MINIMIZING ENERGY CONSUMPTION IN BARE-METAL K8S CLUSTERS

GMX WEB.DE mail.com united internet media

Dr. David Meder-Marouelli

- Lead Architect Delivery Platform
- Expert for CI/CD
- 18 years of IT architecture
- Since 2015 with Mail & Media

Marco Schröder

- Head of Automation Services
- 19 years in IT operations
- Believes in Infrastructure automation and containers
- Leading the Kubernetes team at Mail & Media since 2018

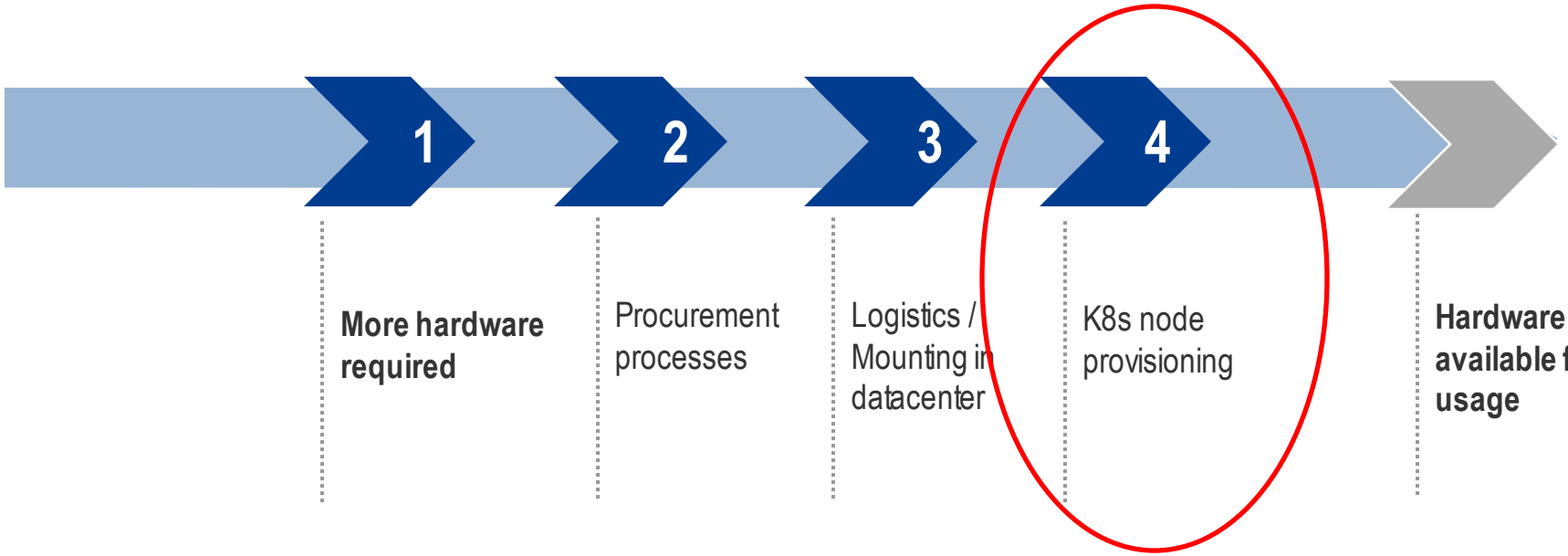
Context

- Large e-Mail provider, >42 million active users
- Microservice landscape
- Multi-tenant Kubernetes platform
- On-premise (IONOS data centers)
- Bare metal (scale appropriate)
  - Each 100-250 cores, 800G RAM, 200-1000W

GMX WEB.DE mail.com united internet media

How much hardware do you need?

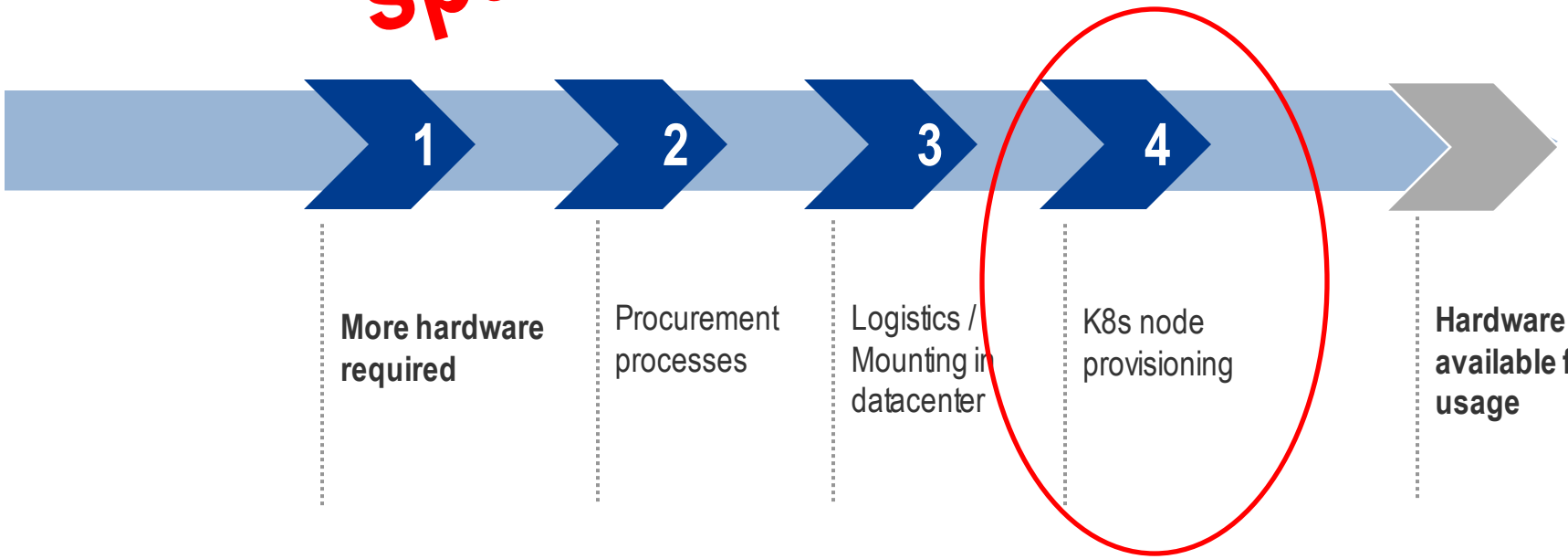
Scaleout is not as dynamic as with public clouds



How much hardware do you need?

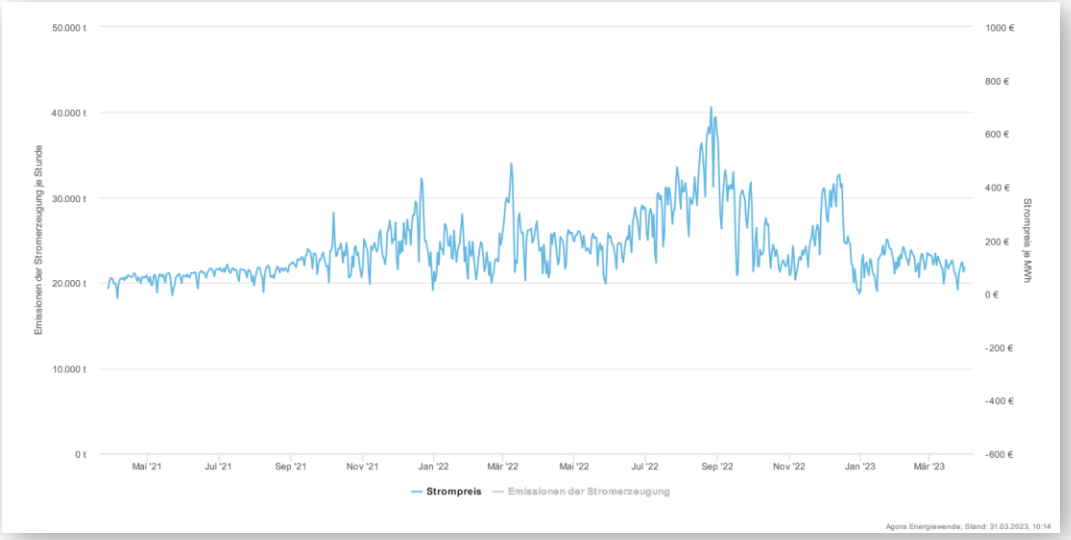
Scaleout is not as dynamic as with public clouds

We need „enough“ spare capacity



Energy Saving Motivation

- Save CO_2 – save the planet
- Energy crisis
- Energy prices skyrocket
- Save costs



Source: Agora Energiewende

Easy solution:

0 Servers == 0 Watt

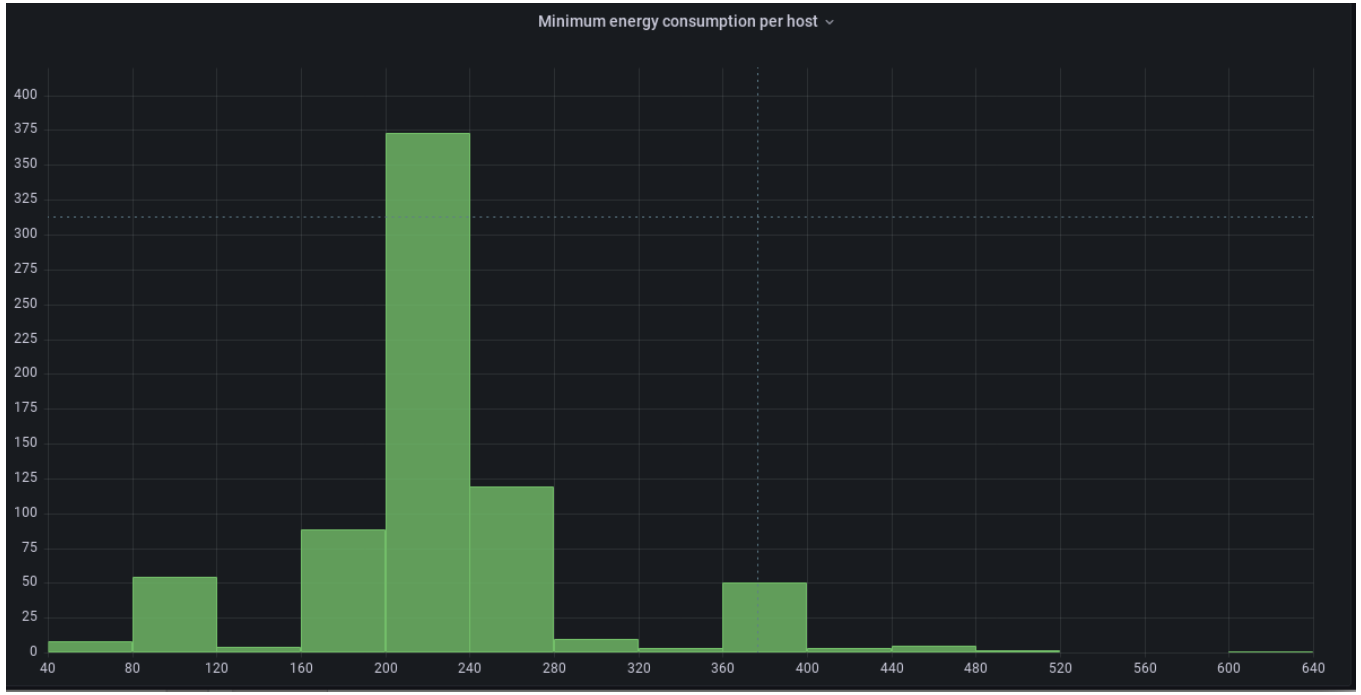
What to Minimize and how to Measure it?

- **Measure wrt. specific aspects**
- **KPIs**
  - Reliable & repeatable
  - Robust
- **Set of KPIs needed**
- **Note: More abstraction = more assumptions**

KPI Proposal: Baseline

Server Idle Power

- Optimize base consumption
- Evaluate components
- Tune configuration



Consumption related to:

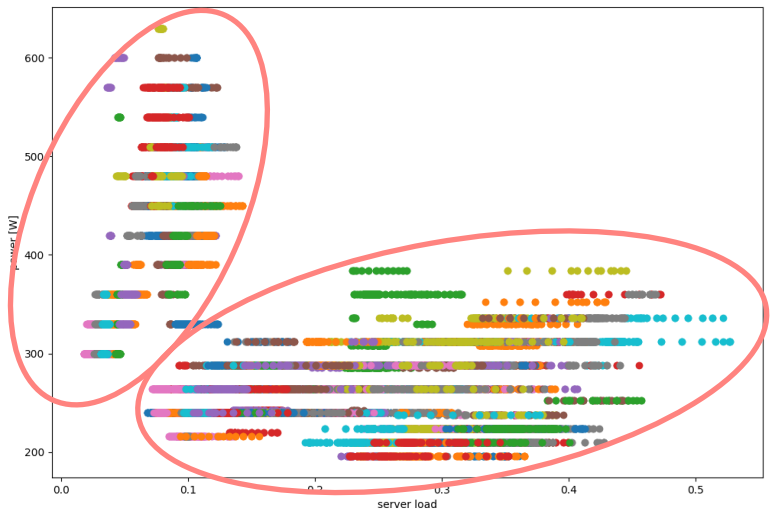
- **Idle servers**
- Servers with load
- Applications
- Requests (external)
- "Average user" / Product / ...

KPI Proposal: Raw Load Performance

Power Performance

- **Behavior under load**
- **CPU optimizations**
- **Thermal tuning**

- **Normalization?**
  - CPU model/generation/brand
  - Clock frequency



Consumption related to:

- Idle servers
- **Servers with load**
- Applications
- Requests (external)
- "Average user" / Product / ...

KPI Proposal: Cluster Performance

Cluster Power Performance

- **Interplay in a cluster**

- **Power vs. load distribution**
- **Load composition**
- **Utilization optimization**

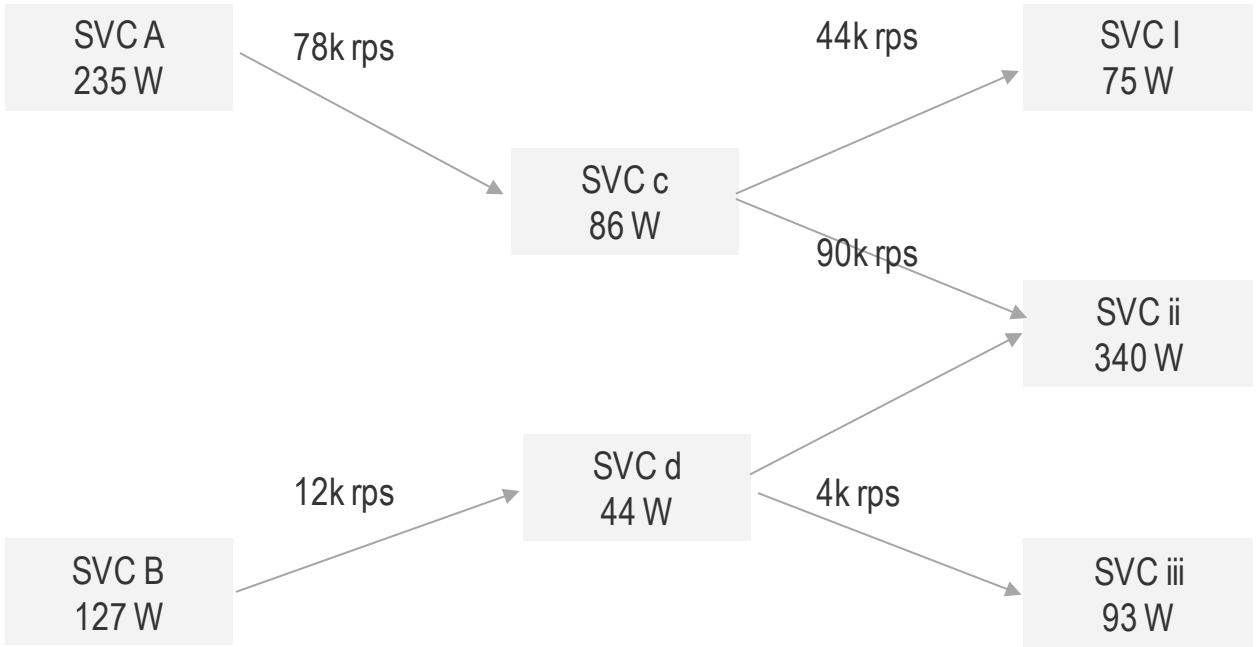
Consumption related to:

- Idle servers
- **Servers with load**
- Applications
- Requests (external)
- "Average user" / Product / ...

KPI Proposal: Application Performance

Application Performance

- **"Power per request"**
- **Bridge to business**
- **Fuzzy: many assumptions required**



Consumption related to:

- Idle servers
- Servers with load
- **Applications**
- **Requests (external)**
- "Average user" / Product / ...

What reserves do we have?

1. Scale-out reserves
2. Geo-redundancy reserves
3. Peak performance reserves

Reduce idle costs

1. **Cutting Scale-out reserves**
   - Power-off nodes with low usage
   - Host specific infrastructure generates cost (kubelet, fluentd, kube-proxy, coredns, node_exporter, …)
   - How fast can we re-enable spare hardware?
   - Infrastructure automation is key
   - Immutable infrastructure => no config drift
2. **Geo-redundancy reserves**
3. **Peak performance reserves**

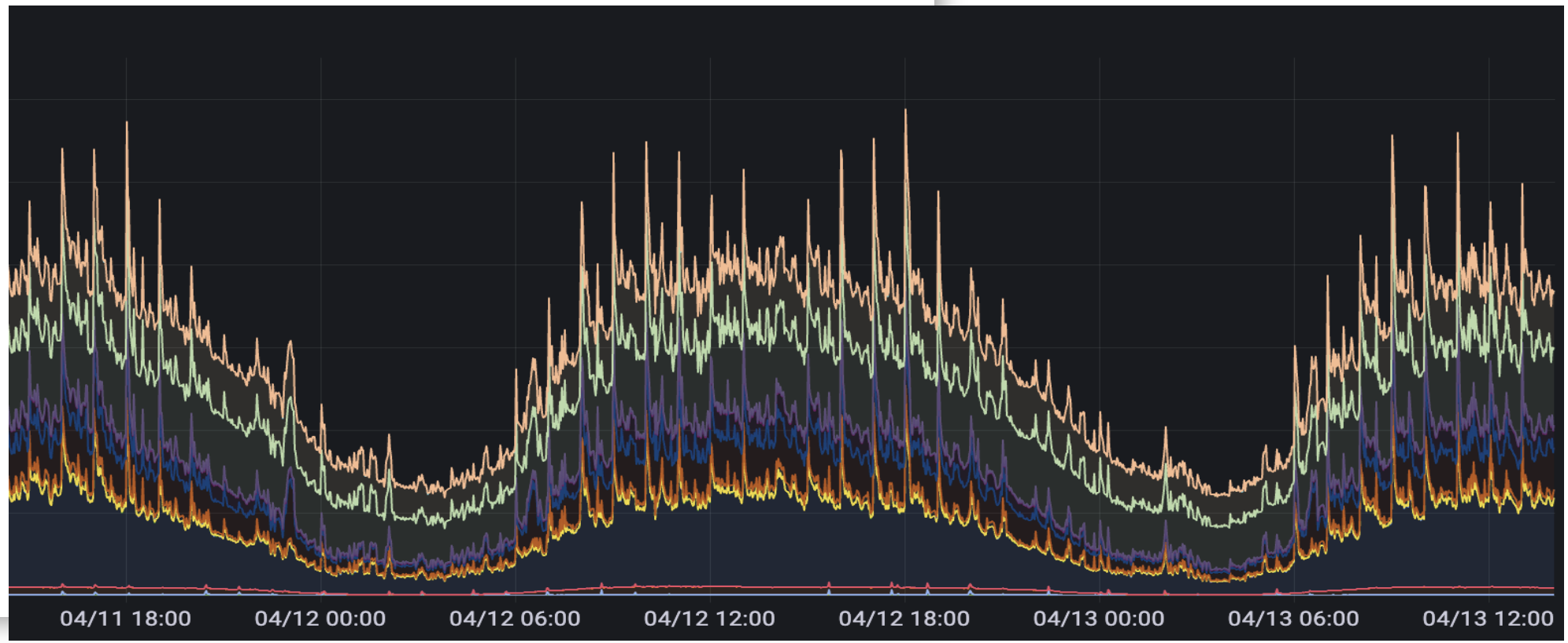
GMX WEB.DE mail.com united internet media

Optimize redundancy overhead

1. Cutting Scale-out reserves
2. **Cutting Geo-redundancy reserves**
   - Again: Infrastructure automation, but faster
   - Negative impact on time-to-recovery
   - Need management buy-in
   - Create transparency about risk (probability) vs. cost-savings
   - Regular emergency drills create confidence
3. Peak performance reserves

What reserves do we have?

1. Cutting Growth reserves
2. Cutting Geo-redundancy reserves
3. **Cutting Peak performance reserves**



GMX WEB.DE mail.com united internet media

What reserves do we have?

1. Cutting Growth reserves
2. Cutting Geo-redundancy reserves
3. **Cutting Peak performance reserves**
   - More difficult
   - Reliable automation needed, zero manual interaction
   - Shift batch load to times of low-usage
   - Daily variations: Consider nightly shutdown of nodes
   - Hourly peaks = tenant (peak) resource reservation

GMX WEB.DE mail.com united internet media

Peak performance overhead / cluster optimization

- HPA allows nightly shutdown of nodes because of automatic reduction of workload replicas
- VPA gives recommendation based on actual usage metrics to tune resource requests for deployments

GMX WEB.DE mail.com united internet media

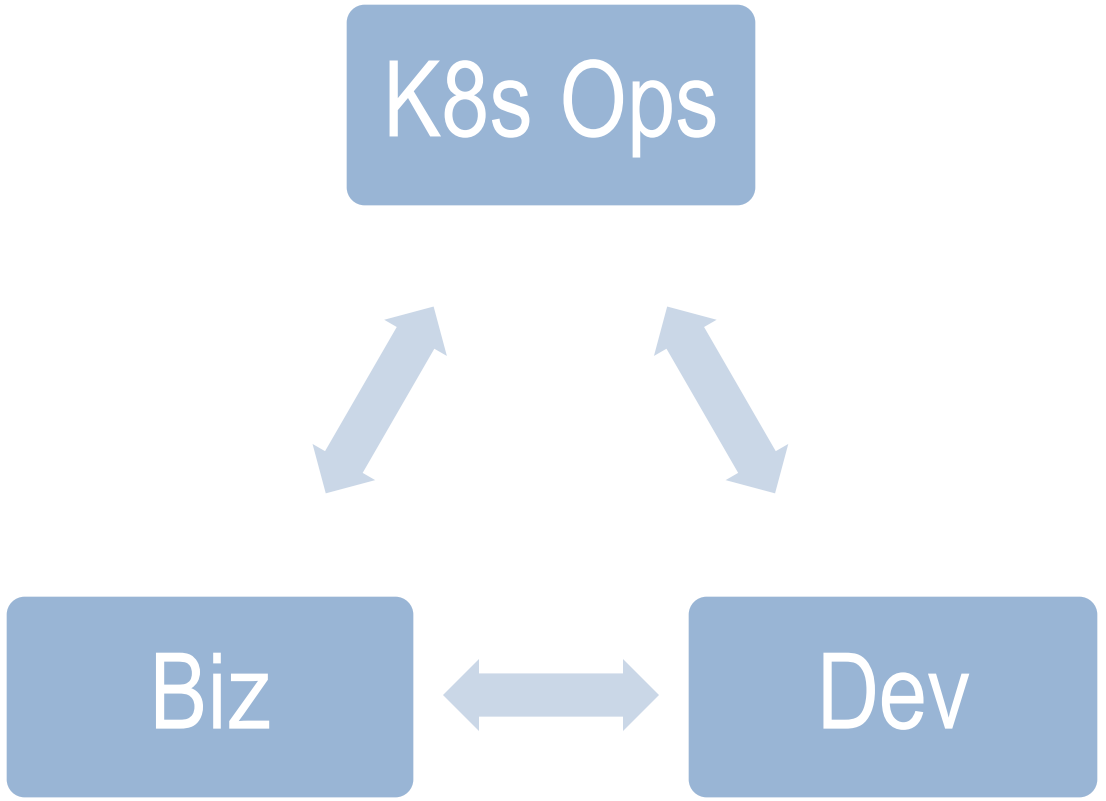
Reduce Idle Consumption as much as possible

- Reducing fan speed of servers saved us 15 W per node (15KW = 10MWh / month)

GMX WEB.DE mail.com united internet media

Beyond Single Measurements

- More abstract KPIs
  - Applications
  - Requests
  - Users
  - Products
- Link between business areas
- Transparency for
  - Product Owners
  - Developers
  - Platform Management / Capacity planning



GMX WEB.DE mail.com united internet media

1&1 Mail & Media GmbH
Brauerstraße 48
76135 Karlsruhe
Tel. +49 721 960 97 40
www.mail-and-media.com
info@mail-and-media.com