neuland

ecoCompute

Anita Schüttler

Making Spot Instances work for Sustainability

Anita Schüttler

Software Engineer
- >15 years at neuland
- >10 years in large scale e-commerce systems

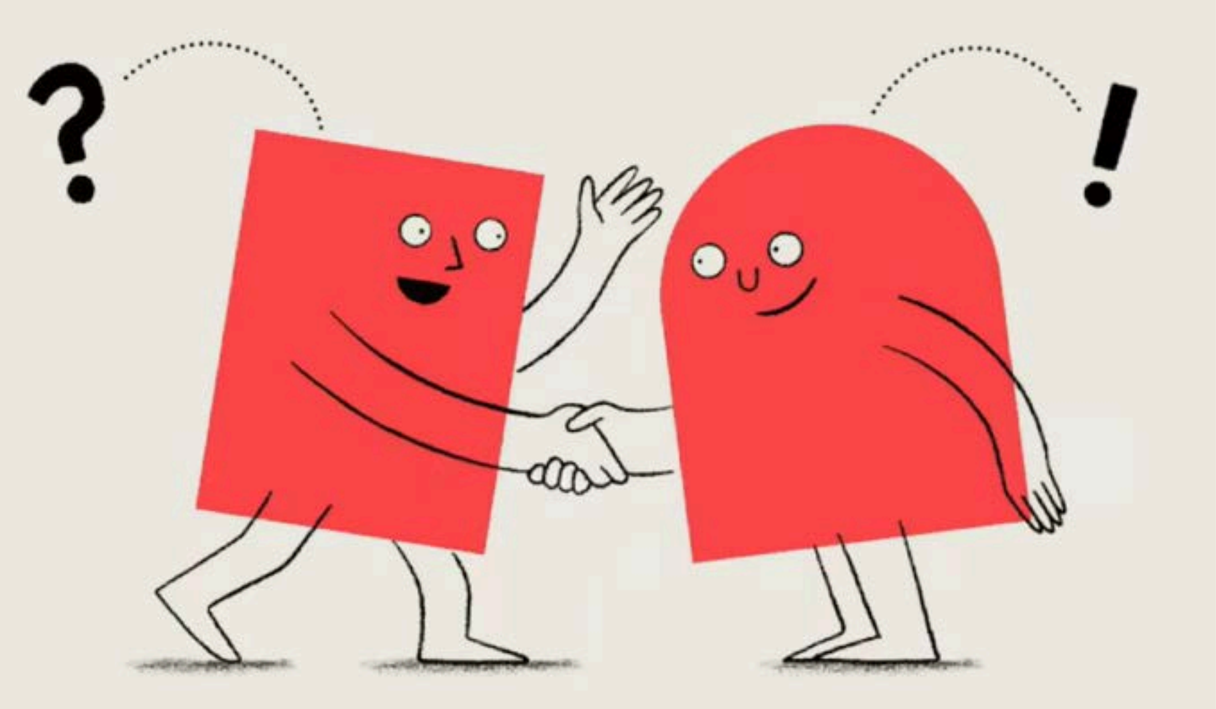
Sustainability Professional
- Sustainability in IT, Green Software
- Circular Economy

Co-Chair Bundesverband Green Software
Auditor for the Blue Angel for Software
Champion of the Green Software Foundation

neuland

custom software and consulting
large scale e-commerce
sustainability and longevity
long-time collaborations

https://www.neuland-bfi.de/



Bundesverband
green.software

INTERSPORT breuninger bonprix. DOUGLAS OTTO OBI dm
Höffner BLUME2000 FRANKONIA MANUFACTUM. helline FRICKE

Agenda

01

The case for using what’s tried and tested

02

What exactly are spot instances and why are they a GreenOps tool?

03

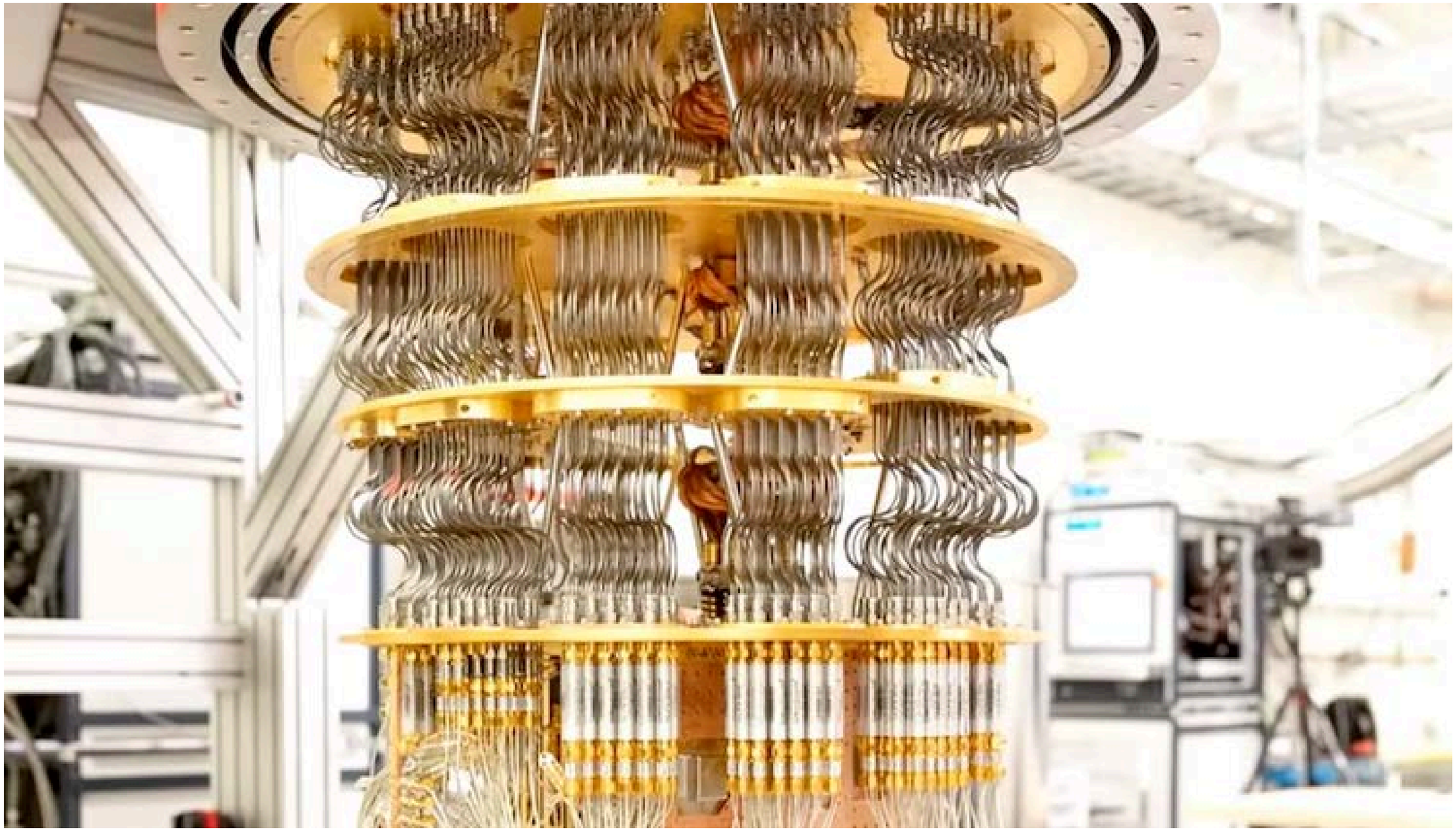
The tricky bits and how to get them right

04

How can spot instances help advance the energy transition?

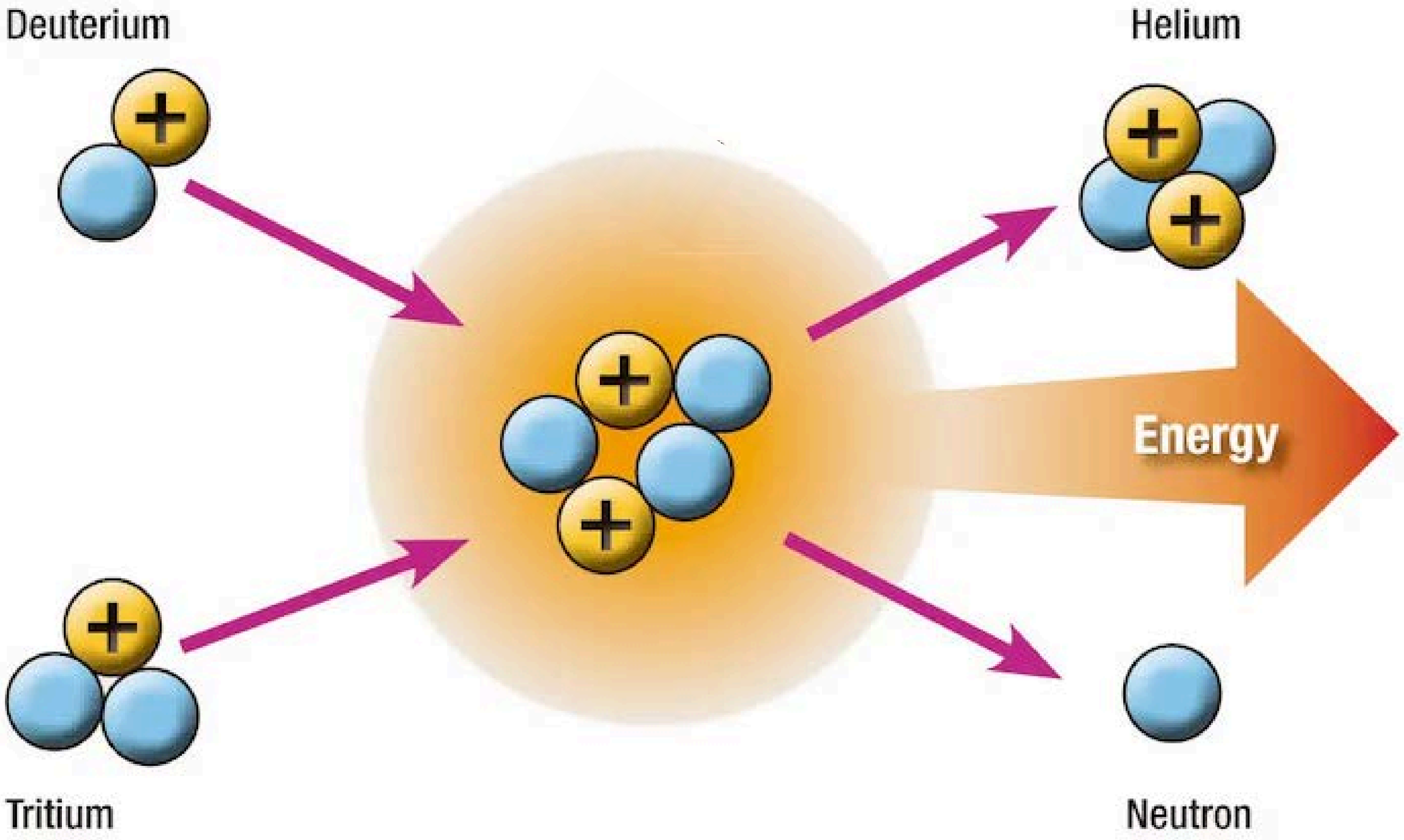
01
The case for using what’s tried and tested

Let’s start with a riddle:
guess what’s on these images!

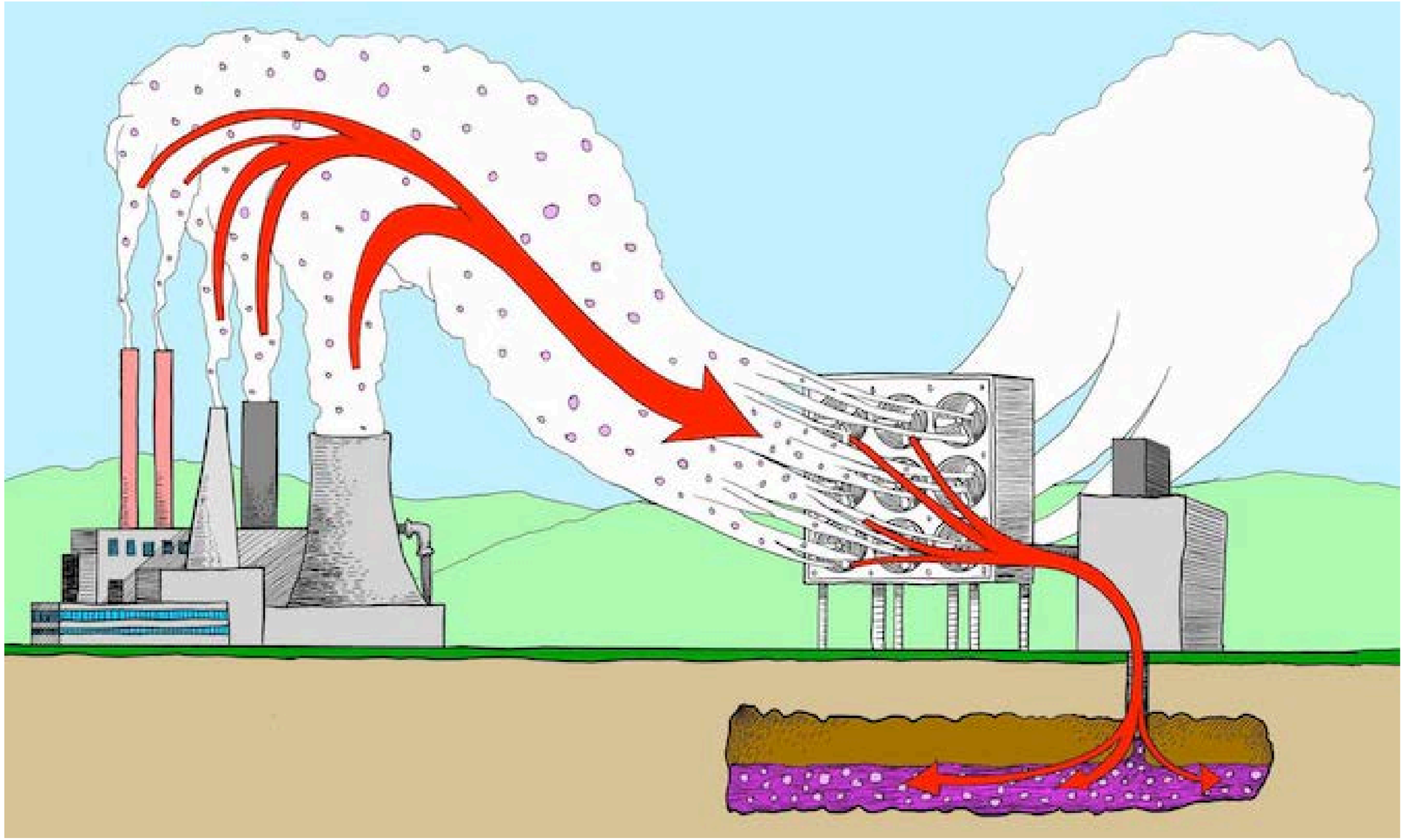


source: Ecosia image search

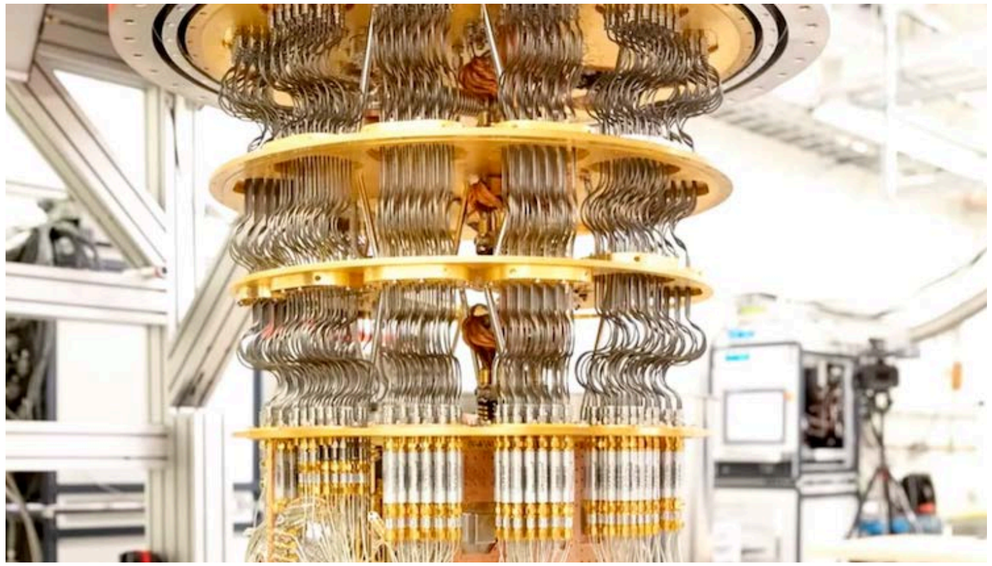
source: Ecosia image search

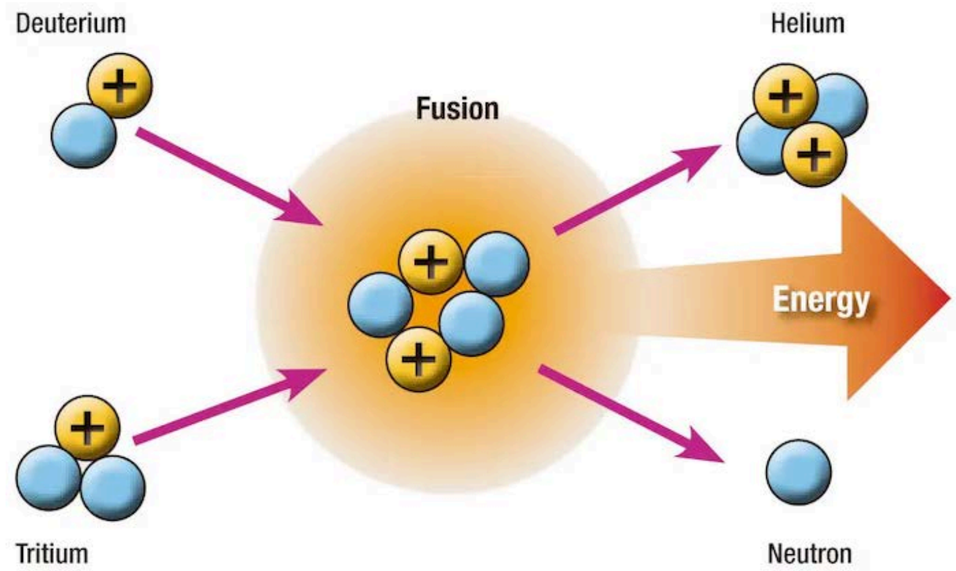


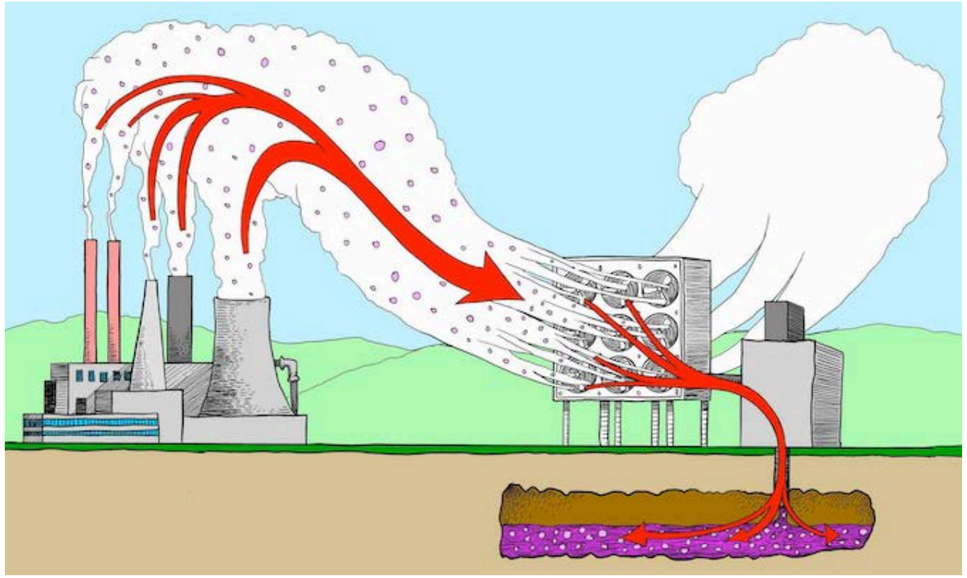
source: Ecosia image search



source: Ecosia image search







source: Ecosia image search

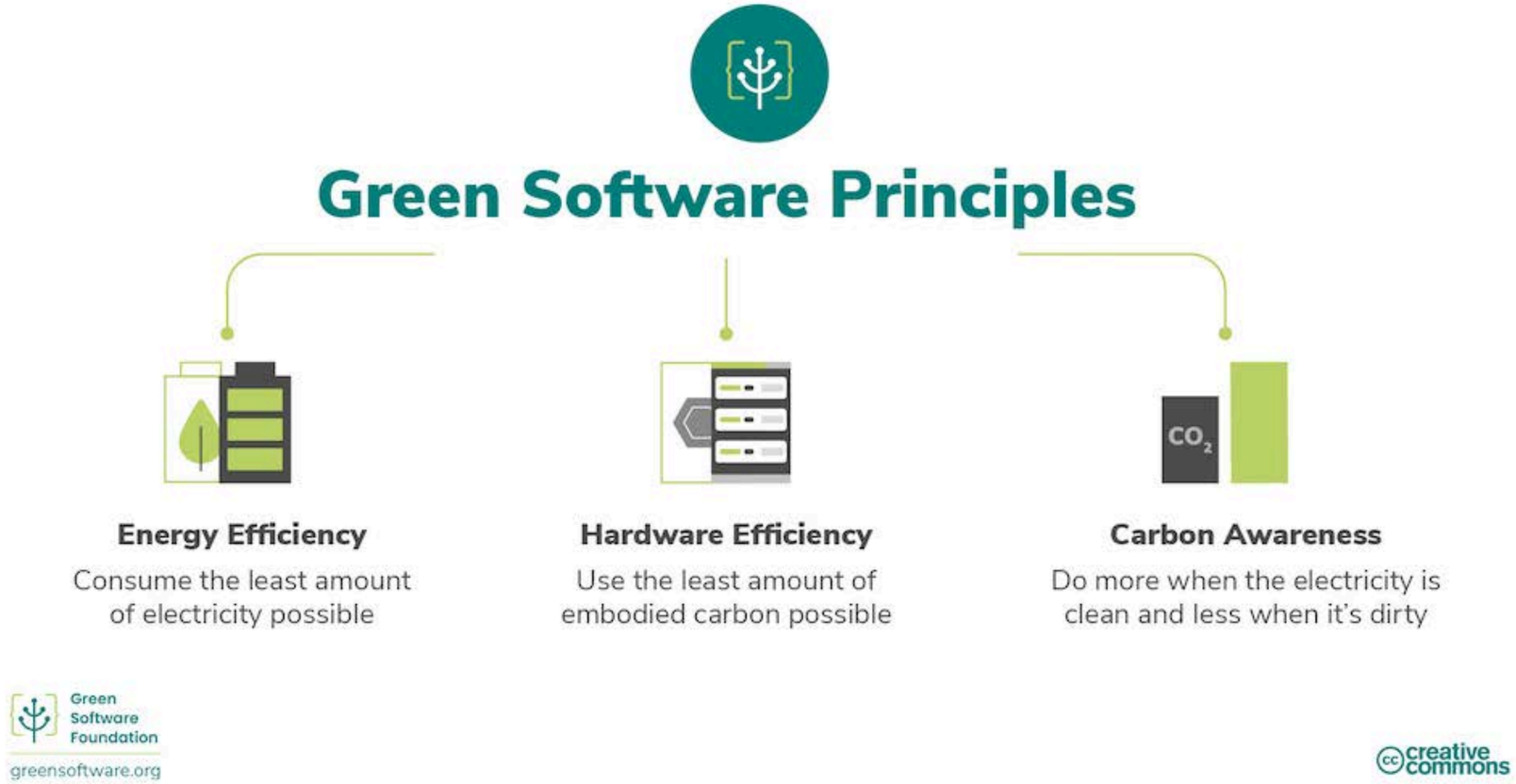
Technology that is promised to solve the climate crisis - once it’s working properly.

- tech bros

source: https://www.vecteezy.com

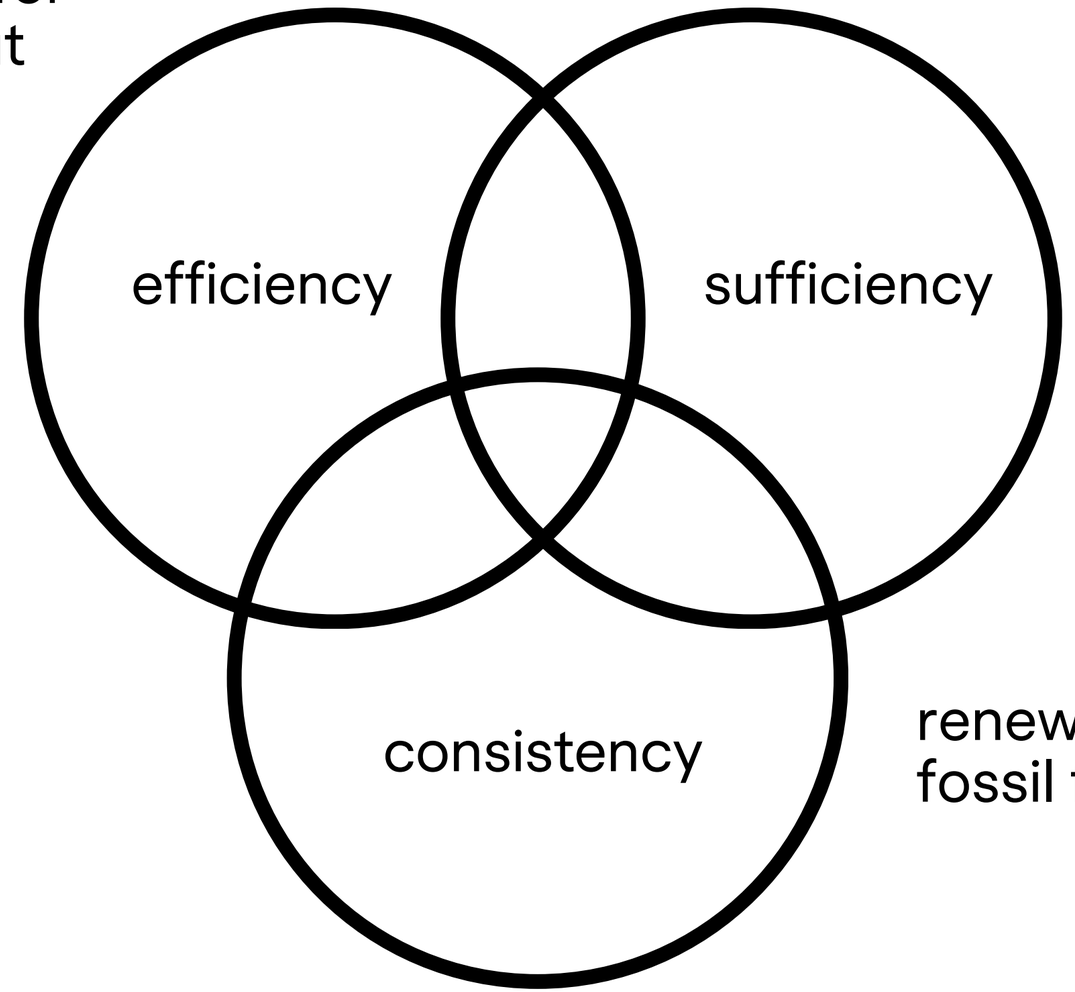
“Once the tools are mature enough,
we can really get going...”

It’s about saving carbon, water, resources...



less resources for
the same output

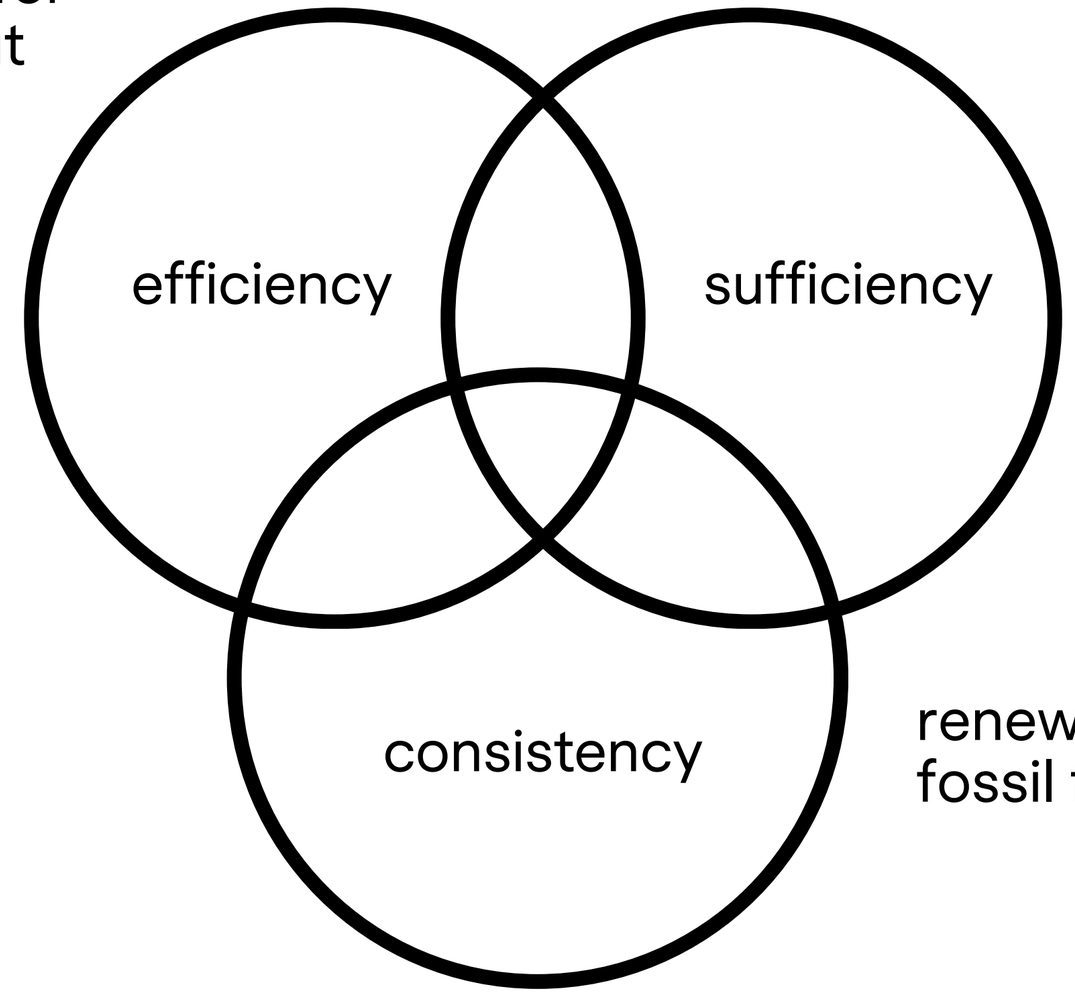
less consumption
overall



renewables instead of
fossil fuels, circular economy

less resources for
the same output

rebound effect /
Jevons paradox

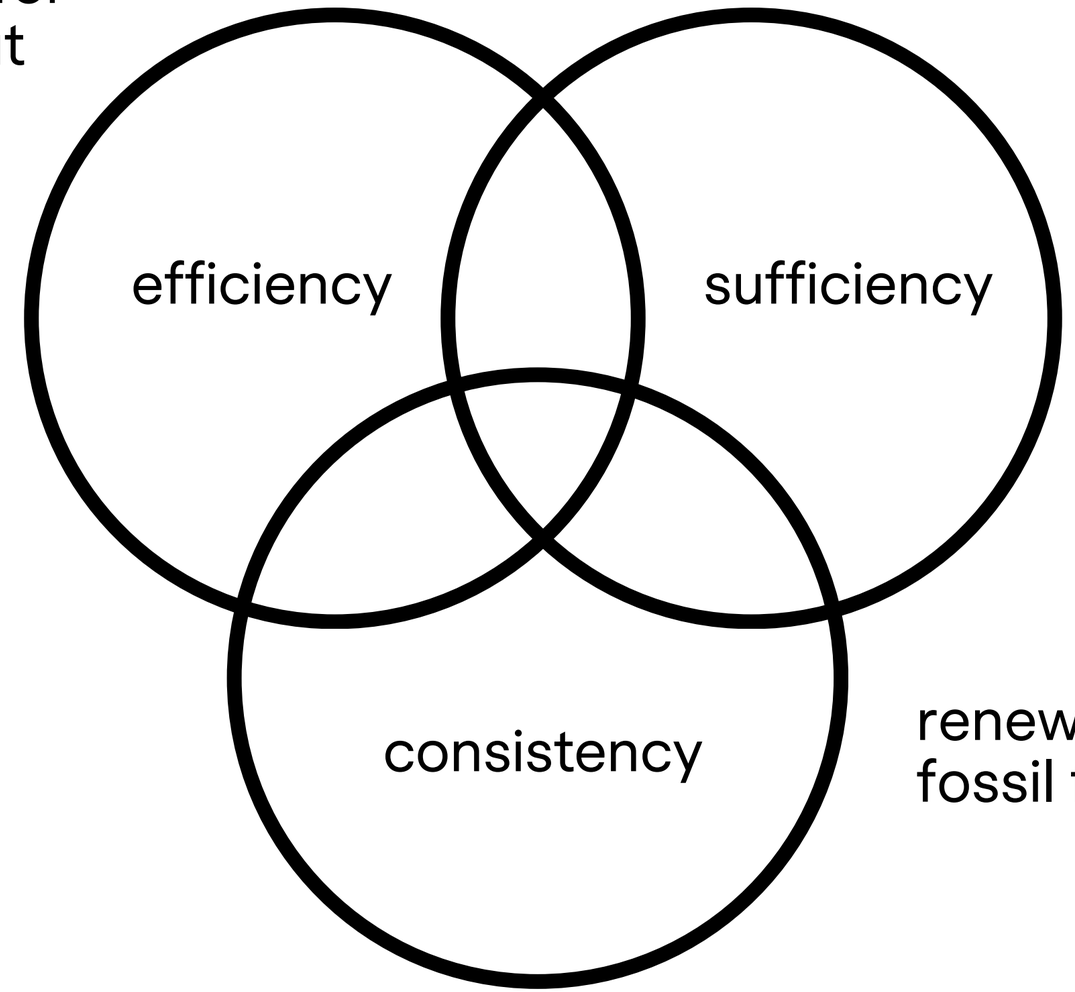


less consumption
overall

renewables instead of
fossil fuels, circular economy

less resources for
the same output

rebound effect /
Jevons paradox

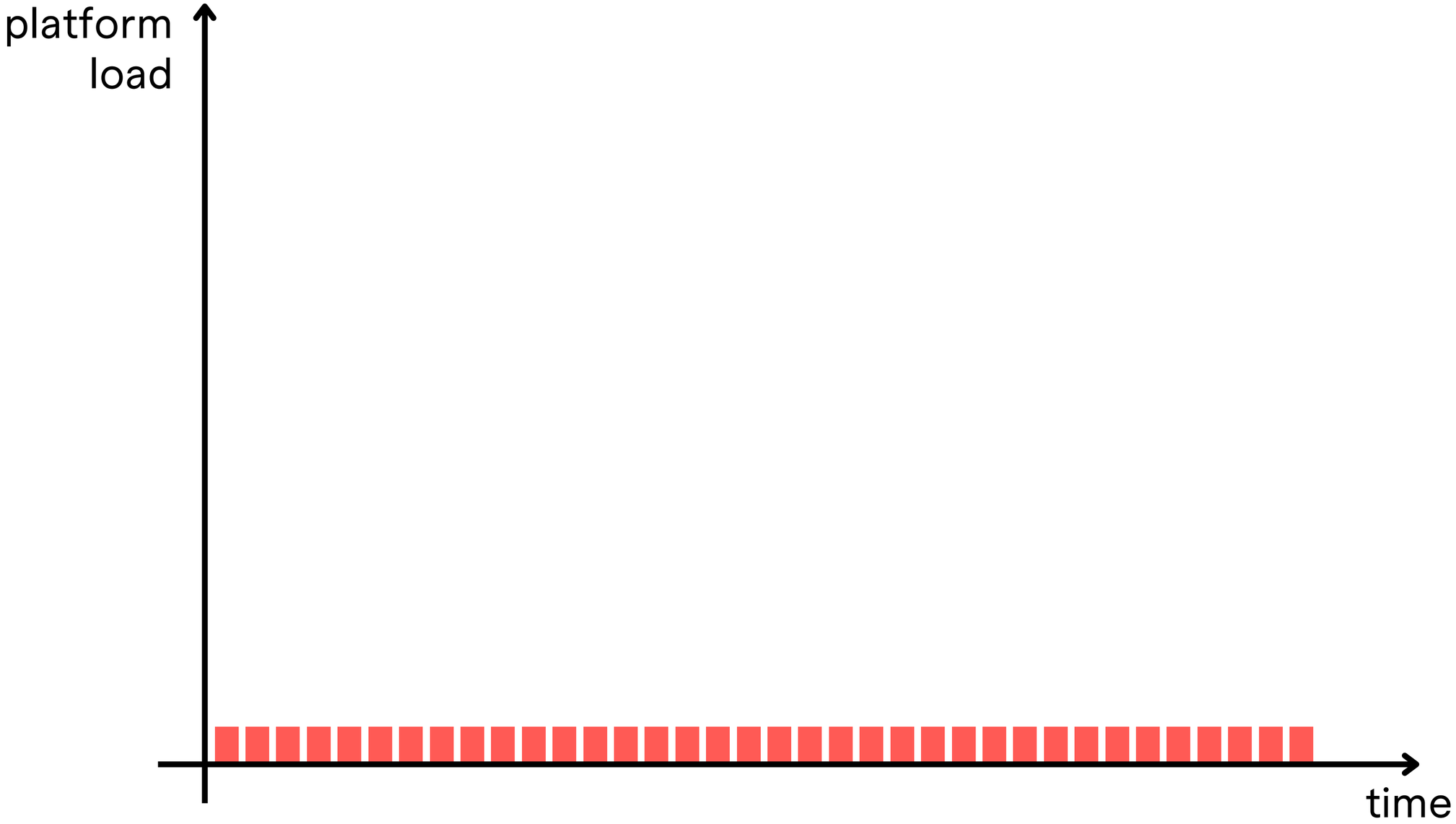


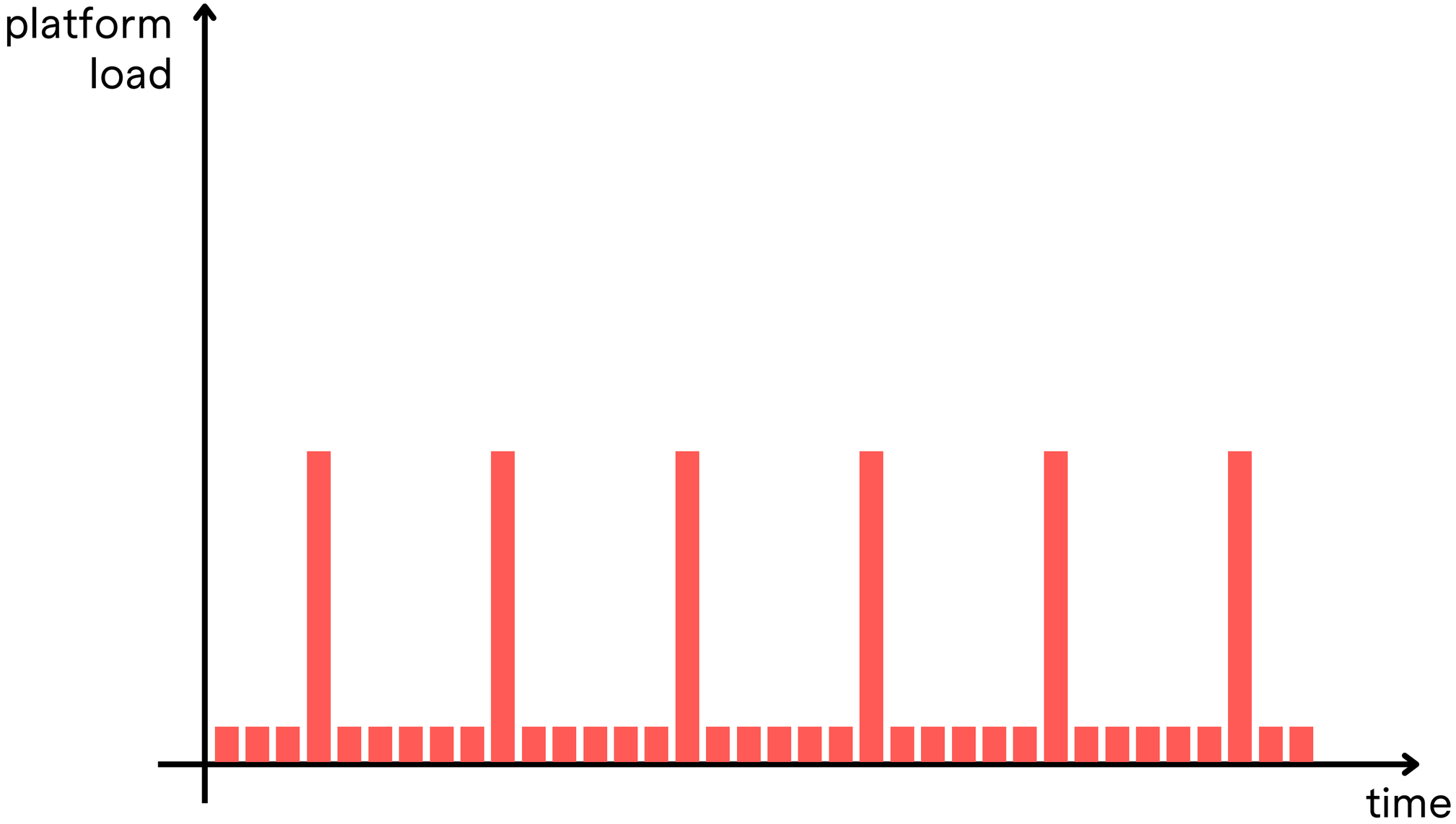
less consumption
overall

this currently has
the biggest effect

renewables instead of
fossil fuels, circular economy

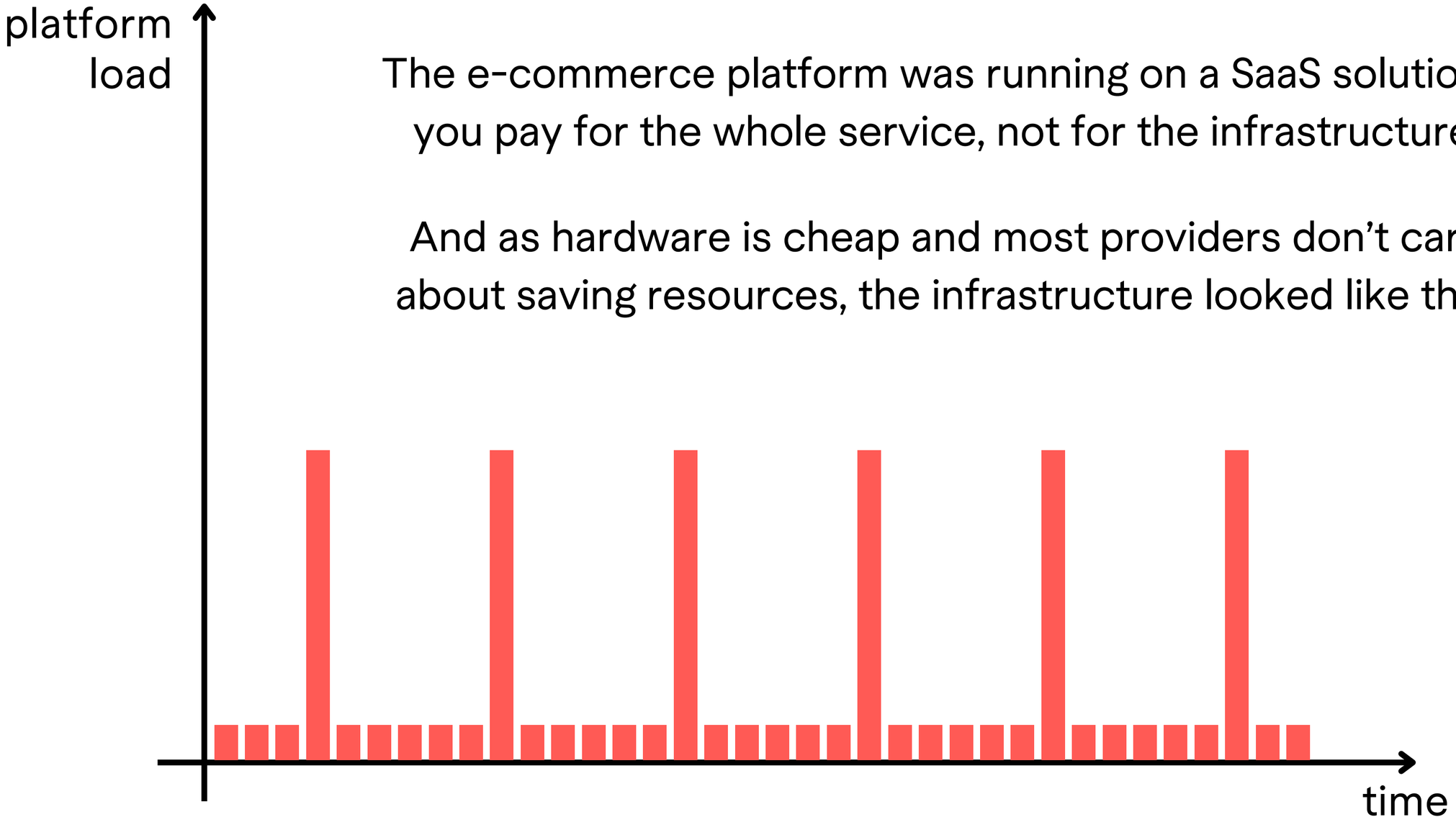
source: https://imgflip.com/i/375i6v

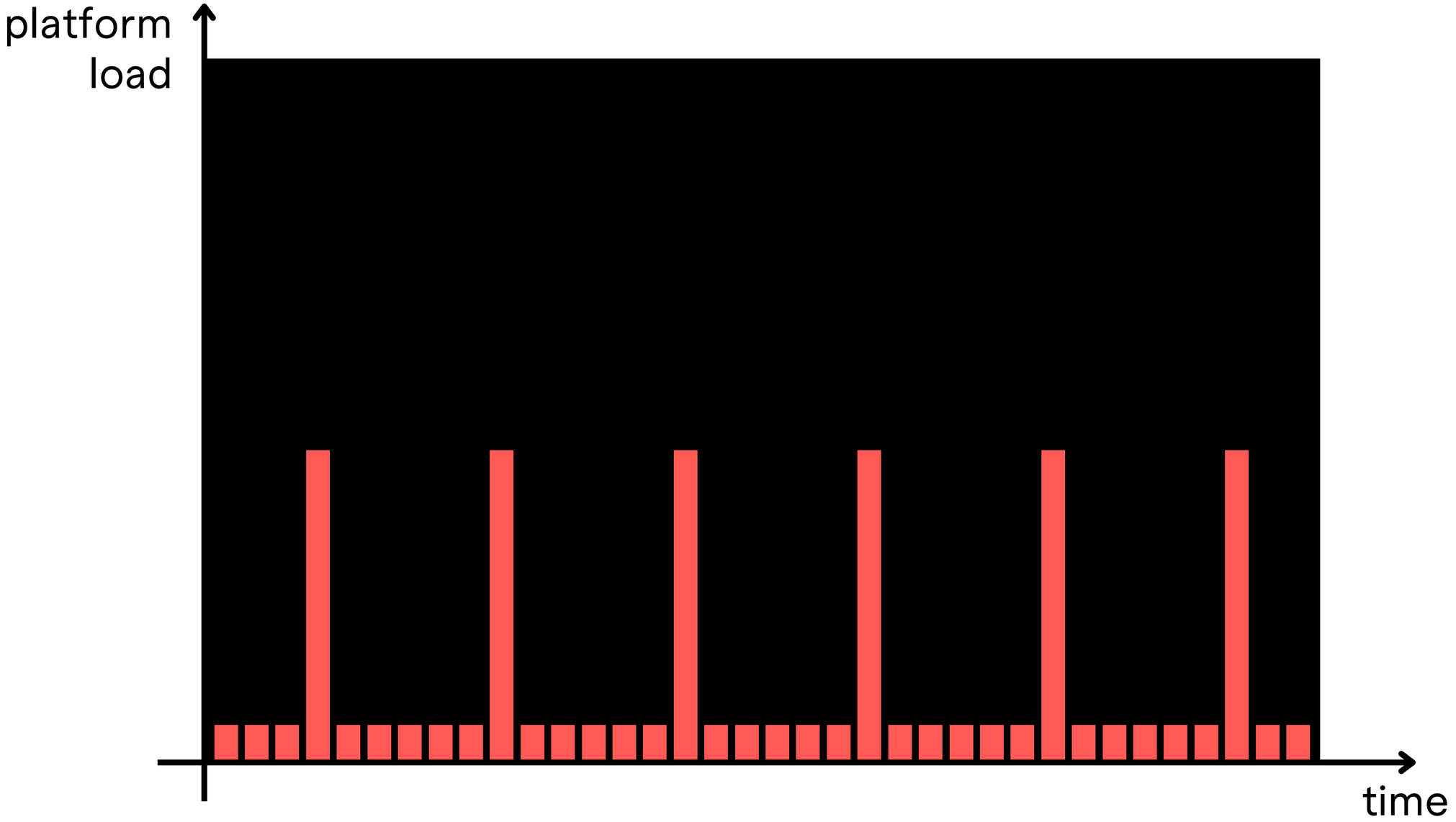

platform
load
time


platform
load
time

The e-commerce platform was running on a SaaS solution - you pay for the whole service, not for the infrastructure.

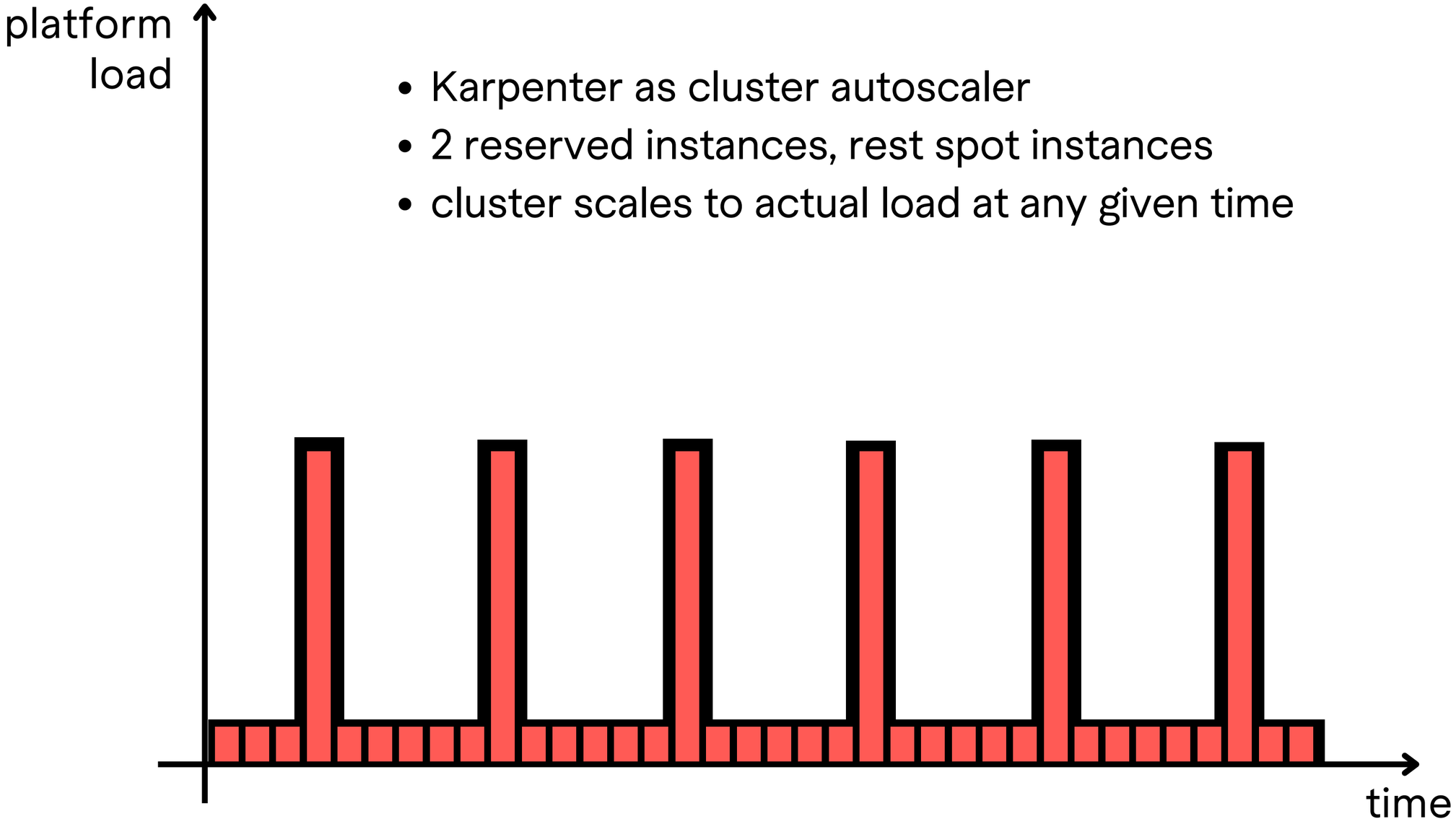
And as hardware is cheap and most providers don’t care about saving resources, the infrastructure looked like this:

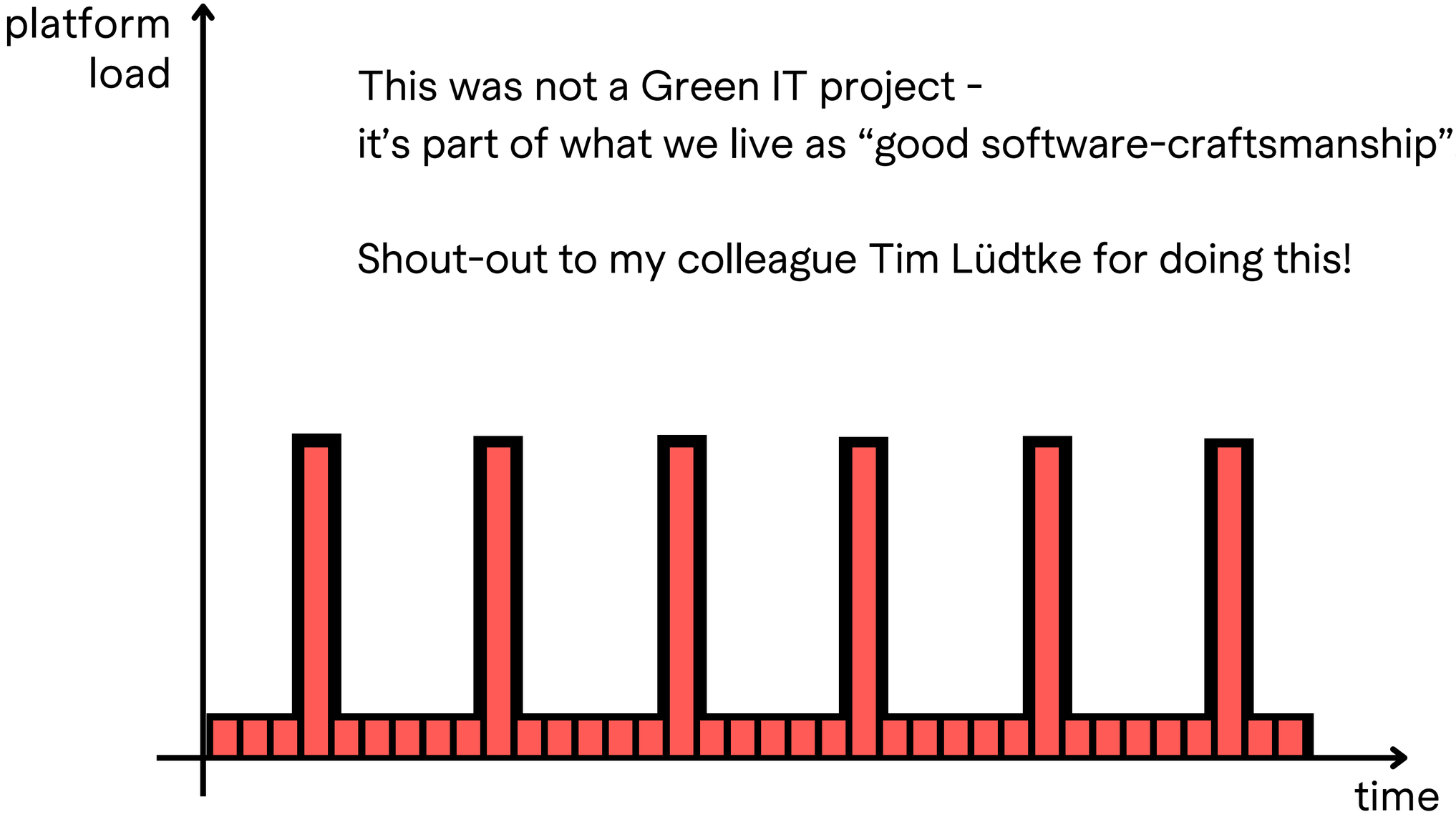



platform
load
time

- way smaller instances
- queries cleaned up + caching
- speed increase

- Karpenter as cluster autoscaler
- 2 reserved instances, rest spot instances
- cluster scales to actual load at any given time





This was not a Green IT project -
it’s part of what we live as “good software-craftsmanship”.

Shout-out to my colleague Tim Lüdtke for doing this!

02
What exactly are spot instances and why are they a GreenOps tool?

- reserved instances
- savings plans
- **spot instances**

FinOps: cost

- reserved instances
- savings plans
- **spot instances**

the dirty “only saves money, not carbon” corner

FinOps: cost

- reserved instances
- savings plans
- **spot instances**

FinOps: cost

- green cloud location
- energy-efficient CPU architecture
- carbon aware computing

GreenOps: carbon

- reserved instances
- savings plans
- **spot instances**

FinOps: cost

- turning things off
- right-sizing
- autoscaling
- storage classes
- (auto-deleting) data
- SLA commitment reduction

both: reducing waste

- green cloud location
- energy-efficient CPU architecture
- carbon aware computing

GreenOps: carbon

- reserved instances
- savings plans
- spot instances

FinOps: cost

- turning things off
- right-sizing
- autoscaling
- storage classes
- (auto-deleting) data
- SLA commitment reduction
- **spot instances**

both: reducing waste

- green cloud location
- energy-efficient CPU architecture
- carbon aware computing

GreenOps: carbon

“Potentially, using spot instances could be one of the greenest ways to operate a system.”

Anne Currie, Sara Bergman, Sarah Hsu - “Building Green Software”

source: https://v1.maturitymodel.security.aws.dev/en/1.-quickwins/choose-regions/

AWS

- currently 38 regions, 120 availability zones
- several hundred data centres (no up to date numbers)
- number of servers:
  - 2014: at least 1.4 million
  - 2025: estimated at ~5 million

source: https://www.perplexity.ai/search/latest-official-statements-fro-4byy3OCETkKV19s9Gwzb.Q, https://aws.amazon.com/de/blogs/aws/cloud-computing-server-utilization-the-environment/

AWS

- currently 38 regions, 120 availability zones
- several hundred data centres (no up to date numbers)
- number of servers:
  - 2014: at least 1.4 million
  - 2025: estimated at ~5 million
- “hot buffer”: servers that are constantly up and running without being sold, in case they’re needed to guarantee service levels
  - ~10-20% of server fleet
  - very rough estimation: 500'000 - 1'000'000 servers
  - spot instances: a portion of this hot buffer

source: https://www.perplexity.ai/search/latest-official-statements-fro-4byy3OCETkKV19s9Gwzb.Q, https://aws.amazon.com/de/blogs/aws/cloud-computing-server-utilization-the-environment/

reserved instance:

- guaranteed for you
- you pick the instance type and size
- it stays the same for the time you reserved it (typically 1 - 3 years)
- well suited for long-term, predictable workloads
- up to 72% discount compared to on-demand instances

reserved instance:

- guaranteed for you
- you pick the instance type and size
- it stays the same for the time you reserved it (typically 1 - 3 years)
- well suited for long-term, predictable workloads
- up to 72% discount compared to on-demand instances

on-demand instance:

- guaranteed for you
- you pick the instance type and size
- flexibility
- higher price

reserved instance:

- guaranteed for you
- you pick the instance type and size
- it stays the same for the time you reserved it (typically 1 - 3 years)
- well suited for long-term, predictable workloads
- up to 72% discount compared to on-demand instances

spot instance:

- can be revoked with 2 mins notice
- availability of what you need is not guaranteed
- flexibility: you can always adapt to what you need at that given time
- instances are auctioned: up to 90% discount, but price can rise up to on-demand instance prices over time

reserved instance:

- guaranteed for you
- you pick the instance type and size
- it stays the same for the time you reserved it (typically 1 - 3 years)
- well suited for long-term, predictable workloads
- up to 72% discount compared to on-demand instances

=> **prevents** hardware efficiency

spot instance:

- can be revoked with 2 mins notice
- availability of what you need is not guaranteed
- flexibility: you can always adapt to what you need at that given time
- instances are auctioned: up to 90% discount, but price can rise up to on-demand instance prices over time

=> **promotes** hardware efficiency

03
The tricky bits and how to get them right

termination

risk:

- Spot instances can be revoked with 2 mins notice when the cloud provider needs the resources for on-demand customers.

termination

risk:

- Spot instances can be revoked with 2 mins notice when the cloud provider needs the resources for on-demand customers.

how to deal with it:

- stateless or fault-tolerant workloads (e.g. interactions)
- start / stop workloads (e.g. machine learning training)
- termination trigger to save state
- retries / rescheduling
- checkpointing (e.g. for CI/CD pipelines)

Termination by AWS is *very* rare!

no stop / restart mechanism

risk:

- Spot instances cannot be stopped and restarted (like on-demand instances), only terminated. Stored data is lost on termination.

no stop / restart mechanism

risk:

- Spot instances cannot be stopped and restarted (like on-demand instances), only terminated. Stored data is lost on termination.

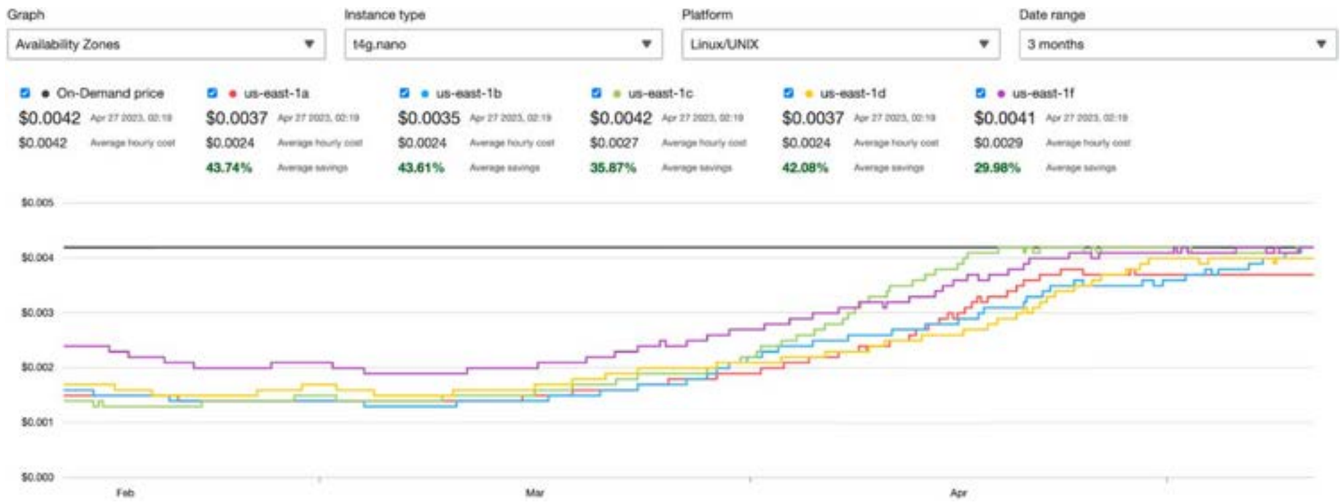
how to deal with it:

- save state and data on external persistent storage volumes
- attach volume upon instance creation or swap the boot volume

bidding mechanism: price and availability can vary a lot

risk:

- Low capacity for specific instance types in a cloud region can cause delays or result in inability to launch spot instances.
- Prices can rise a lot over time.



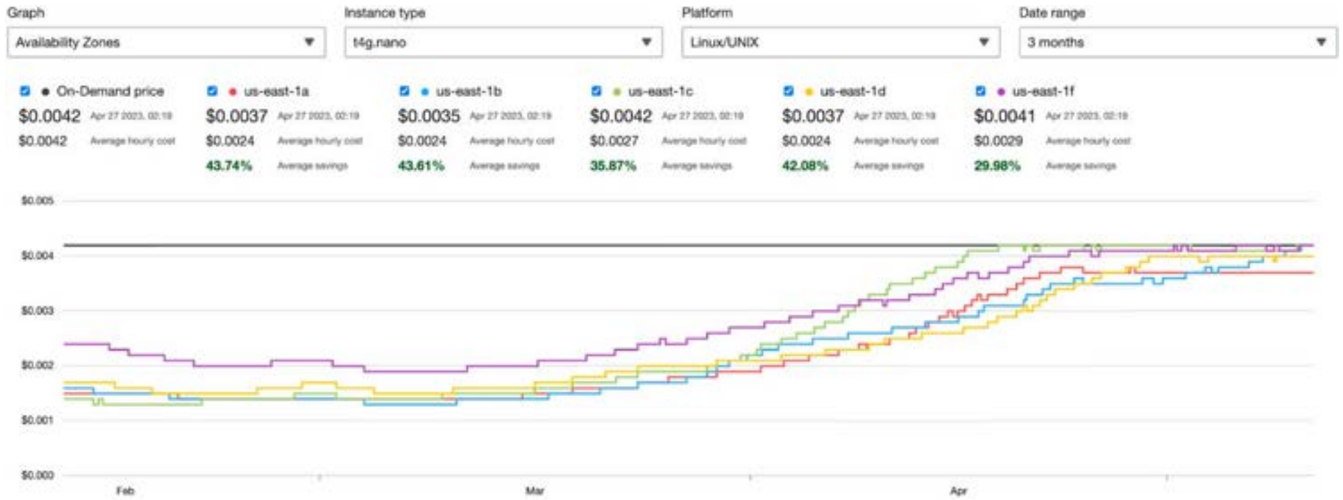
bidding mechanism: price and availability can vary a lot

risk:

- Low capacity for specific instance types in a cloud region can cause delays or result in inability to launch spot instances.
- Prices can rise a lot over time.

how to deal with it:

- diversify instance types
- ranges in size etc.
- monitor price
- smart automation



smart automation

pod autoscaler

- e.g. Horizontal Pod Autoscaler / KEDA
- creates more pods or deletes them, based on service needs

smart automation

pod autoscaler

- e.g. Horizontal Pod Autoscaler / KEDA
- creates more pods or deletes them, based on service needs

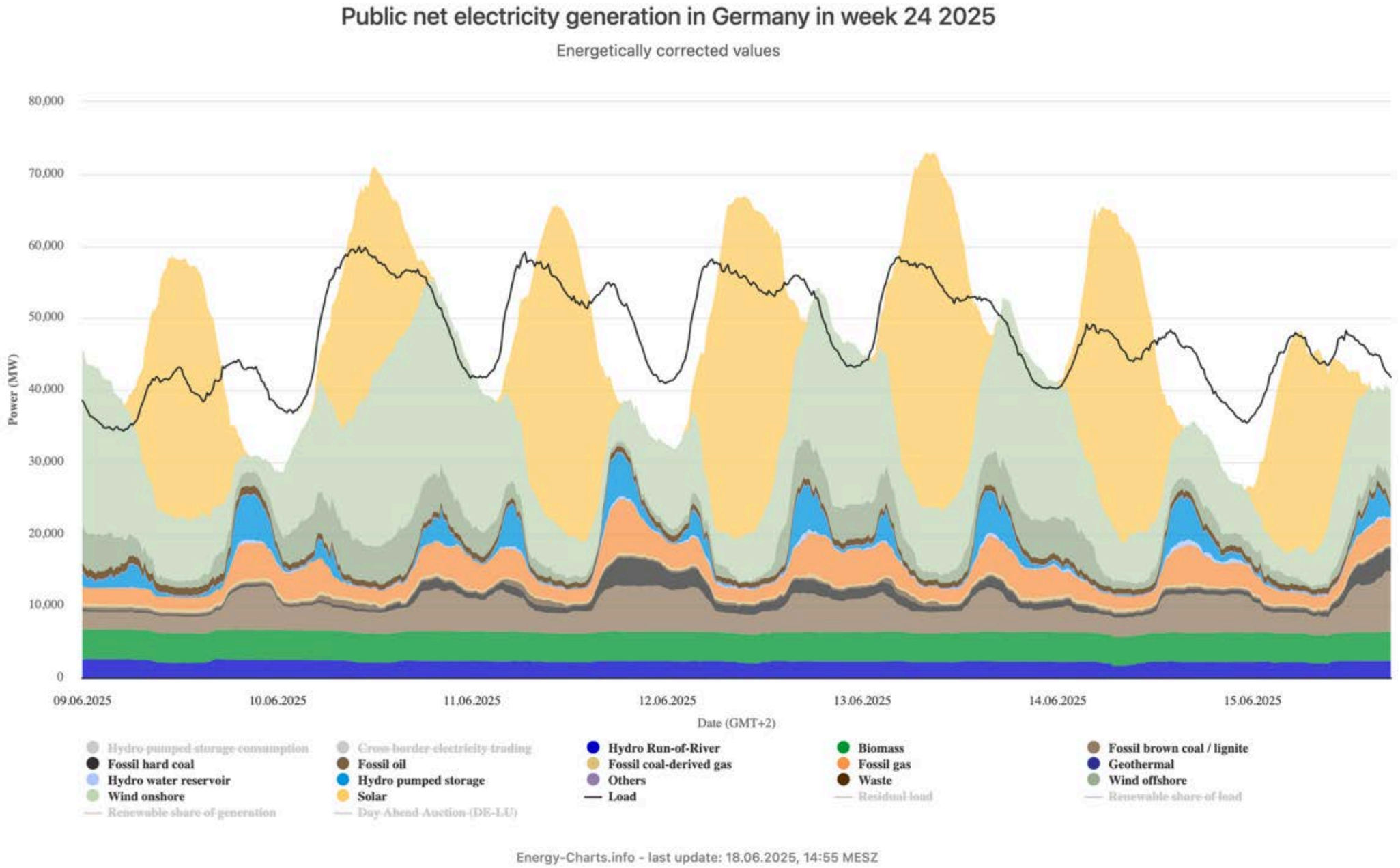
cluster autoscaler

- e.g. Karpenter
- automatically provisions nodes based on specifications
- very fast
- optimizes for price, capacity, reliability
- combine with minimum of 2 on-demand / reserved instances for cluster stability
- e.g. small instances for DEV/TEST, bigger ones for PROD

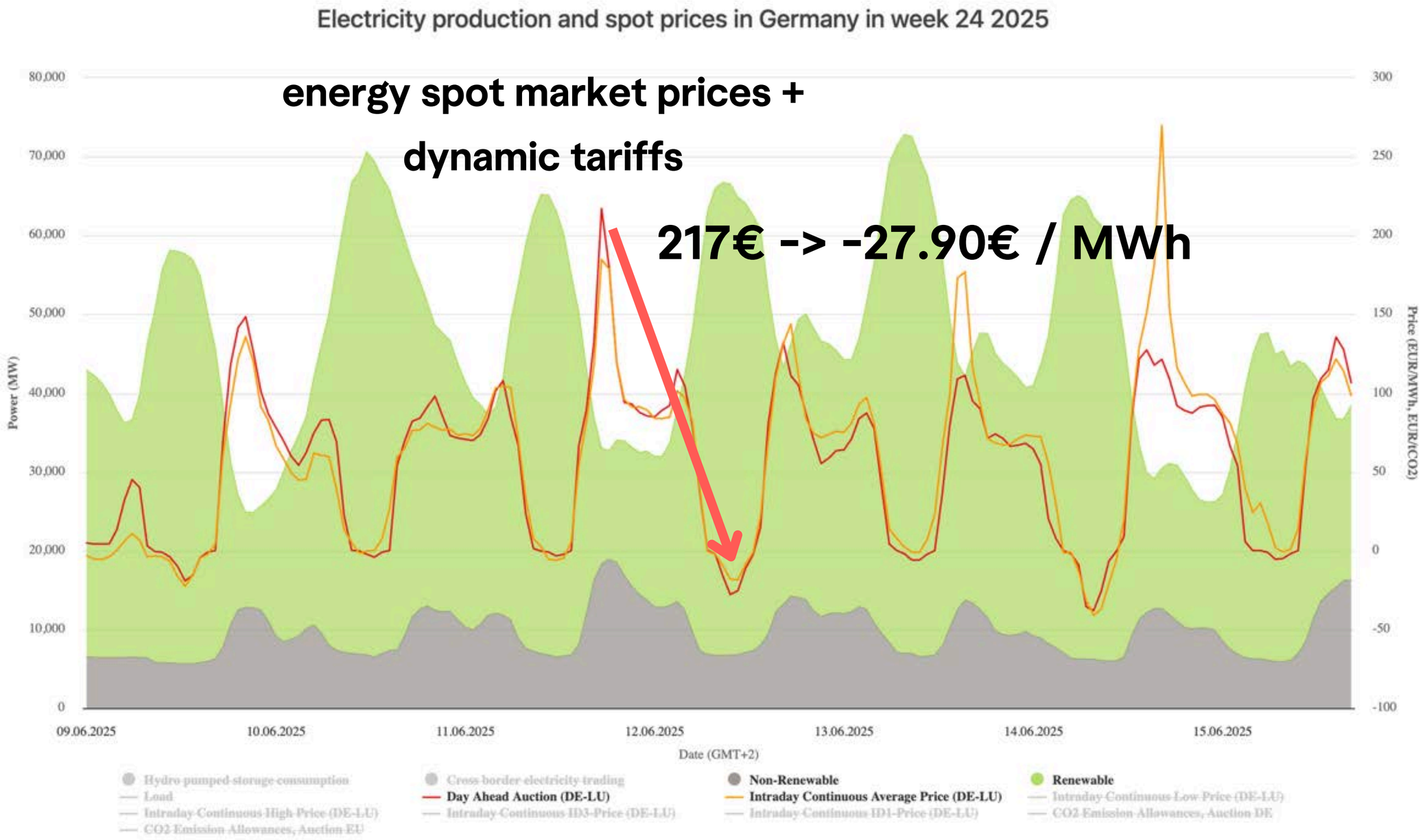
Karpenter

```yaml
apiVersion: karpenter.sh/vlbetal
kind: NodePool
metadata:
  name: default
spec:
  weight: 99
  template:
    spec:
      requirements:
        - key: karpenter.sh/capacity-type
          operator: In
          values: [ "spot" ] ## <- "on-demand"
        - key: topology.kubernetes.io/zone
          operator: In
          values: [ "eu-central-1a", "eu-central-1b" , "eu-central-1c"
        - key: kubernetes. 10/arch
          operator: In
          values: [ "amd64"]
        - key: kubernetes.io/os
          operator: In
          values: [ "linux" ]
        - key: karpenter.k8s.aws/instance-category
          operator: In
          values: [ "c", "m" ]
        - key: karpenter.k8s.aws/instance-cpu
          operator: In
          values: [ "4", "8" ]
      nodeClassRef:
        apiVersion: karpenter.k8s.aws/vlbetal
        kind: EC2NodeClass
        name: default
  limits:
    cpu: 80
  disruption:
    consolidationPolicy: WhenUnderutilized
```

04
How can spot instances help advance the energy transition?

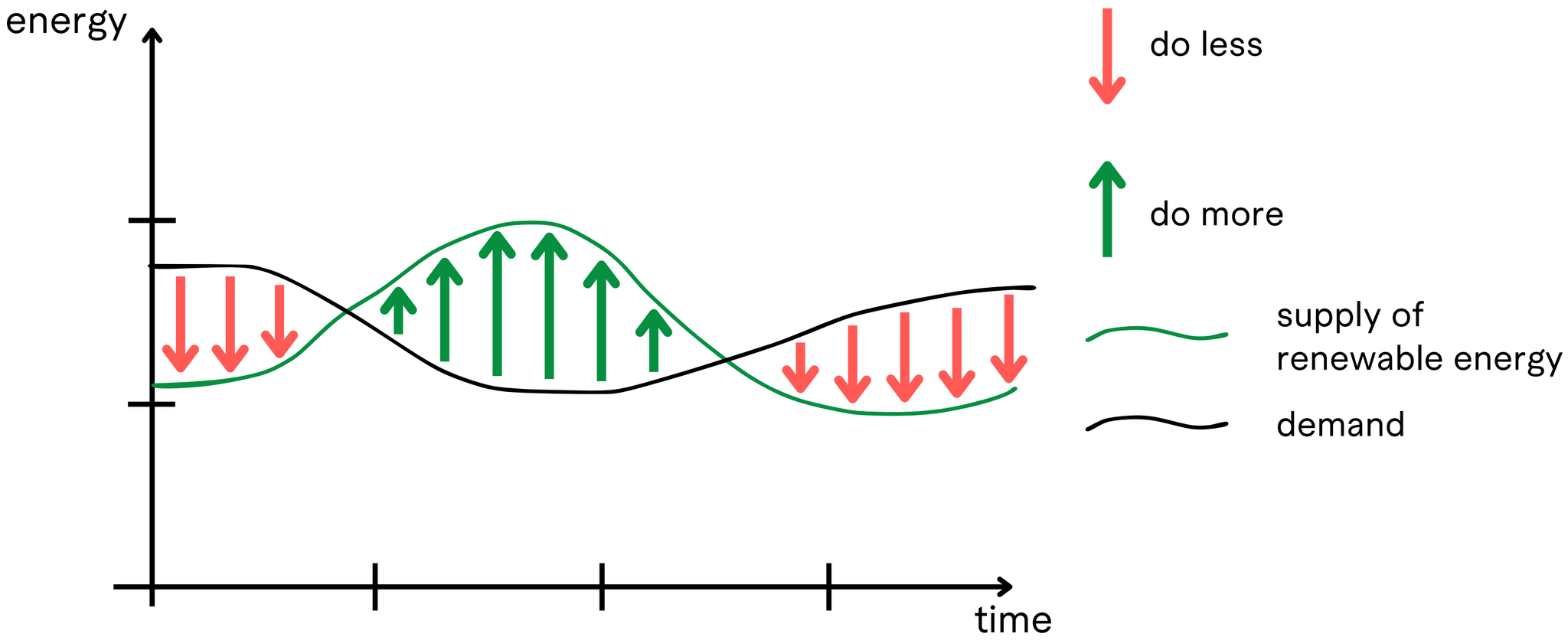


Source: https://www.energy-charts.info/charts/power/chart.htm?c=DE&week=02&l=en



Source: https://www.energy-charts.info/charts/price_spot_market/chart.htm?c=DE&week=02&l=en

Data centres will become flexible players on the grid.

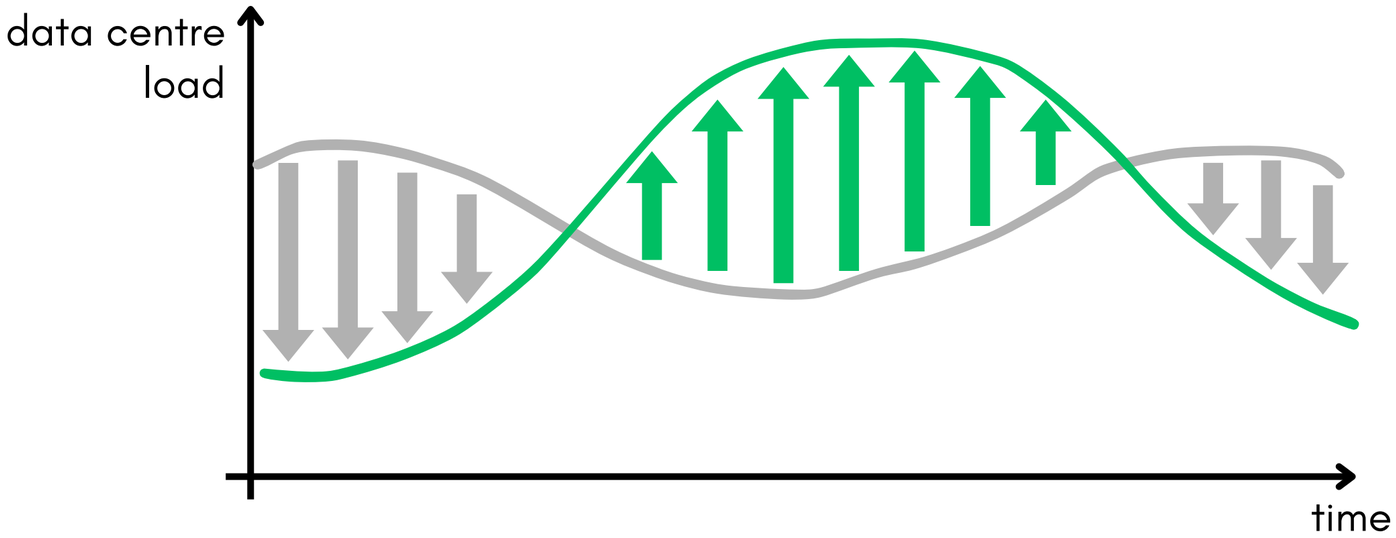


Data centre energy consumption will be shaped via resource pricing.

server capacity in the data centre

grid capacity and availability of renewable energy

spot market electricity prices



Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

location independence

- execution of workloads in different geographic regions / data centers
- support for cross-location data replication and synchronisation

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

location independence

- execution of workloads in different geographic regions / data centers
- support for cross-location data replication and synchronisation

modularity and statelessness

- services that can easily be moved and scaled
- service cuts based on scaling needs
- outsourced state management to enable easy relocation of compute

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

flexible scheduling of jobs

- done within a timeframe / until a given time
- orchestrated sequencing
- support of asynchronous processes

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

flexible scheduling of jobs

- done within a timeframe / until a given time
- orchestrated sequencing
- support of asynchronous processes

fault tolerance and resilience

- ability to reliably handle interruptions or postponements
- support for checkpointing and resumption of long-running processes

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

- efficient data management
  - data lifecycle management to minimise storage and transfer of unnecessary data
  - data compression and efficient formats

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

efficient data management

- data lifecycle management to minimise storage and transfer of unnecessary data
- data compression and efficient formats

efficient utilisation of resources

- waste elimination, minimal resources for software + ops
- fast scalability according to demand and prices
- containers, Cloud-native technologies

Architecting software to be suited for spot instances also helps prepare the software for data centre demand shaping.

configurable compromises between performance and cost savings

- tasks with different profiles / multiple operating modes
- ability to customise quality of service
- dynamic and automatic adjustments based on current conditions

Sometimes, the “old stuff” we already have is just fine - we just need to use it!

neuland

Thank you!
Questions?

Anita Schüttler
anita.schuettler@neuland-bfi.de