Large-scale Cluster Management at Google with Borg

By Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune and John Wilkes

Presented by Matthias Erdmann, Simon Buchacher, and Spencer Johann Thellmann

Overview

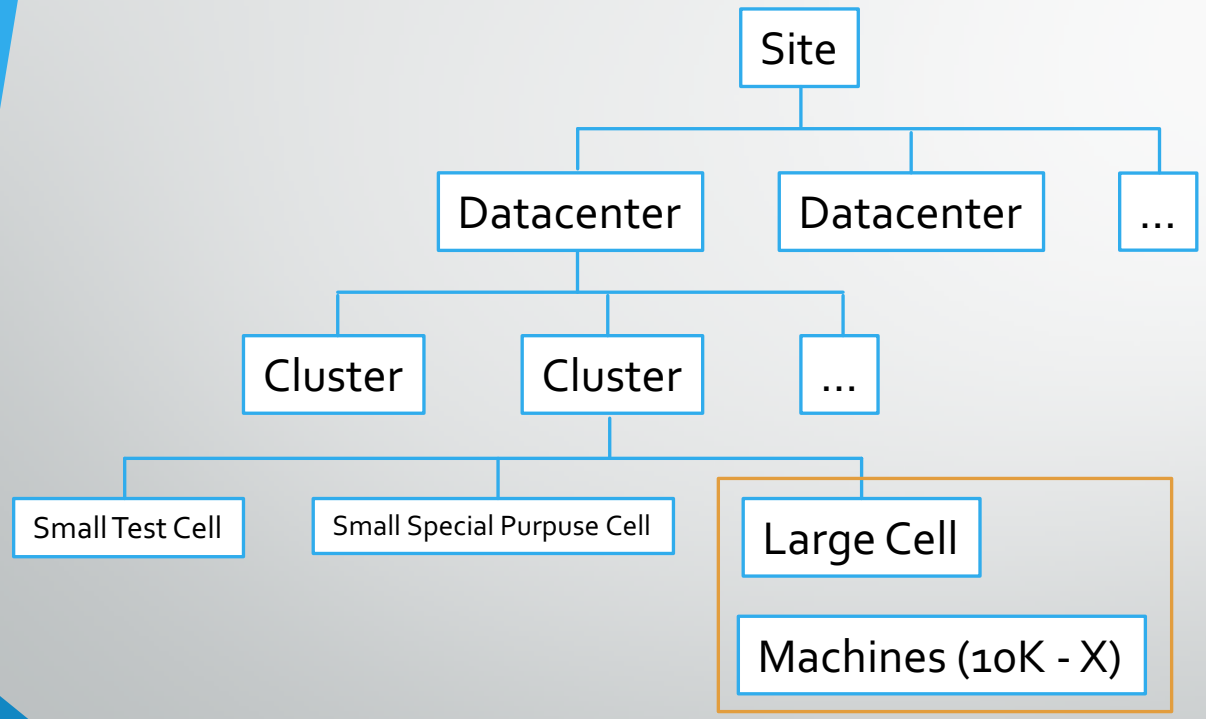
I. Borg
II. Evaluation
III. Opinion


A Brief Introduction to
Borg

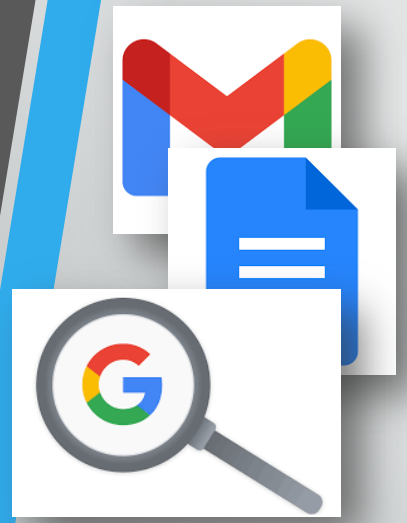
Analogy -

- Lectures
- Practicals
- Papers
- (essential & recommended) Readings
- Exams
- Free time

Google infrastructure hierarchy



Cells: heterogenous workload



Long-running services

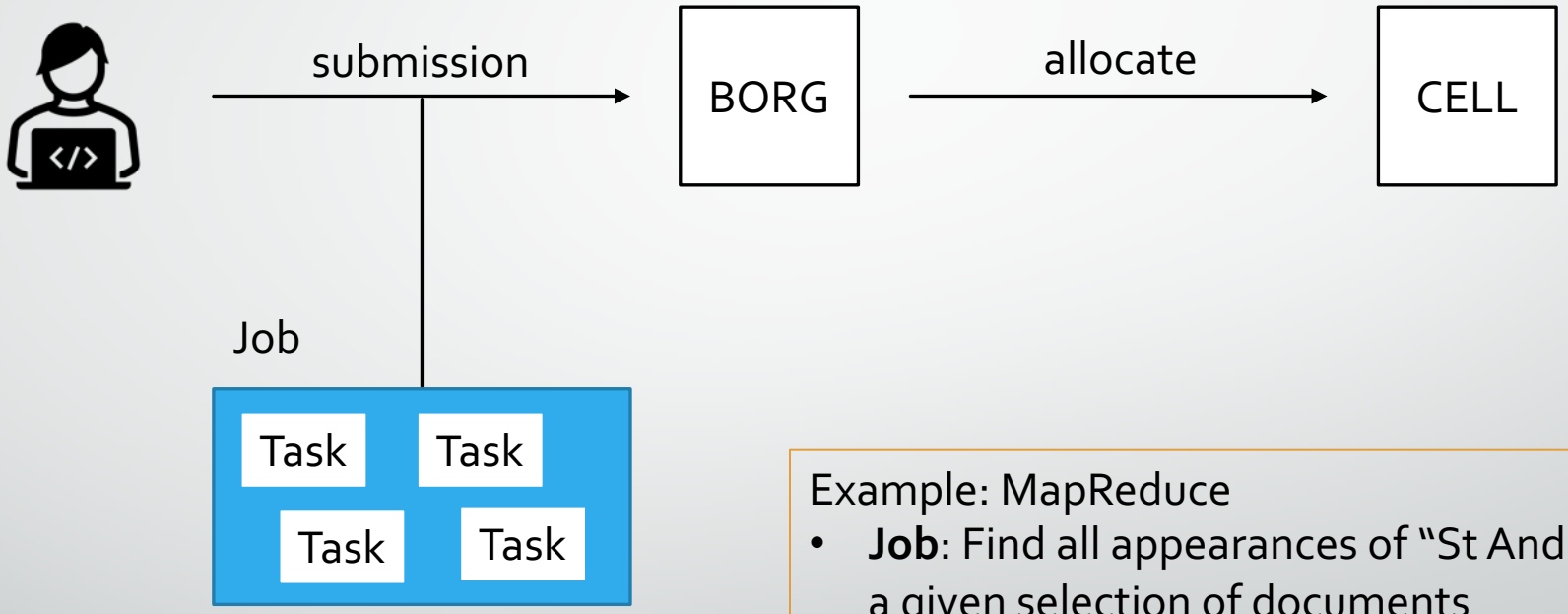
- (mostly) prod
- Constant availability

Batch jobs

- (mostly) non-prod
- Short to long jobs (seconds - days)

→ Variations in mix of prod and non-prod jobs

USP: Decoupled scheduling



Example: MapReduce

- **Job**: Find all appearances of "St Andrews" in a given selection of documents
- **Tasks**: Replicas of Mapping and Reducing Tasks + Master Task

Allocation of jobs

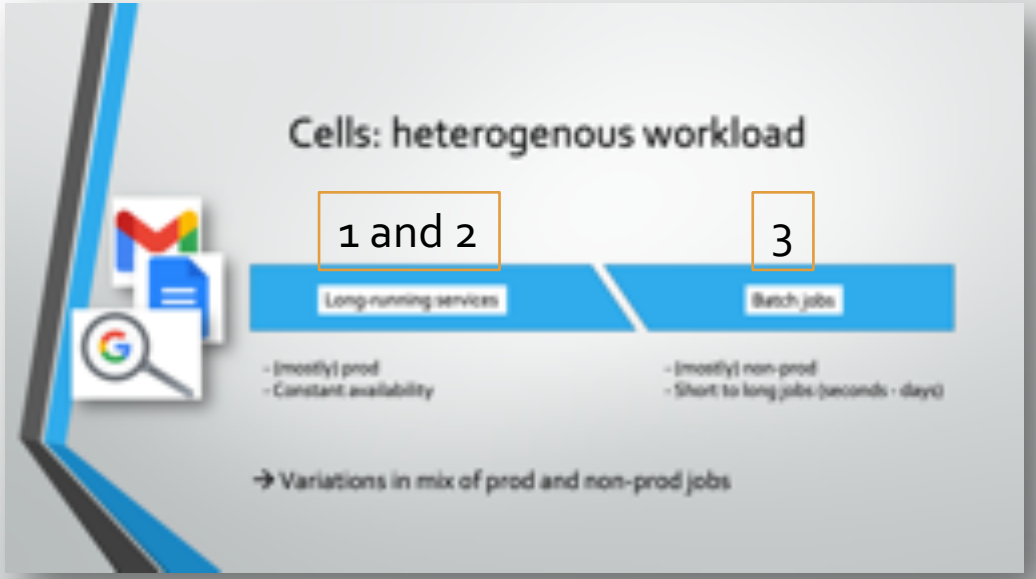
Alloc: resources on a machine that are obtained to run a job

- **Heterogeneity** in clusters: size, processor type, performance, capabilities
- **Importance of job (→ priority)**
- **Capacities (→ quota)**

Priority

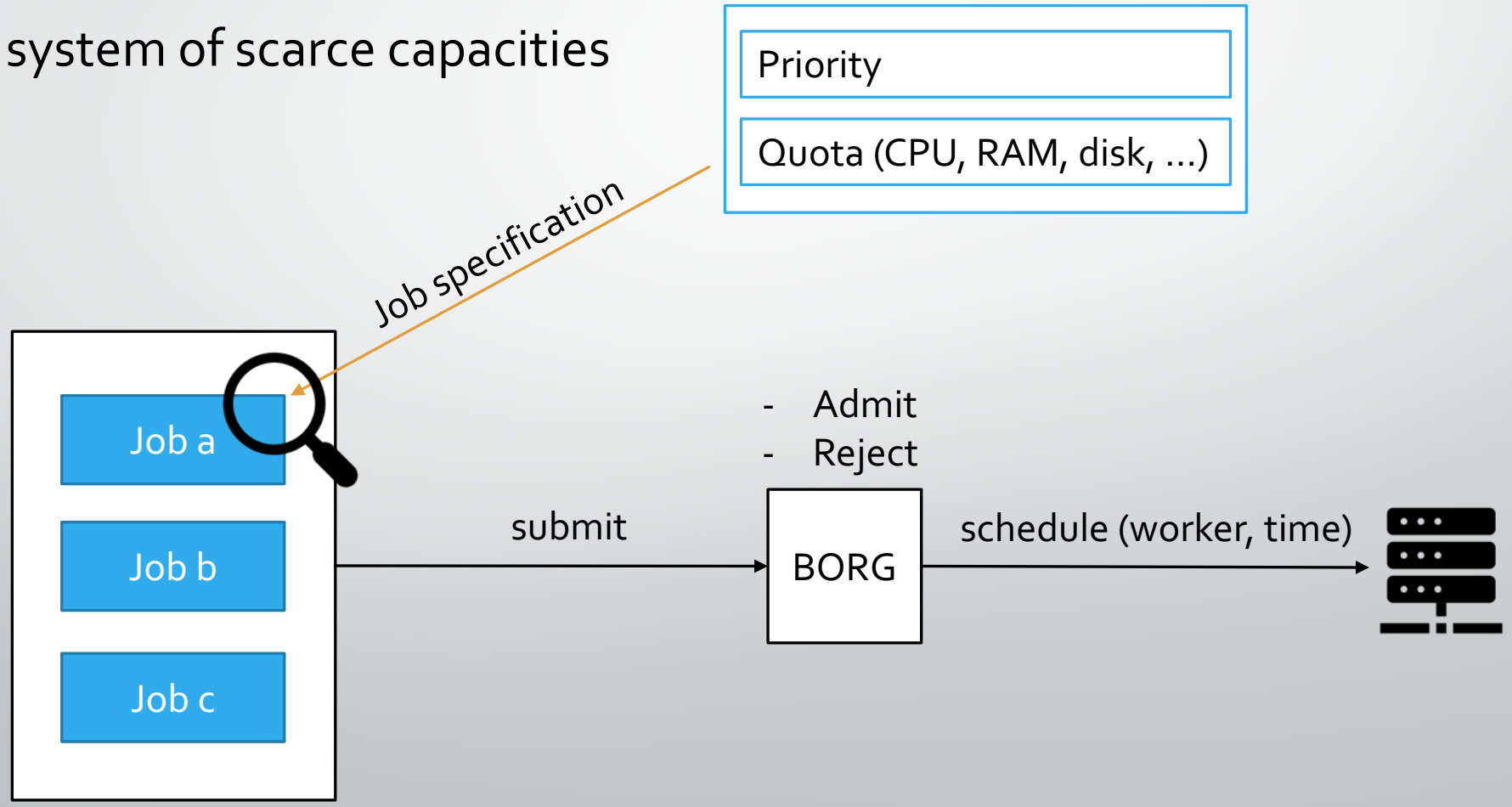
Non-overlapping priority bands:

1. Monitoring
2. Production
3. Batch
4. Best effort (testing)

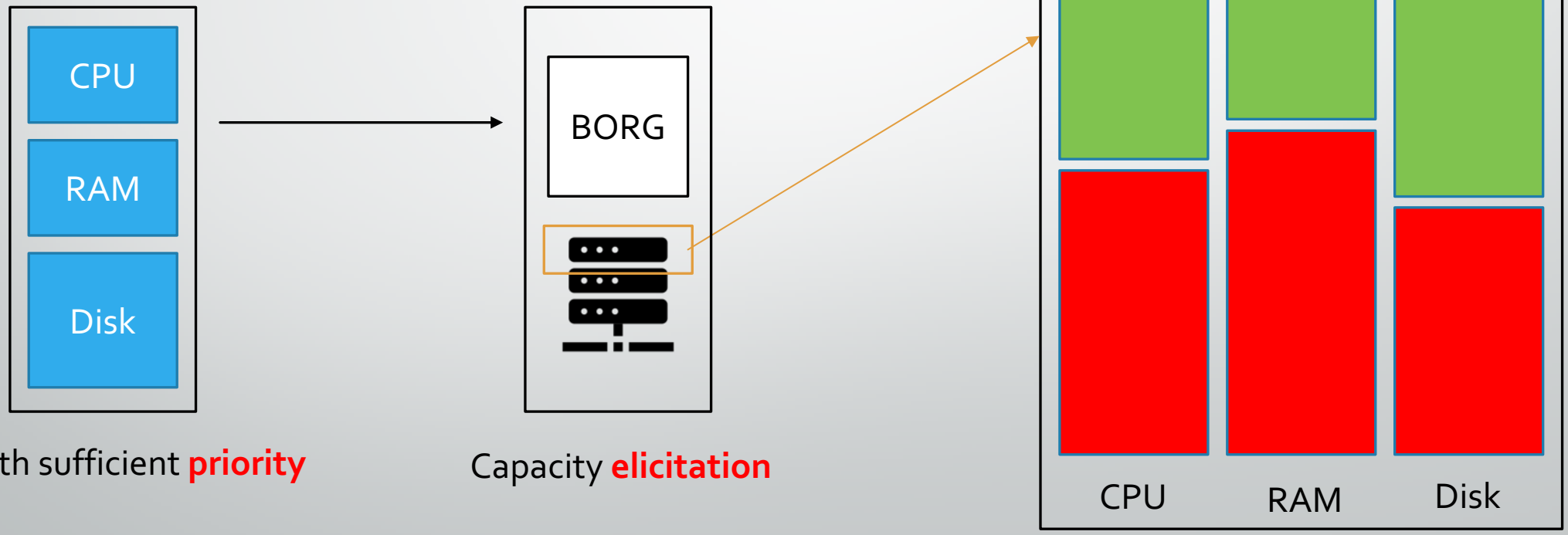


Admission control – an internal auction

→ In a system of scarce capacities



Scheduling jobs…

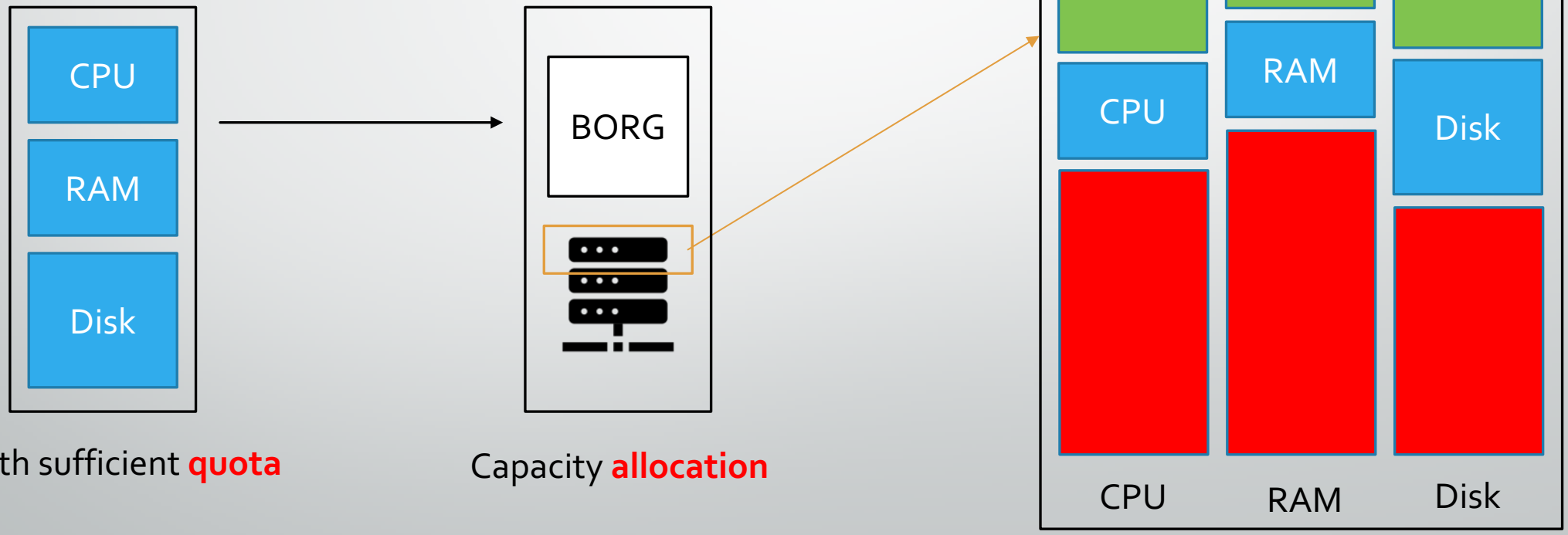


Job with sufficient **priority**

Capacity **elicitation**

Worker: capacities for a certain **period of time**

...Scheduling jobs



Job with sufficient **quota**

Capacity **allocation**

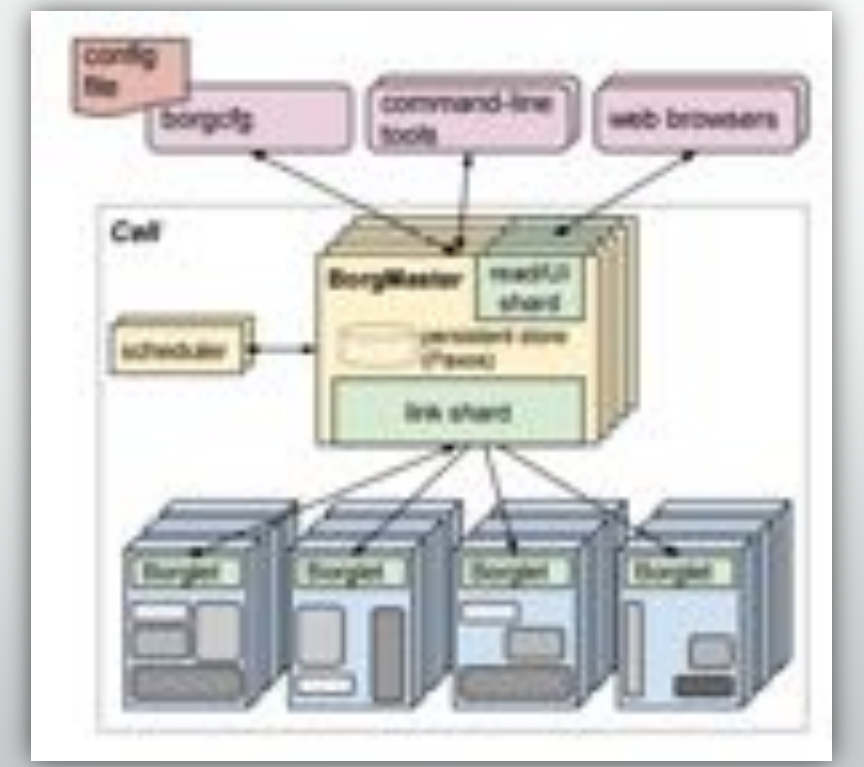
Worker: capacities for a certain **period of time**

User Tradeoffs:

- Specify priority: admission vs cost
- Specify quota: admission vs sufficient capacities

"Cost" – We assume that users receive a budget of real or some "virtual currency" they can spend on submitting jobs.

Borg Architecture



In a nutshell

BORG

Evaluating Borg

Questions

How much can you save by sharing machines between jobs?

What happens when you decrease cell size?

What happens when users choose memory and compute freely?

How much can you save by sharing machines between jobs?

How much can you save by sharing machines between jobs?

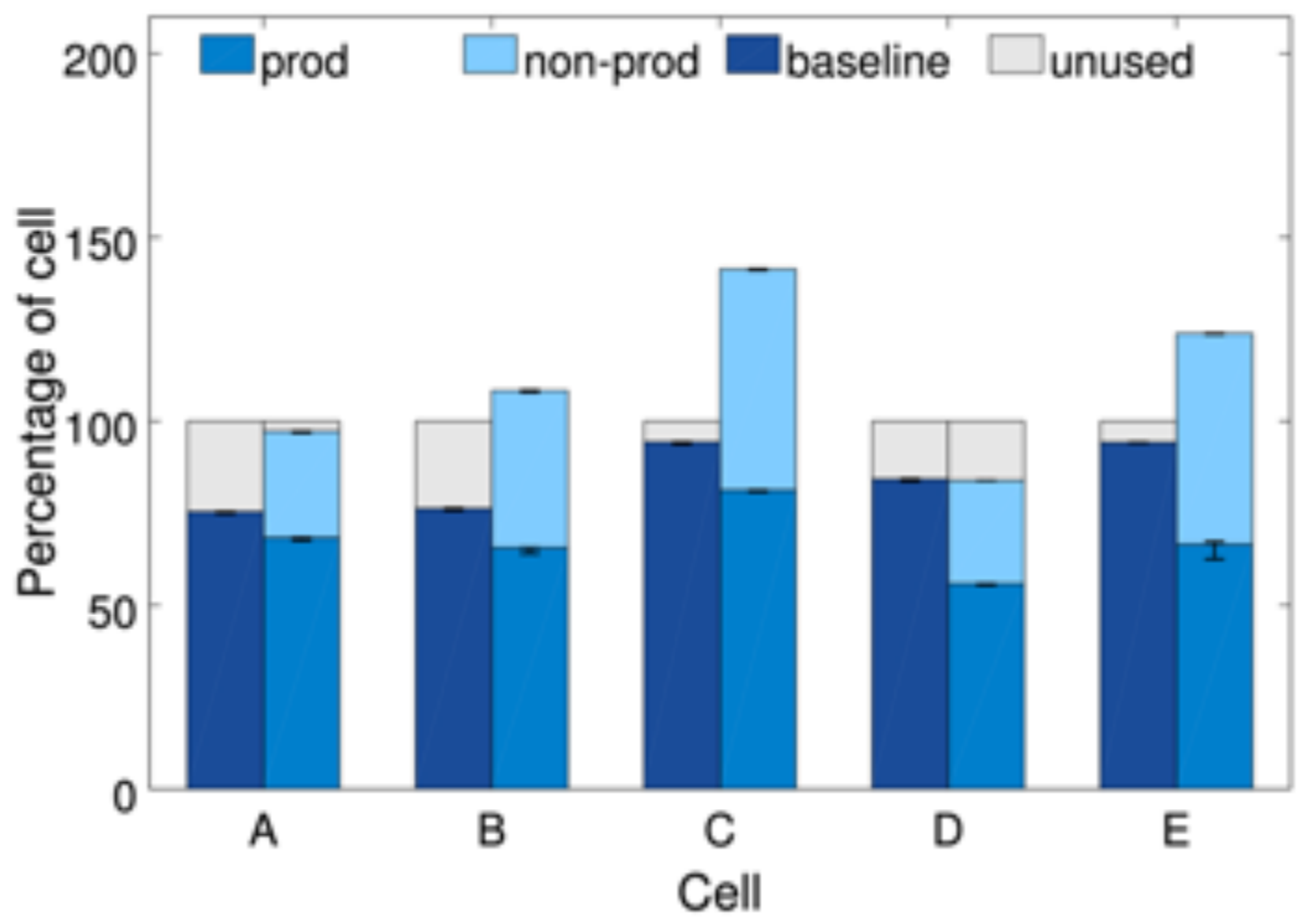
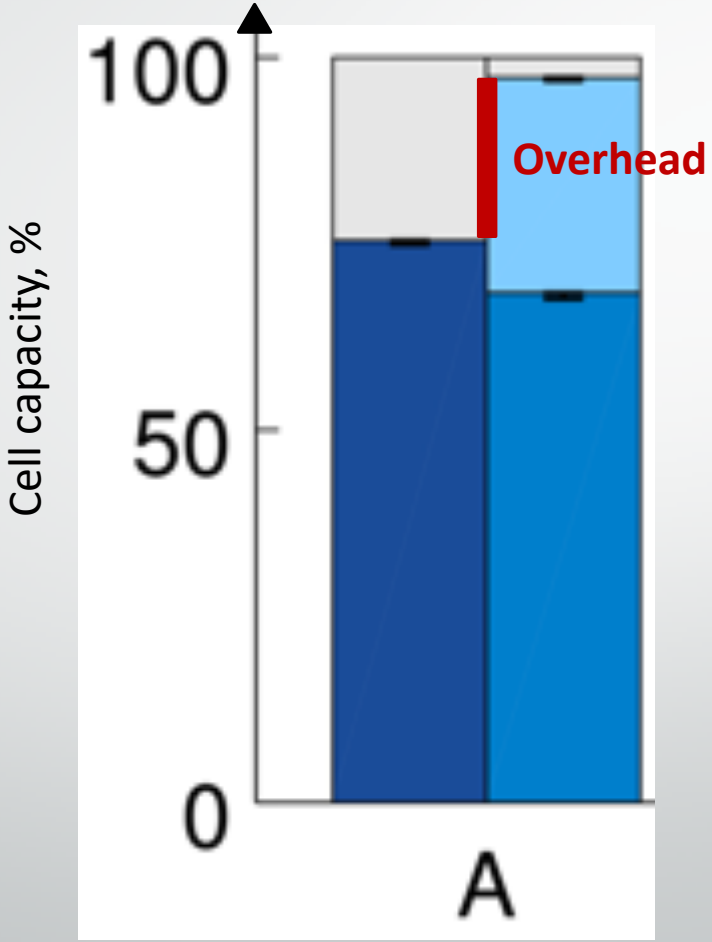


Figure 5(a)

How much can you save by sharing machines between jobs?



How much can you save by sharing machines between jobs?

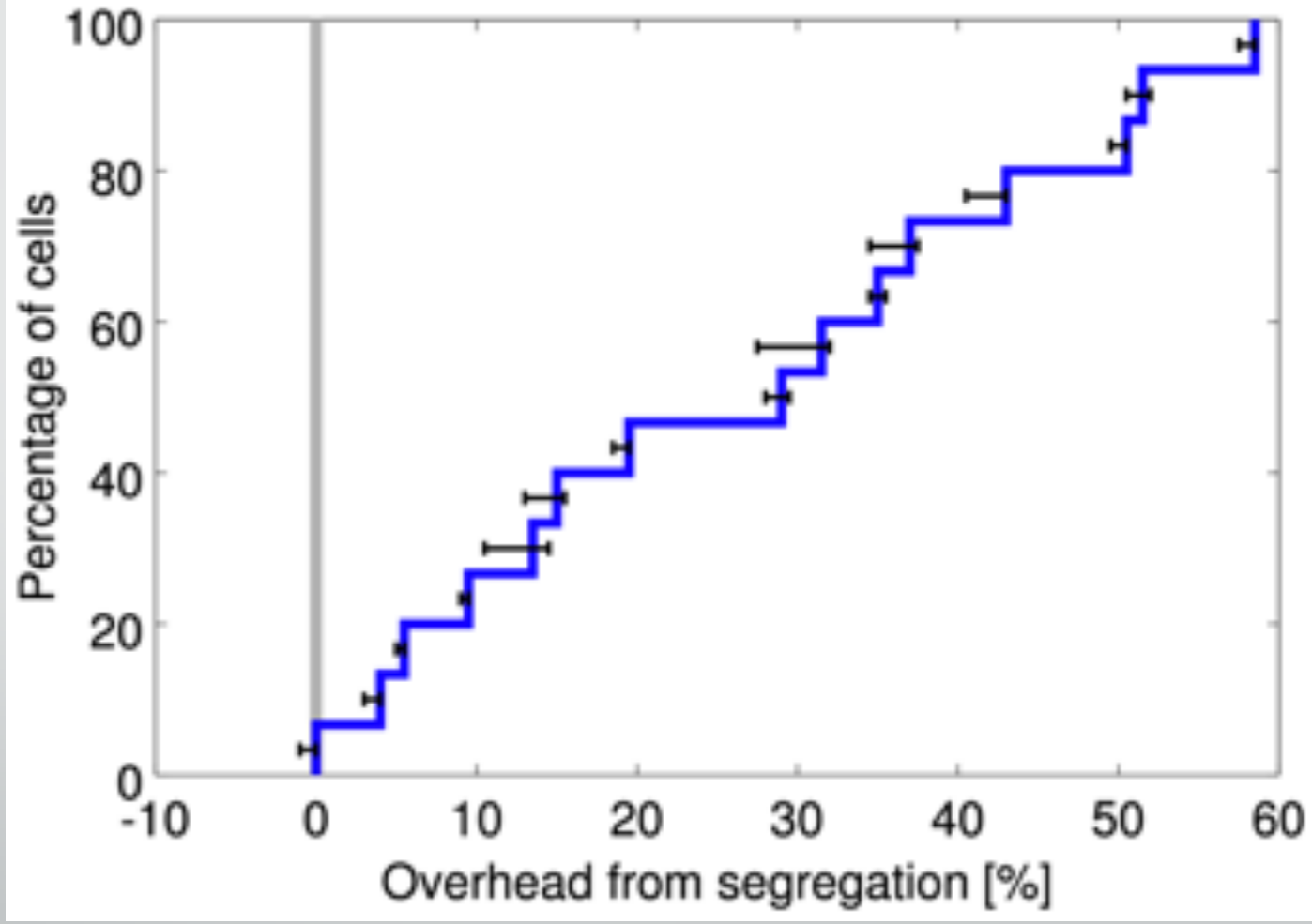


Figure 5(b)

How much can you save by sharing machines between jobs?

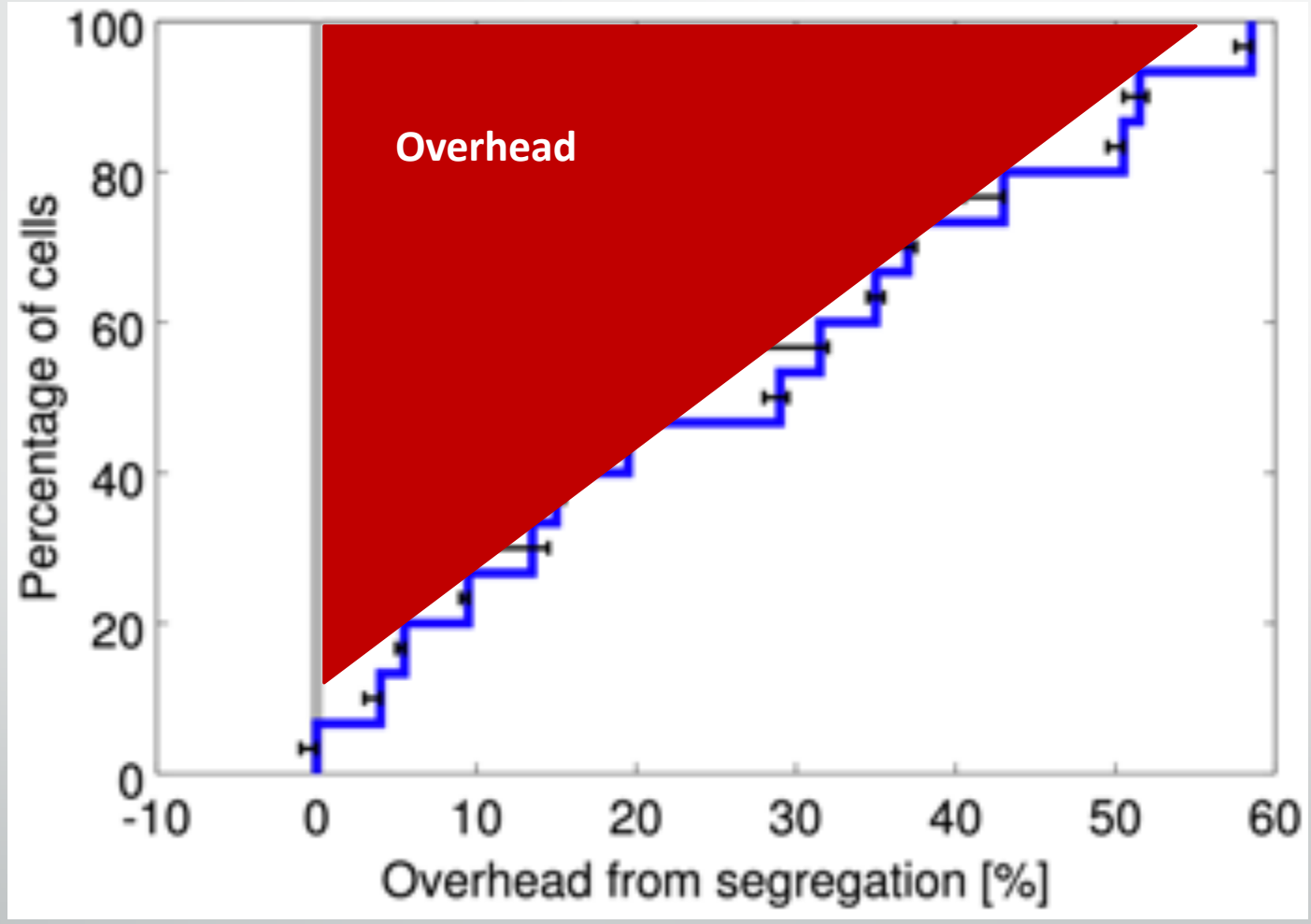


Figure 5(b)

What happens when you decrease cell size?

What happens when you decrease cell size?

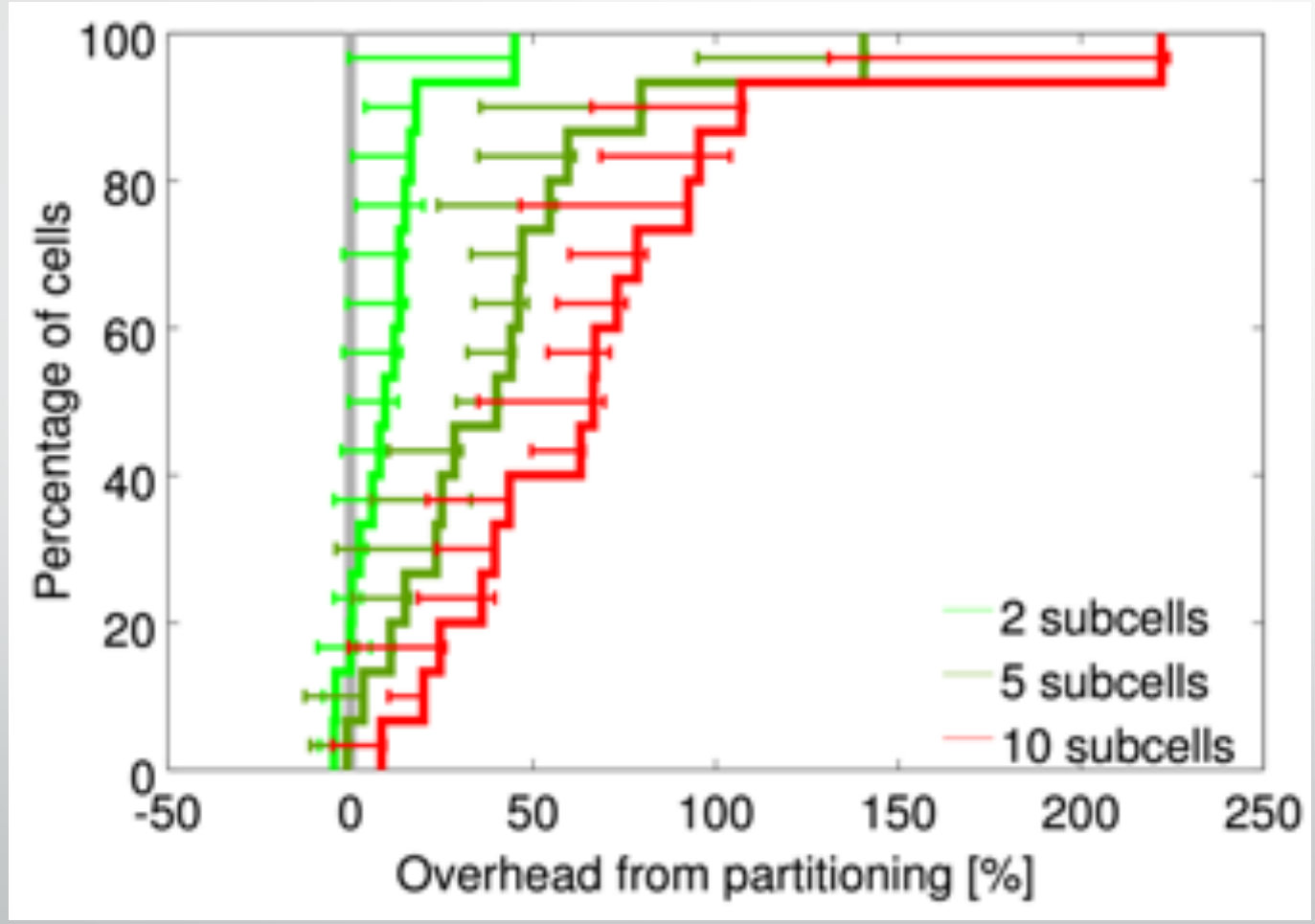


Figure 7(b)

What happens when you decrease cell size?

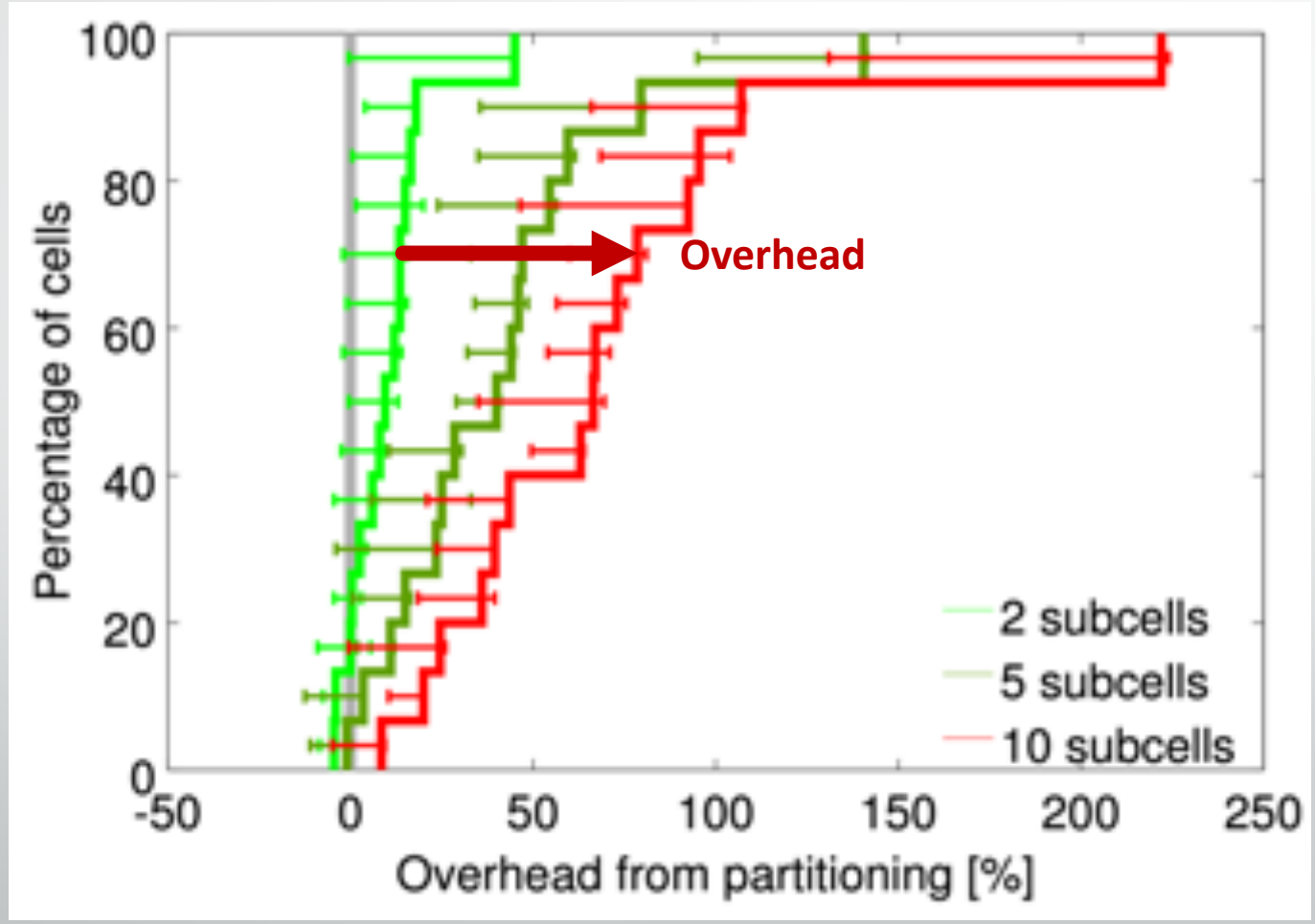


Figure 7(b)

What happens when users choose memory and compute freely?

What happens when users choose memory and compute freely?

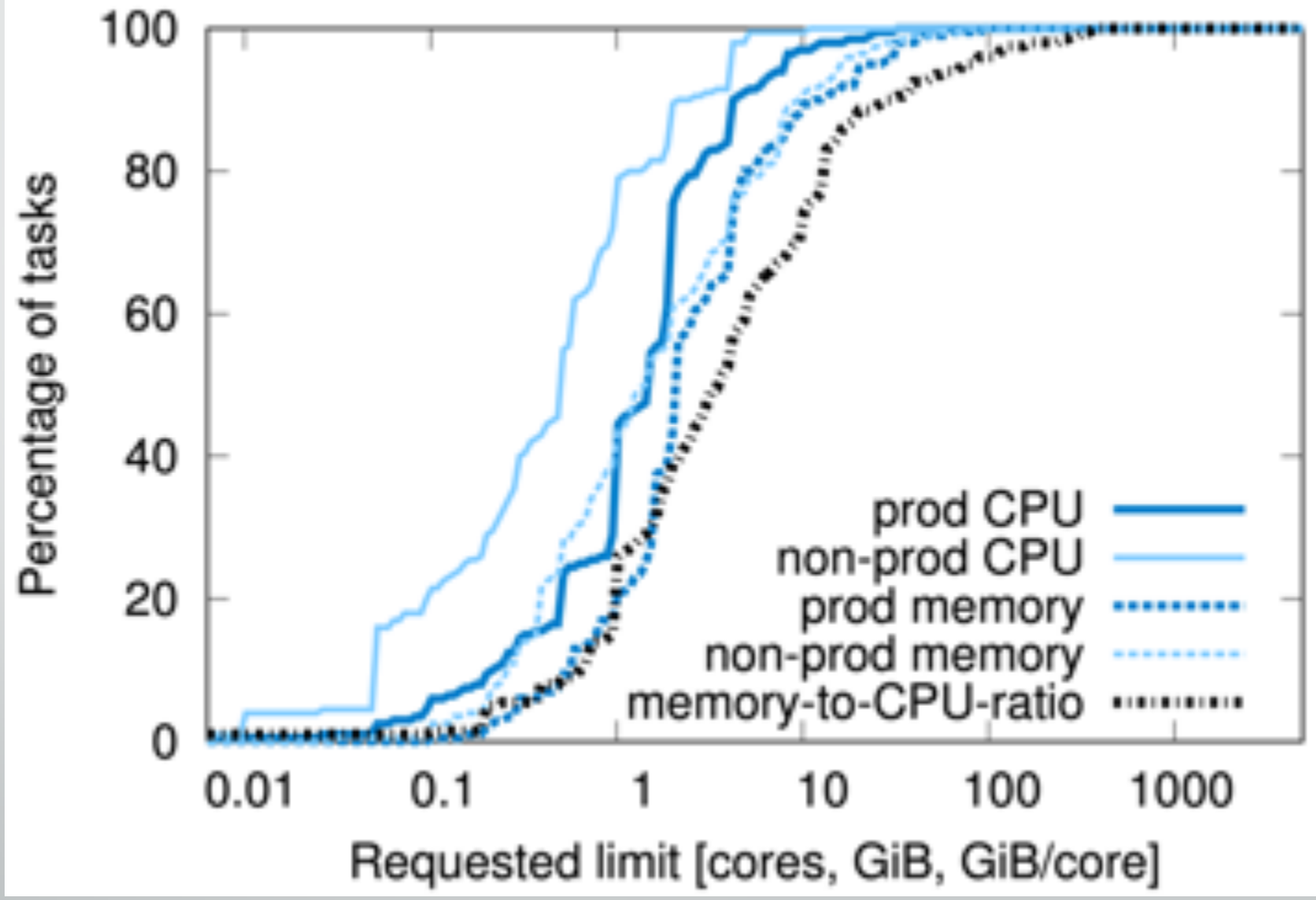


Figure 8

What happens when users choose memory and compute freely?

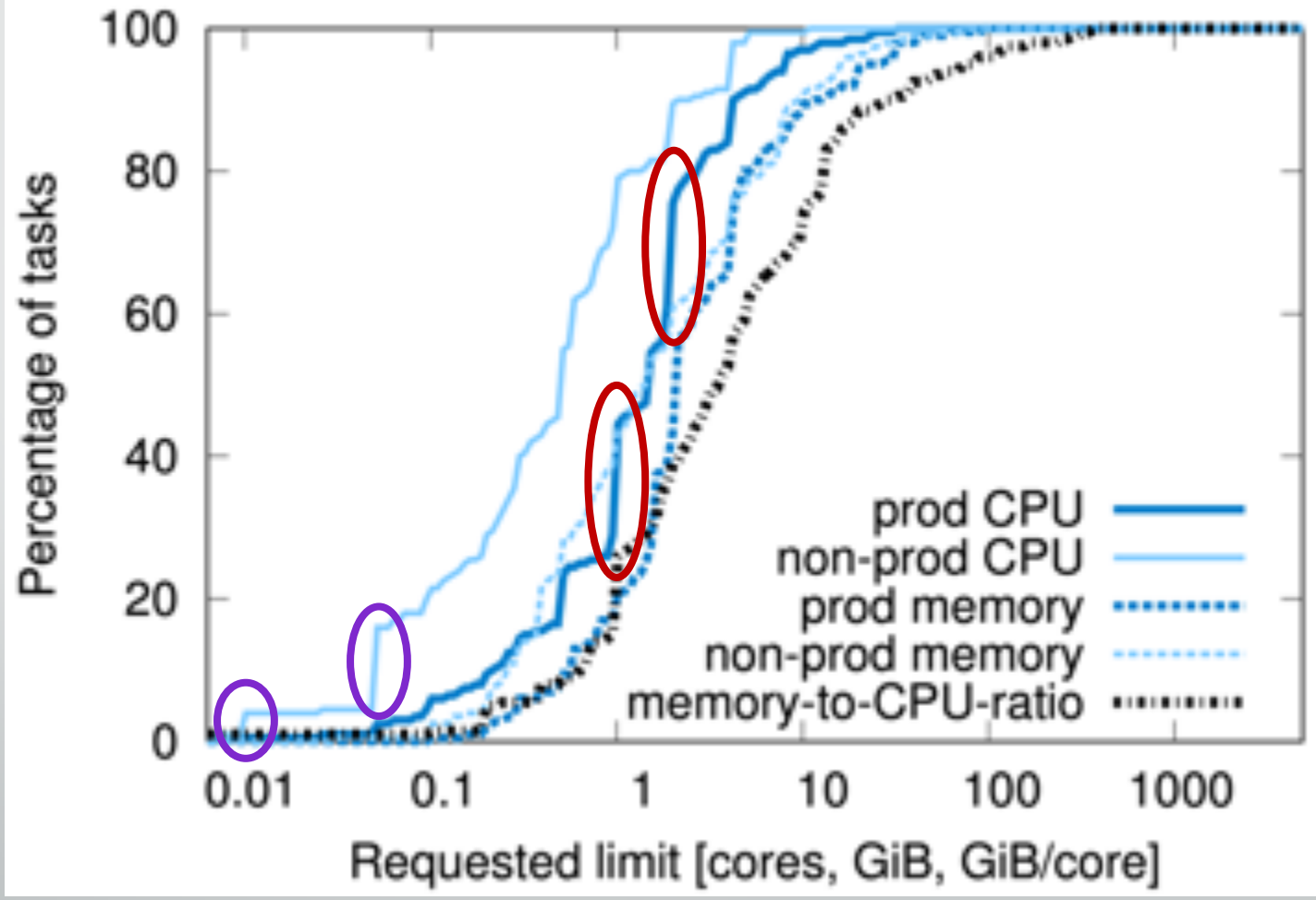


Figure 8

What happens when users choose memory and compute freely?

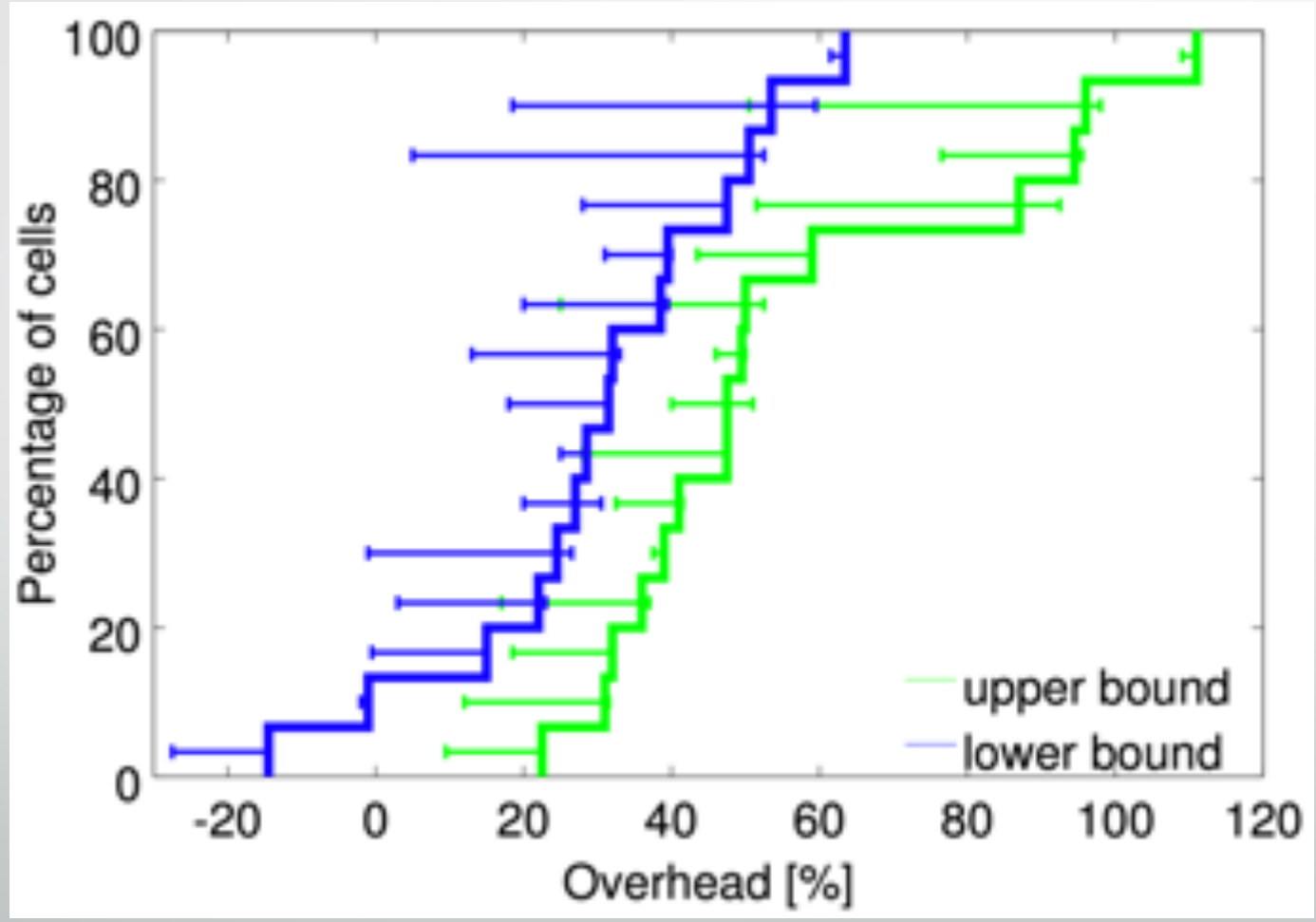


Figure 9

What happens when users choose memory and compute freely?

What happens when users choose memory and compute freely?

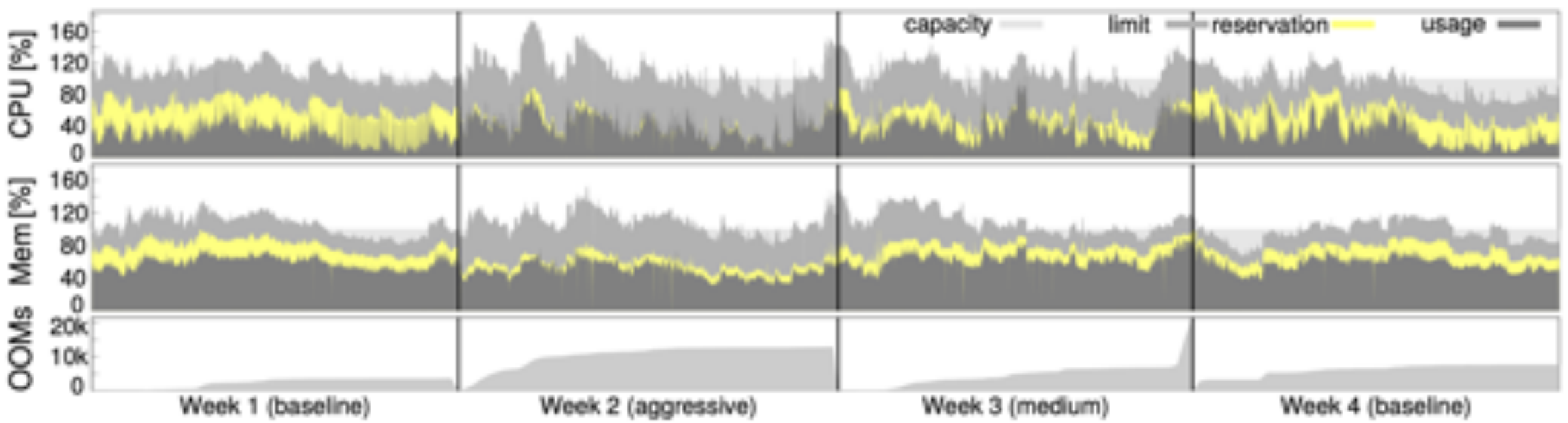
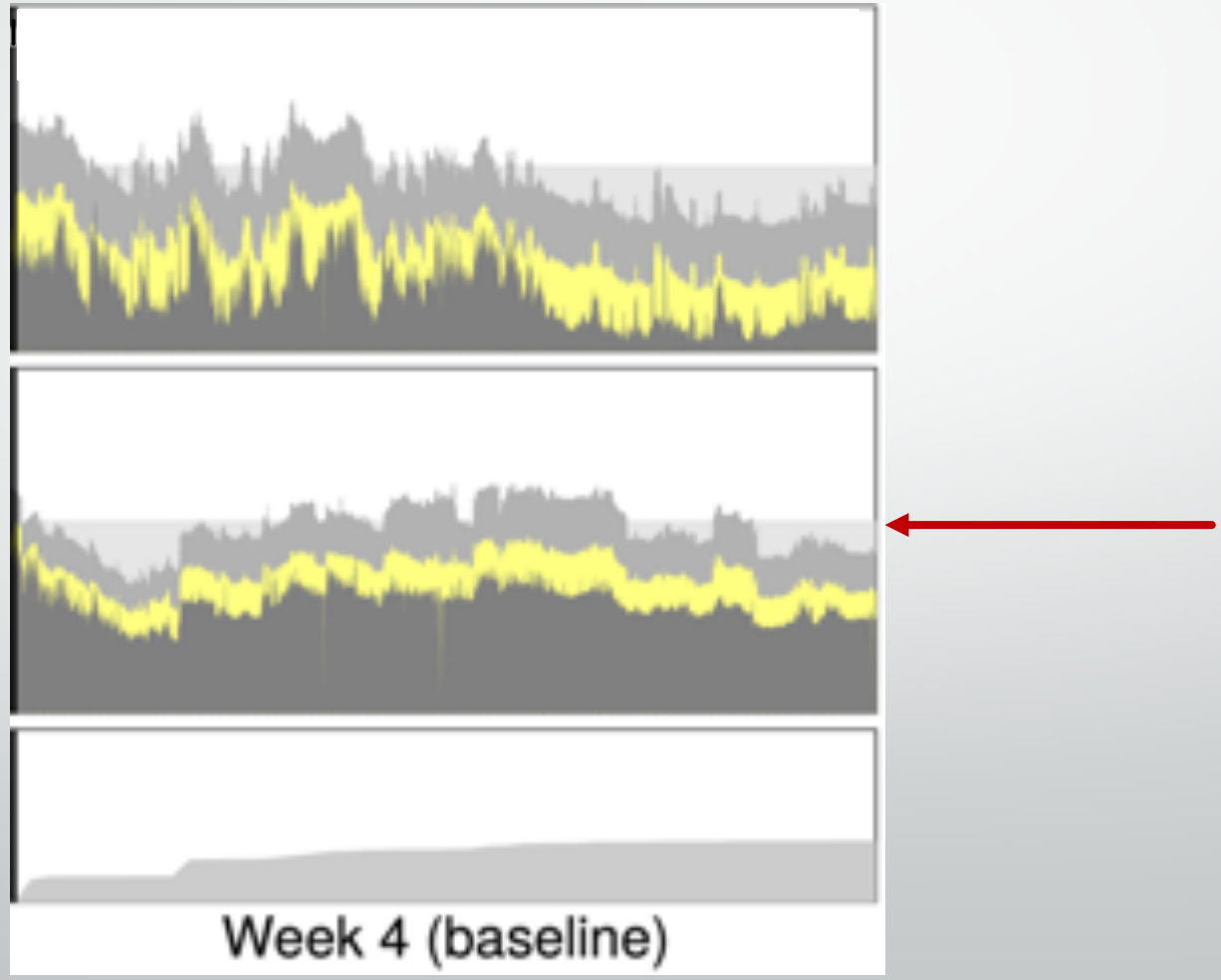
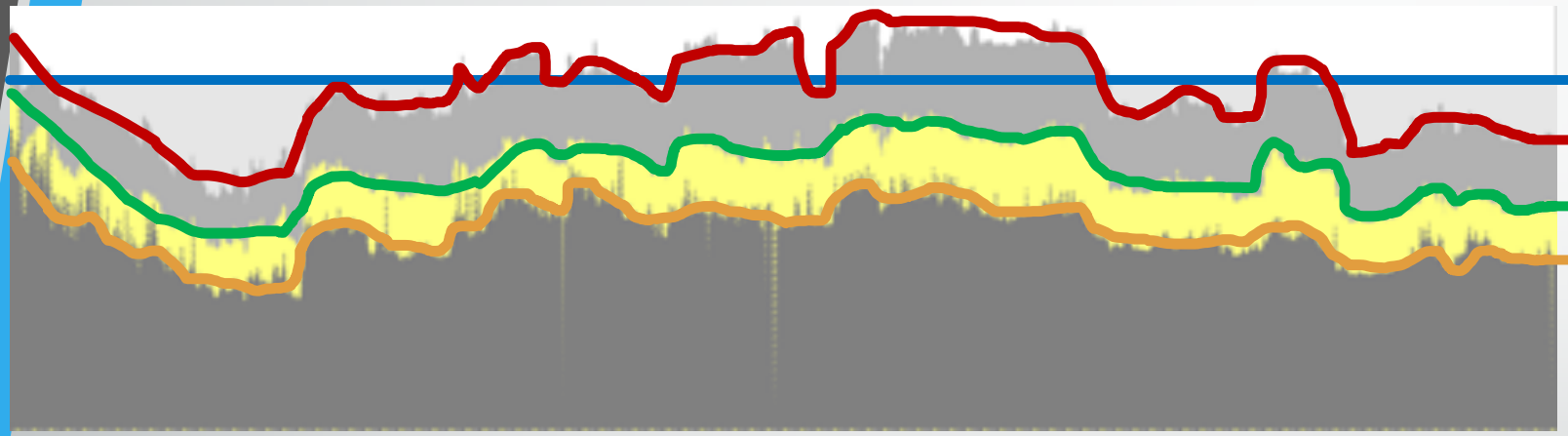


Figure 12

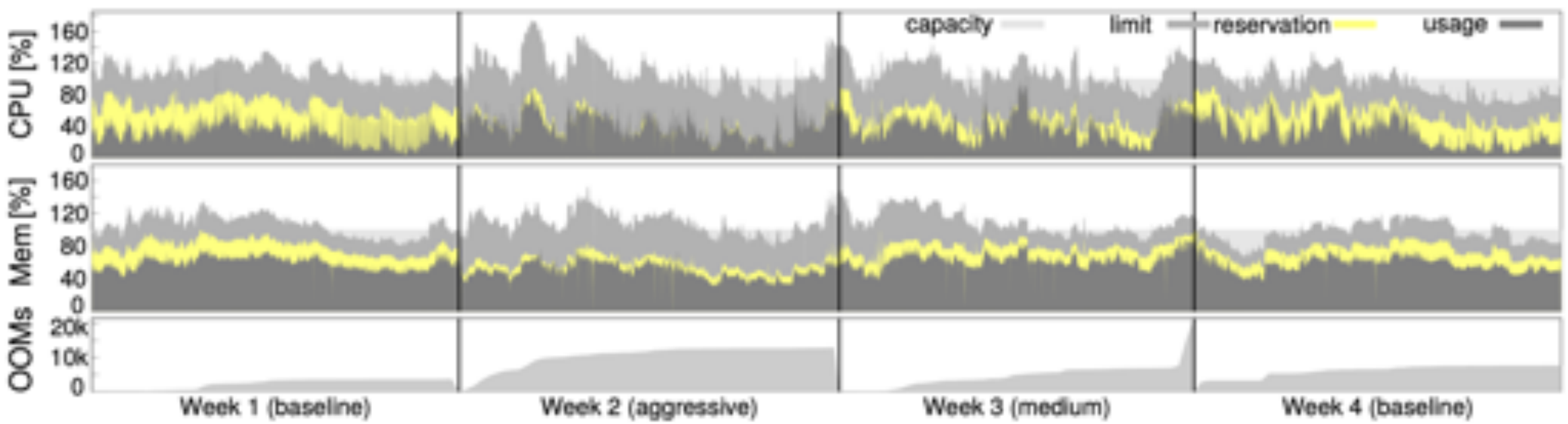
What happens when users choose memory and compute freely?



What happens when users choose memory and compute freely?



What happens when users choose memory and compute freely?



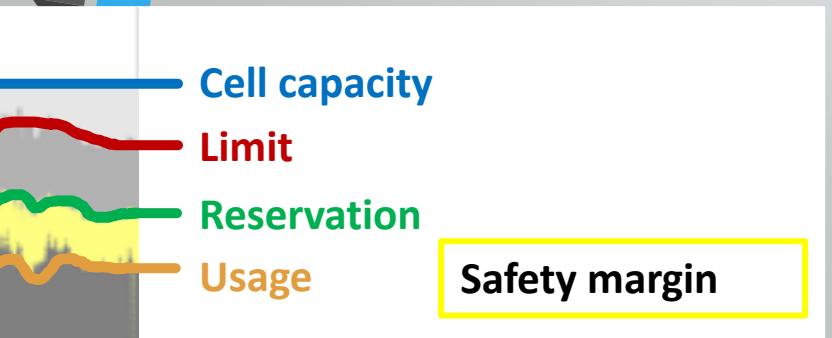
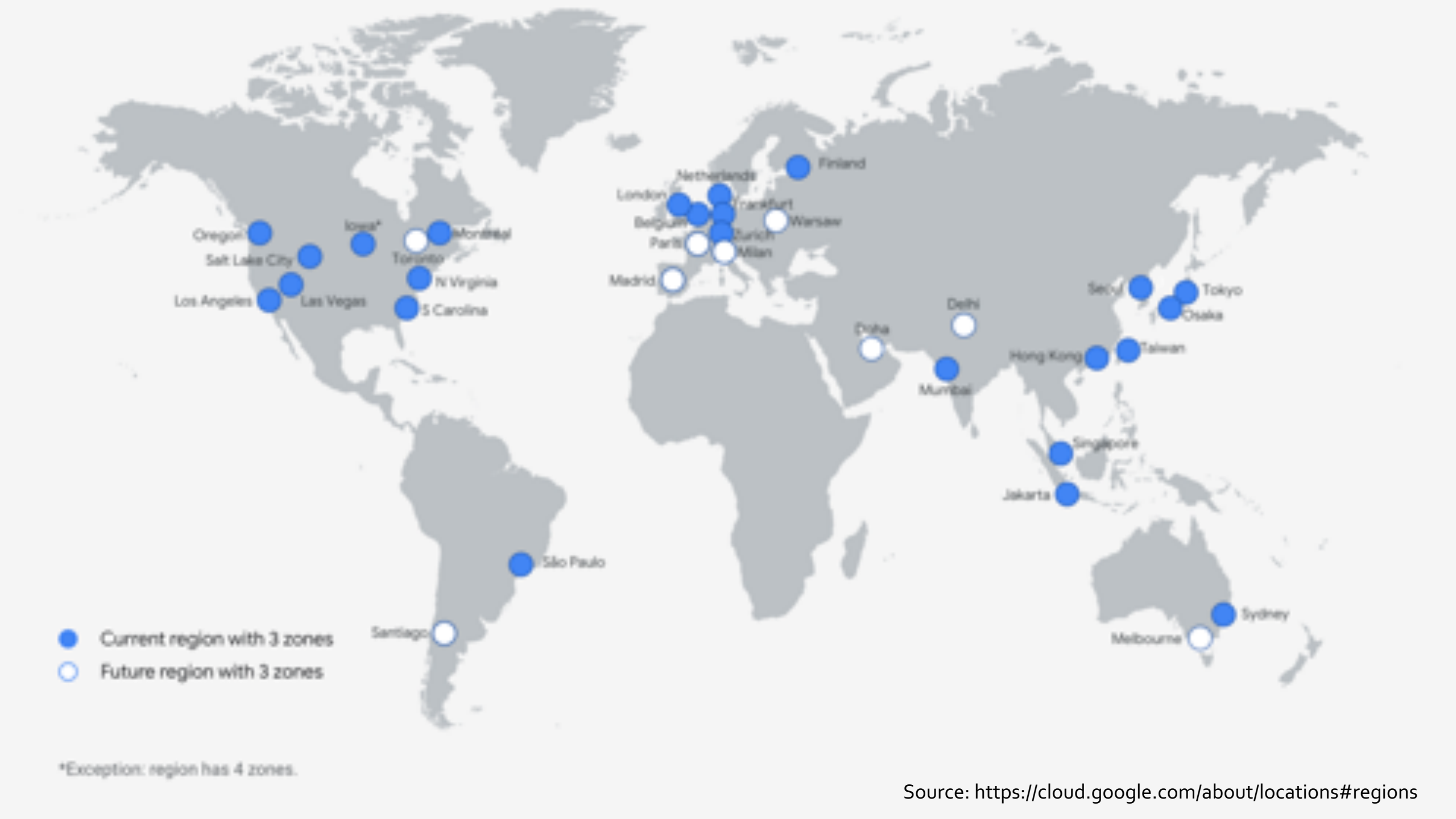


Figure 12



Source: https://cloud.google.com/about/locations#regions

The Future in Borg's Past

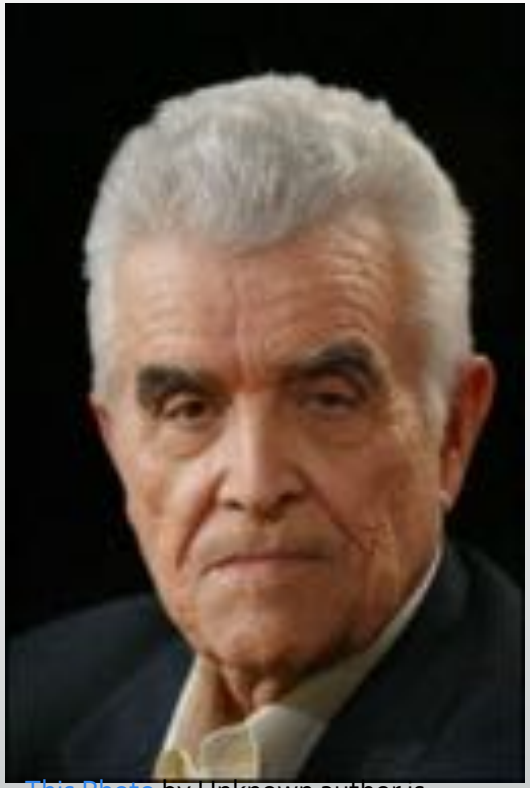
Kubernetes

- K8s is a set of tools which can be used to automate the deployment, scale, and orchestration of containers.
- K8s was released as OSS in 2014.
- Largely derived from the best practices and lessons learned from the Borg project.
- Google deploys Borg and Kubernetes internally.
- A managed Kubernetes service is available in Google Cloud Platform via Google Kubernetes Engine (GKE).

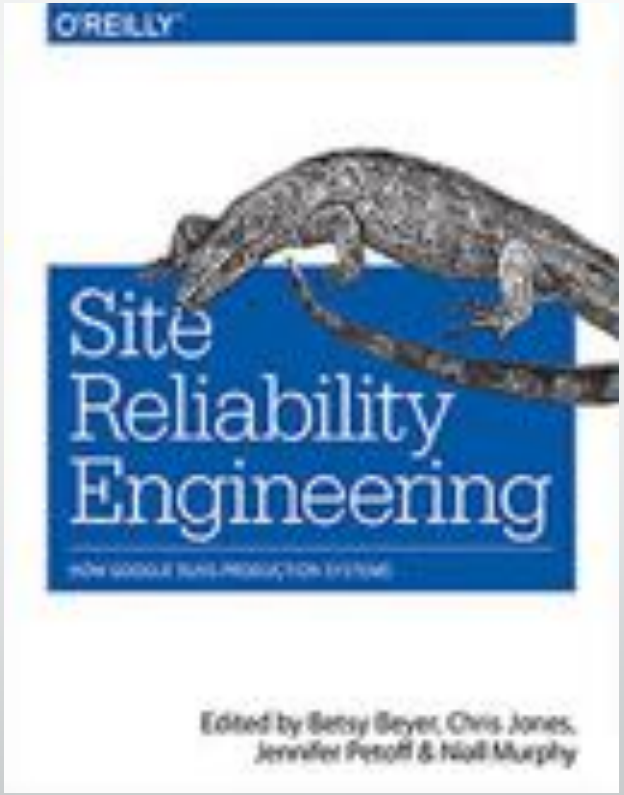
Kubernetes vs. Borg

- Kubernetes pods are largely derived from and equivalent to Borg allocs.
- Kubernetes transcends the Borg Job-Task paradigm through the labelling (key + value) of pods.
- Pod labels enable engineers to manage high-level services, instances of services, and pod subsets at a level of granularity which is not possible within the confines of the Borg Job-Task paradigm.
- Pod labelling enables significantly greater flexibility.
- Borg tasks carry the IP address of the machine they run on whereas Kubernetes pods can each allocated IP addresses.
- The scale attainable via Kubernetes deployments comes at a cost: massively complex configuration.
  - Solution: Google Kubernetes Engine (GKE) Autopilot.

René Girard and the Mimetic Theory of Desire



Site Reliability Engineering



Borg and SRE Principles

- The development and maintenance of Borg largely gave rise to the principles of the Google Site Reliability Engineering methodology.
- Core SRE Principles:
  - Embrace risk
  - Define and adhere to service level objectives
  - Eliminate toil through automation
  - Monitor distributed systems
  - The release process is an engineering problem
  - Embrace simplicity

Strengths

- SRE != SWE: abstracting away infrastructure details.
- Built from first principles to meet Google's unique availability, scale, and workload requirements.
- Clearly defined evaluation metrics.
- Thorough comparison to existing large-scale server cluster systems.

Weaknesses

- Deeply entrenched proprietary systems increase switching costs.
- Borg tasks inherit the IP address of their host machine.
- Job-Task paradigm lacks flexibility.
- Human incentive to over-estimate resource consumption.
- Little insight into the operational dimensions of Borg within Google.

Further Reading on K8s, SRE, and Philosophy

- https://kubernetes.io/
- https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/44843.pdf
- https://sre.google/sre-book/part-II-principles/
- https://sre.google/sre-book/production-environment/
- https://kubernetes.io/blog/2015/04/borg-predecessor-to-kubernetes/
- https://cloud.google.com/kubernetes-engine
- https://www.theregister.com/2021/02/25/google_kubernetes_autopilot/
- https://iep.utm.edu/girard/

Sources

- Google Site / Datacenter: https://www.datacenterknowledge.com/google-alphabet/google-spend-11-billion-new-data-centers-netherlands
- Google Datacenter Cluster: https://www.datacenterknowledge.com/archives/2012/10/17/google-reveals-its-data-centers
- Borg Architecture: https://www.nextplatform.com/2015/05/05/google-omega-to-become-part-of-borg-collective/
- Cluster Management at Google with Borg, John Wilkes: https://youtu.be/0W49z8hVnok

Q & A