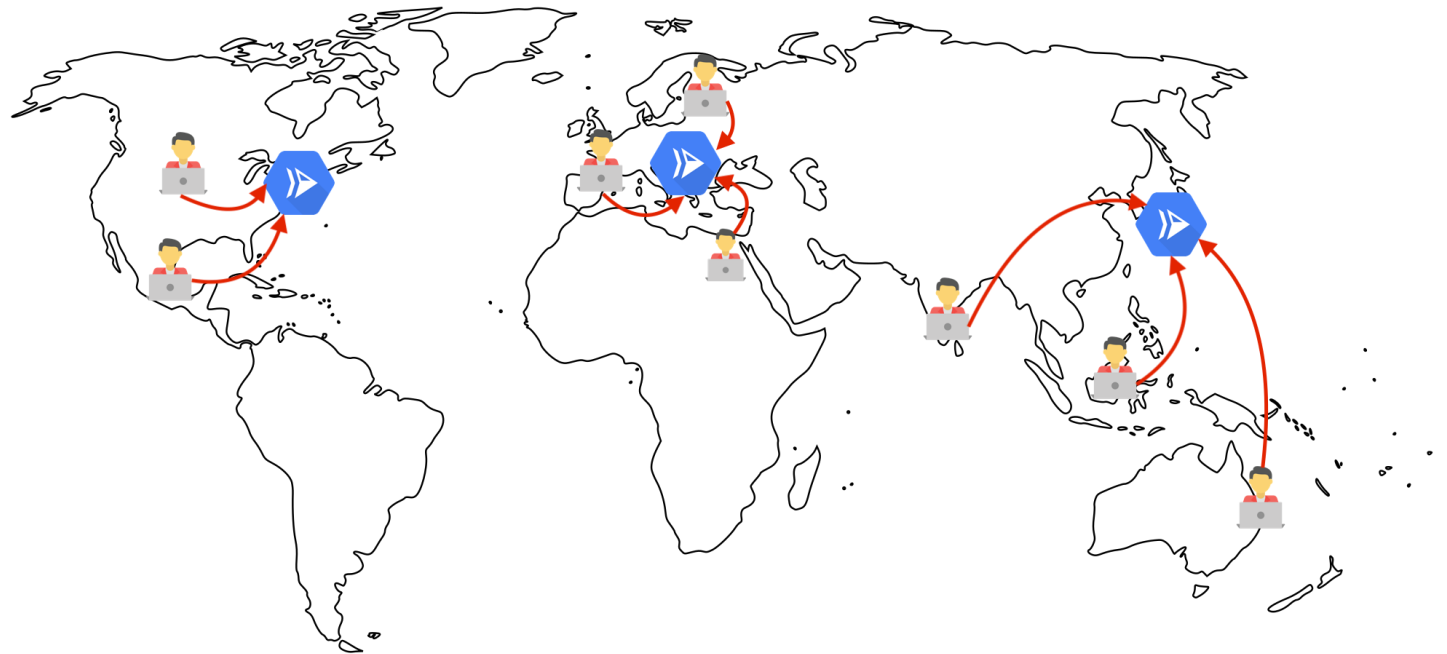


Resource Selection – Locality

- Cloud region also influences energy consumption through **network transmission**
- Energy consumption through network transmission can be optimized by reducing data volume and distance
- Customers often come from different world regions

Solution: Edge deployments

→ deployment of services on different local data centers



Resource Selection – Locality

Advantages:

- less network traffic
- potential coverage of energy demand by renewable energies
- reduced latencies

Disadvantages:

- data duplication if data has to be available worldwide

- **Placement groups** or context awareness should be considered
- Services that communicate a lot with each other should always be operated geographically close

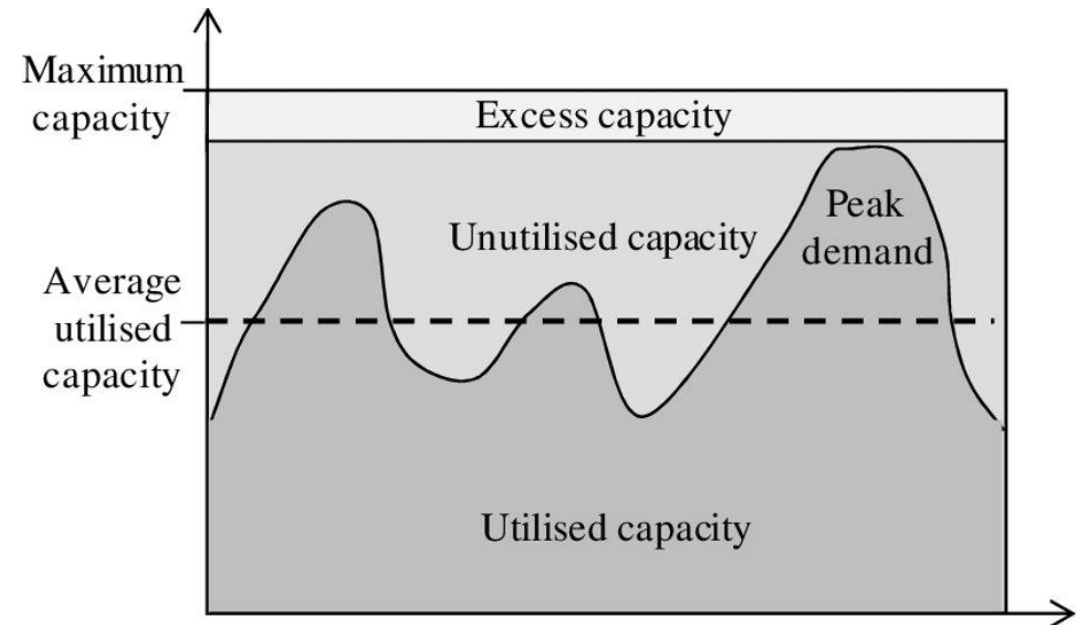
Overprovisioning

Anti-pattern when operating in the Cloud: **Overprovisioning**

- Permanent allocation of resources in order to be able to serve **peak loads**
- On average, resource allocation exceeds actual demand

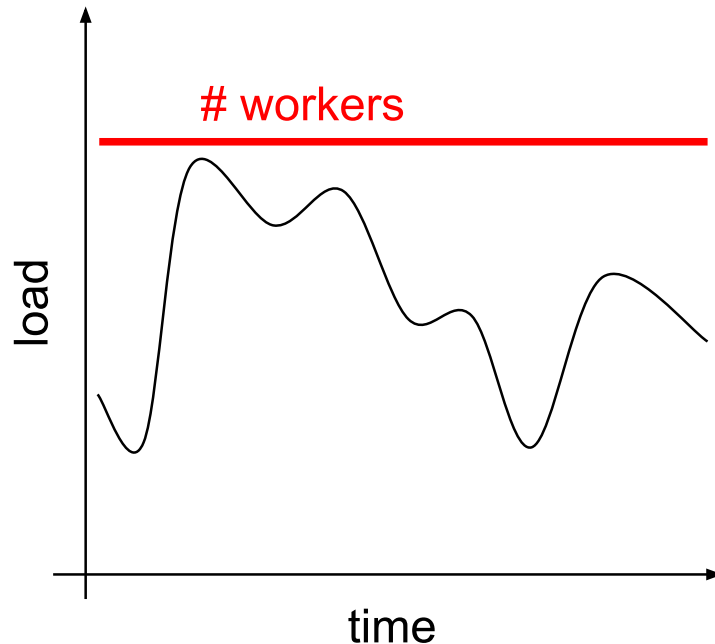
Examples:

- Provision online shop for peak loads in Christmas season
- Provision for execution of scheduled jobs



Overprovisioning

- Overprovisioning often occurs when **static or reactive scaling** is used



Static Scaling

No elasticity → overprovisioning needed to handle peak loads



Reactive Scaling

Resources are not provided fast enough → overprovisioning needed

Scaling Strategies

- Different scaling strategies to avoid overprovisioning and to save resources

Pro-Active /
On-Prediction

Random /
On-Coincidence

Demand Shifting /
On-Availability

Demand Shaping /
On-Availability

Scaling Strategies – Pro-Active

- Pro-active provisioning of resources
- Demand-driven scaling **before** actual demand is present (“On-Prediction”)
- Counteracts overprovisioning that occurs due to excessive startup times of VMs or other instances

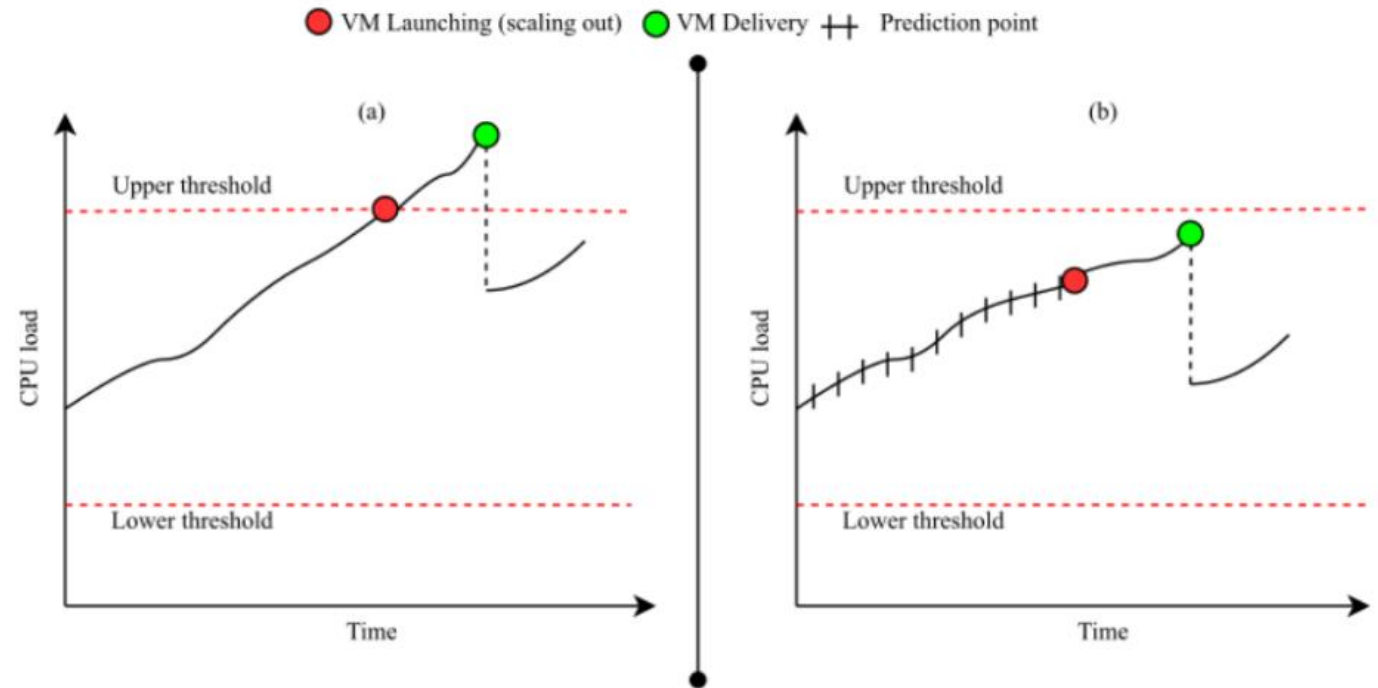


Fig. 1. Elasticity approaches: (a) reactive; (b) proactive.

Scaling Strategies – Random

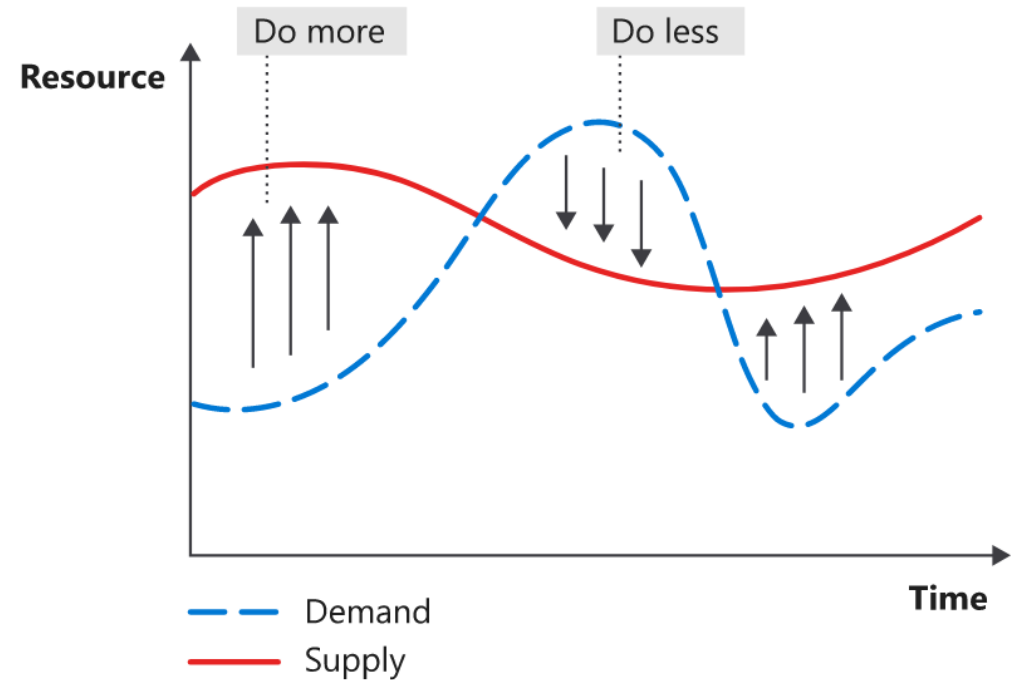
- **Random provisioning** of resources to equalize peak loads
- Regular workloads will be started at random times to better distribute the load on the system
- Only possible for workloads without user interaction

Examples:

- Event-Streaming applications
- Creating a database backup that would otherwise always run at 12 a.m.

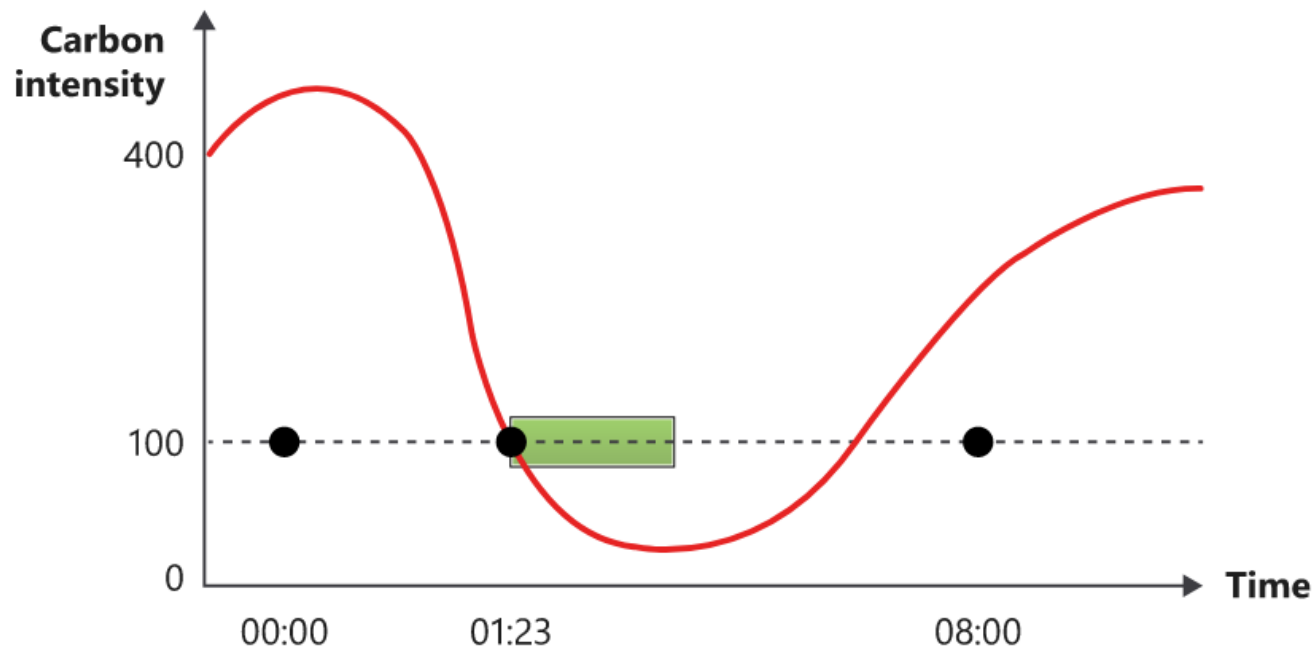
Scaling Strategies – Demand Shaping

- "On-availability" scaling shapes demand according to **available supply**
- Strategy for flexible workloads without time constraints
- Adjusts the provisioned resources to available resources
 - Examples for constrained resources: **CPU capacity, renewable energies ...**
- Workloads can be matched to free capacities of the cloud provider with **on-demand, spot and preemptible instances**



Scaling Strategies – Demand Shifting

- Time-flexible workloads are shifted to **times or regions** where they can be executed with lower carbon emissions



- Alternatively, execution is shifted according to other criteria
 - Times or regions where unused cloud provider resources are available
 - Times or regions where own unused resources are available

Efficient Workload Distribution

- Services change continuously, so distribution must also be adjusted to prevent over-provisioning or under-provisioning
- Orchestrating software is responsible for optimal distribution
- **Orchestrator** searches for the appropriate VM for each service

But:

- No more redistribution of services at runtime
- Inefficient distribution of services and thus waste of resources

Solution:

- System that **dynamically** redistributes workloads

Efficient Workload Distribution

Bin packing algorithm for dynamic redistribution of workloads:

- Bin packing deals with distributing **"items"** to a finite number of **"bins"**
- Goal: find the smallest number of "bins" for the available "items"
- Optimization problem of bin packing is **NP-hard**
- Approximation algorithms solve the problem in acceptable runtimes

Offline approach:

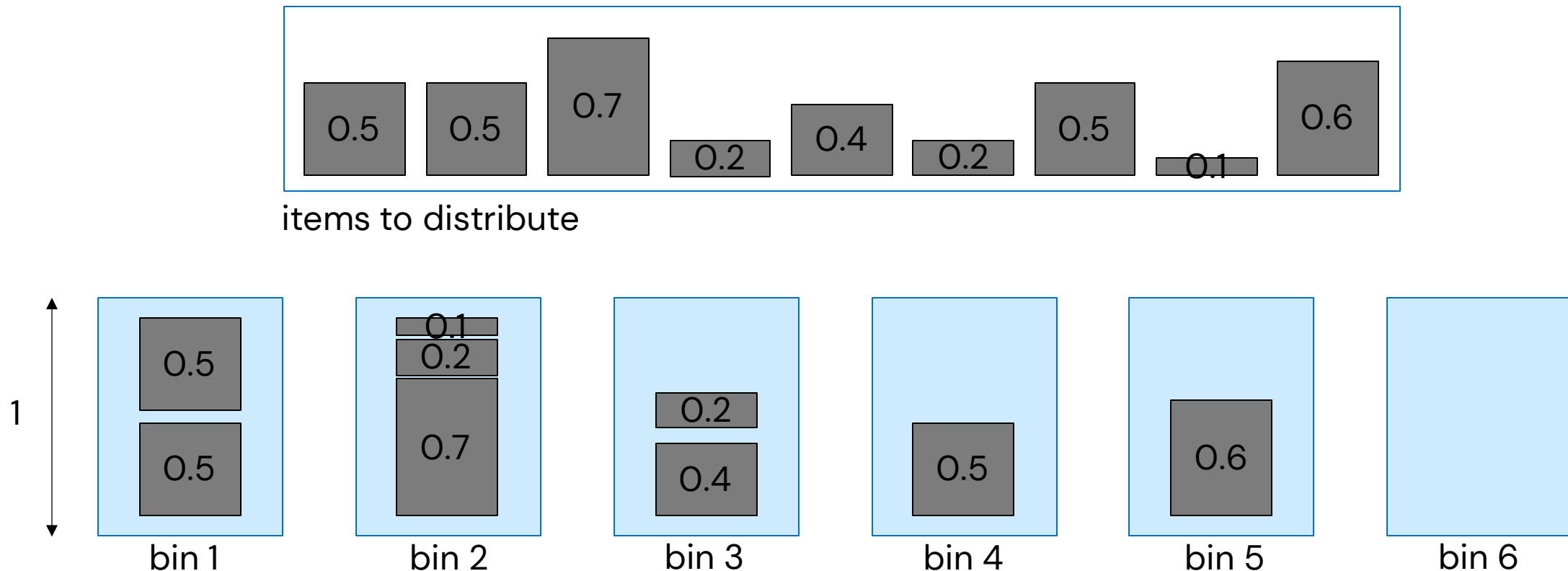
- E.g., "first-fit-decreasing bin packing"
- "Items" are dynamically included in the calculation and can be rearranged

Online approach:

- Incoming "items" are allocated to bins in the best possible way
- Supported by Kubernetes [1]
- "Items" that have already been placed are not rearranged in the process

Dynamic Workload Distribution

„First-Fit“ Bin Packing Algorithm (Online):



Green Cloud Computing

- Energy Consumption by Data Centers
- Metrics & Sustainability of Cloud Providers
- Resource Utilization in the Cloud
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- **Cloud Native Software Development**
- Rebound Effects

Cloud Native Software Development

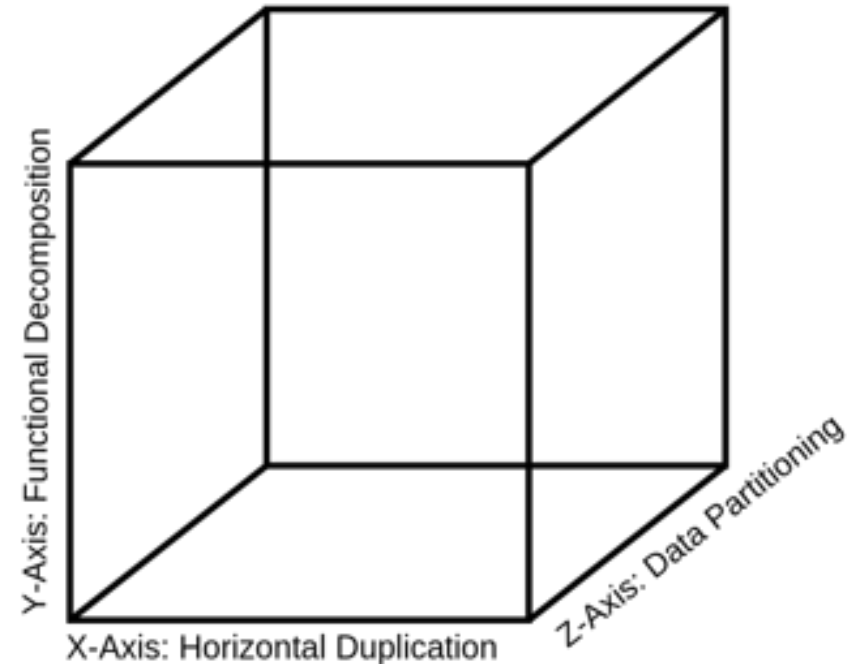
- Cloud-native software development deals with the specialized development of applications in the cloud
- Applications are designed to be...
 - highly scalable,
 - resilient,
 - flexible
- Flexibility of the cloud infrastructure is important
- Cloud development offers many advantages over solutions on-premises in terms of energy efficiency

Advantages can only be used if applications are designed to be operated in the cloud!

Cloud Native Software Development

Prerequisites for taking advantage of all the benefits in terms of energy efficiency:

- **Fast startup times** for flexible scalability
- **Fast "graceful shutdowns"** to be able to shut down applications without data corruption
- **Available failover strategy** to get back online quickly
- Should be **stateless**
- Have good **scalability** according to the scale cube model



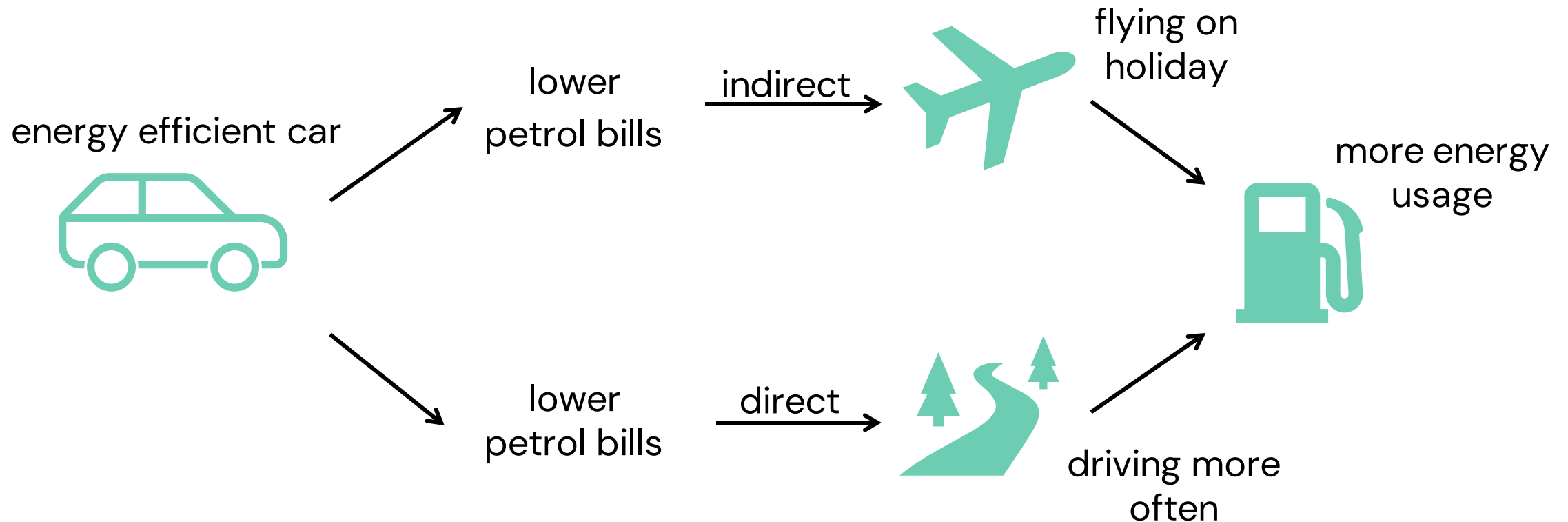
Applications should follow the **12 Factor Method** (<https://12factor.net/>)

Green Cloud Computing

- Energy Consumption by Data Centers
- Metrics & Sustainability of Cloud Providers
- Resource Utilization in the Cloud
- Optimization Measures for IaaS and PaaS
 - Resource Selection – CPU & Location
 - Scaling Strategies
 - Efficient Workload Distribution
- Cloud Native Software Development
- **Rebound Effects**

Rebound Effects

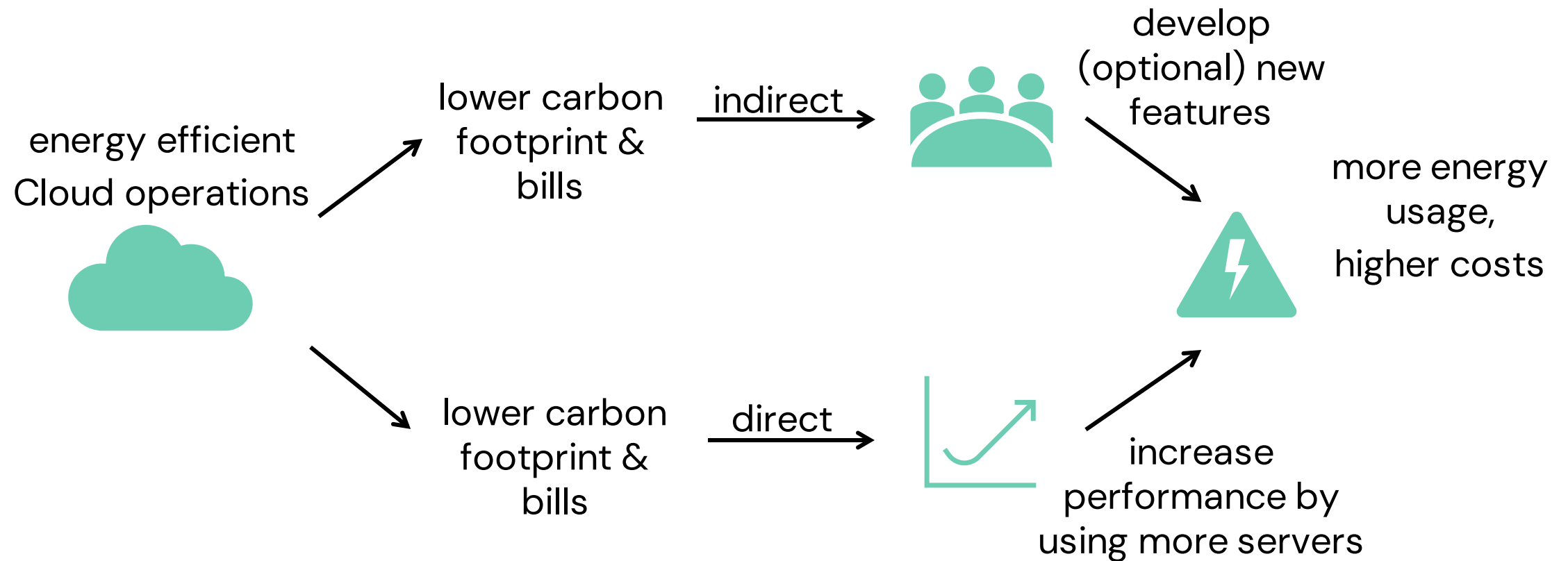
Rebound Effect = energy savings lead to changed behavior and increased energy usage



Adelmeyer, M.; Walterbusch, M.; Biermansk, P.; Seifert, K.; Teuteberg, F. (2017): Rebound Effects in Cloud Computing: Towards a Conceptual Framework, in Leimeister, J.M.; Brenner, W. (Hrsg.): Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, S. 499–513

Rebound Effects

- Cloud migration and optimizations lead to energy and costs savings
- **Risk:** savings encourage changed behavior and lead to increased energy usage



Green Cloud Computing

Can you think of any optimizations that could be made to your project from this semester to improve its energy efficiency?

Questions?

Thank you!

Uwe Eisele

uwe.eisele@envite.de

Nadja Hagen

nadja.hagen@envite.de

envite consulting GmbH

www.envite.de

