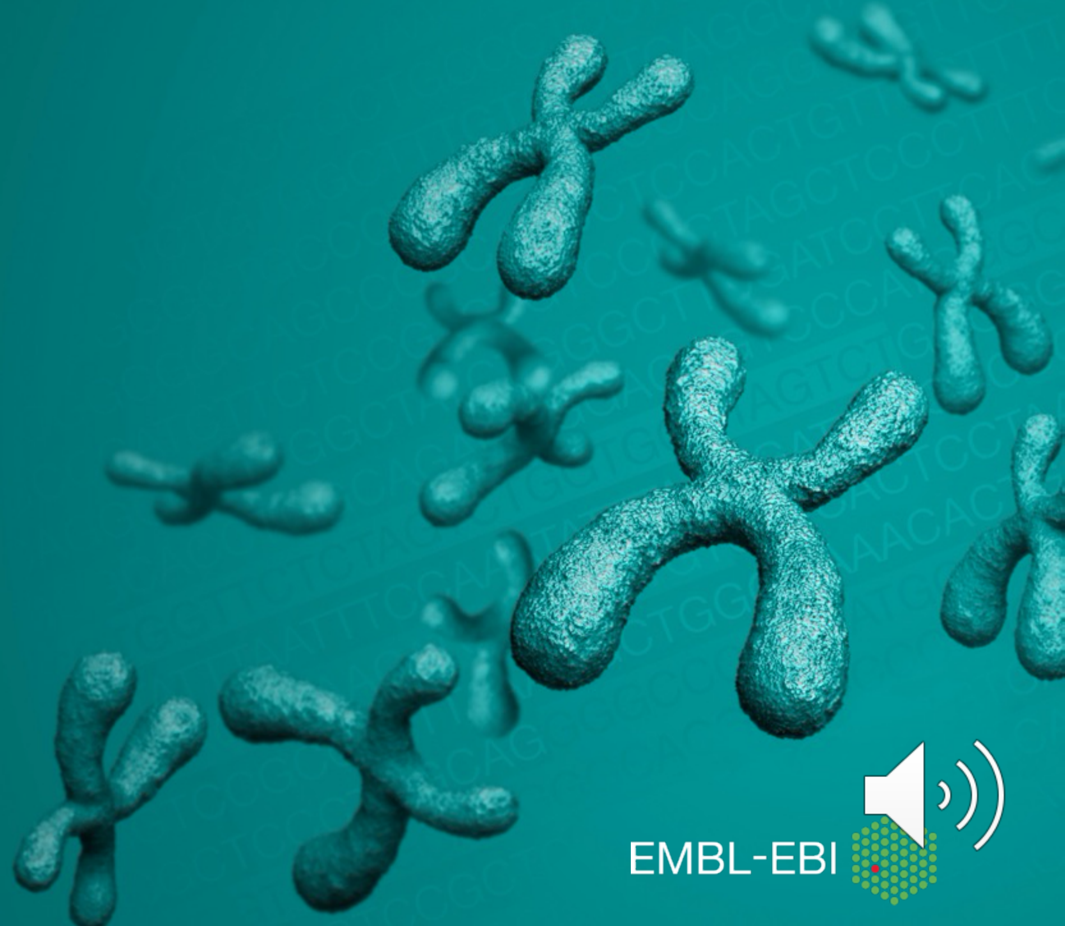
Porting Research Pipelines into Clouds

Architectural considerations (3/3)



David Yuan, Ph.D.

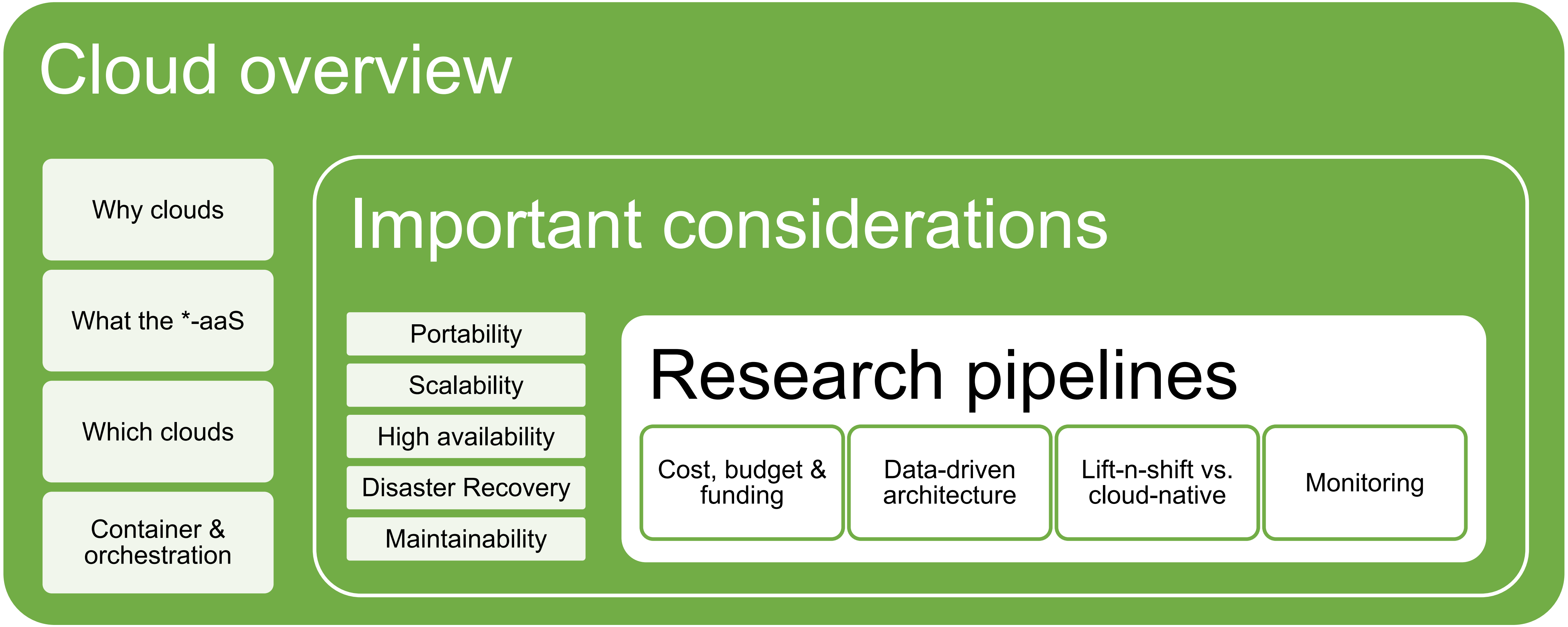
Cloud Bioinformatics Application Architect

Technology and Science Integration

European Bioinformatics Institute, EMBL

EMBL-EBI

Porting into clouds



Cloud overview

- Why clouds
- What the *-aaS
- Which clouds
- Container & orchestration

Important considerations

- Portability
- Scalability
- High availability
- Disaster Recovery
- Maintainability

Research pipelines

- Cost, budget & funding
- Data-driven architecture
- Lift-n-shift vs. cloud-native
- Monitoring

Cost, budget & funding

Current situation

Cloud requirement

Advises

- Research pipelines are usually funded by research grants.
- Funding agency is OK with capital cost but generally do not allow operational cost.
- Pipeline operators generally do not track usage metrics. There is little information to start estimating the cost in the cloud.

Cost, budget & funding

Current situation

Cloud requirement

Advises

Current situation

- Research pipelines are usually funded by research grants.
- Funding agency is OK with capital cost but generally do not allow operational cost.
- Pipeline operators generally do not track usage metrics. There is little information to start estimating the cost in the cloud.

Cloud requirement

- Cloud deployments can outlive 3 – 5 year funding period.
- Public cloud requires little capital investment.
- Cloud providers charge by usage:
  - CPU cycles, active connections, ingress, egress, memory consumption, disk space used and duration, etc.
- Different cloud providers charge very different prices
  - Constantly changing

Cost, budget & funding

Current situation

- Research pipelines are usually funded by research grants.
- Funding agency is OK with capital cost but generally do not allow operational cost.
- Pipeline operators generally do not track usage metrics. There is little information to start estimating the cost in the cloud.

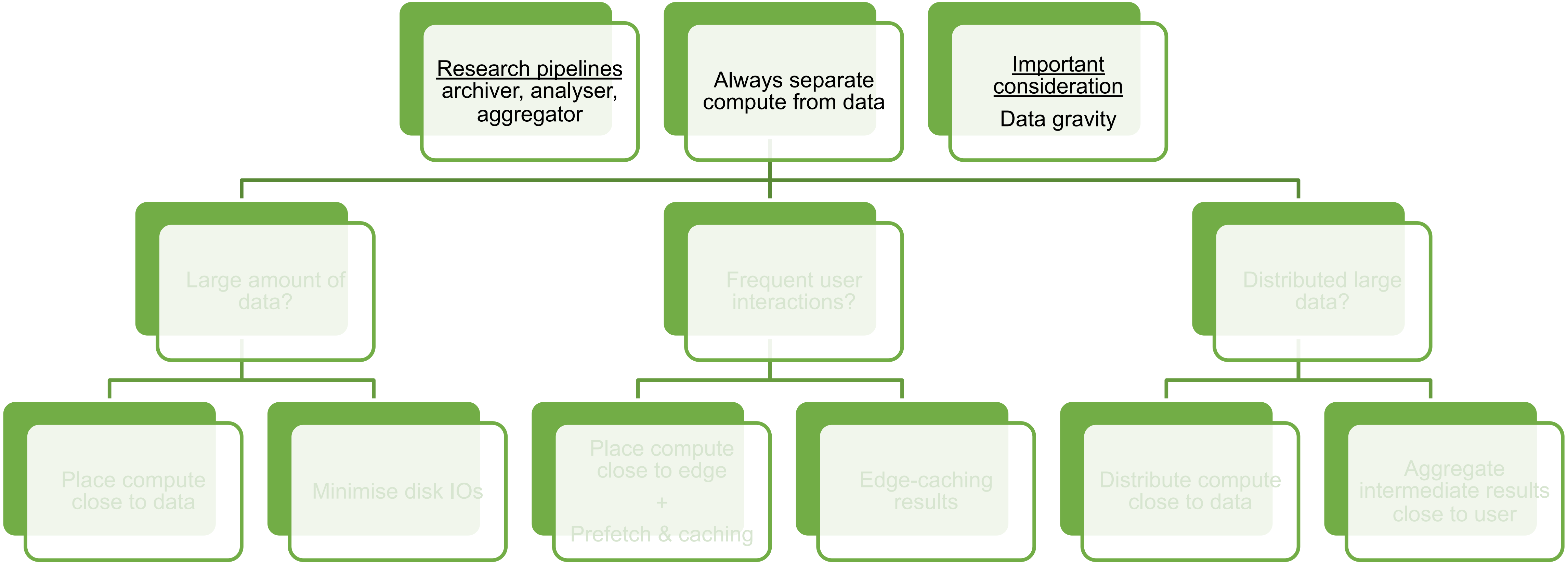
Cloud requirement

- Cloud deployments can outlive 3 – 5 year funding period.
- Public cloud requires little capital investment.
- Cloud providers charge by usage:
    - CPU cycles, active connections, ingress, egress, memory consumption, disk space used and duration, etc.
- Different cloud providers charge very different prices
    - Constantly changing

Advises

- When choosing a cloud platform
    - Go cloud-native to maximize benefit and to minimize cost
    - Take potential funding and cost issues into consideration
    - Shop around – private or public clouds
- To avoid vendor lock-in
    - Ensure portability if technically possible
- To estimate operational cost
    - Compile usage metrics
    - Benchmark / profile pipelines

Data-driven architecture for research pipelines



Research pipelines archiver, analyser, aggregator

Always separate compute from data

Important consideration

Data gravity

Large amount of data?

Frequent user interactions?

Distributed large data?

Place compute close to data

Minimise disk IOs

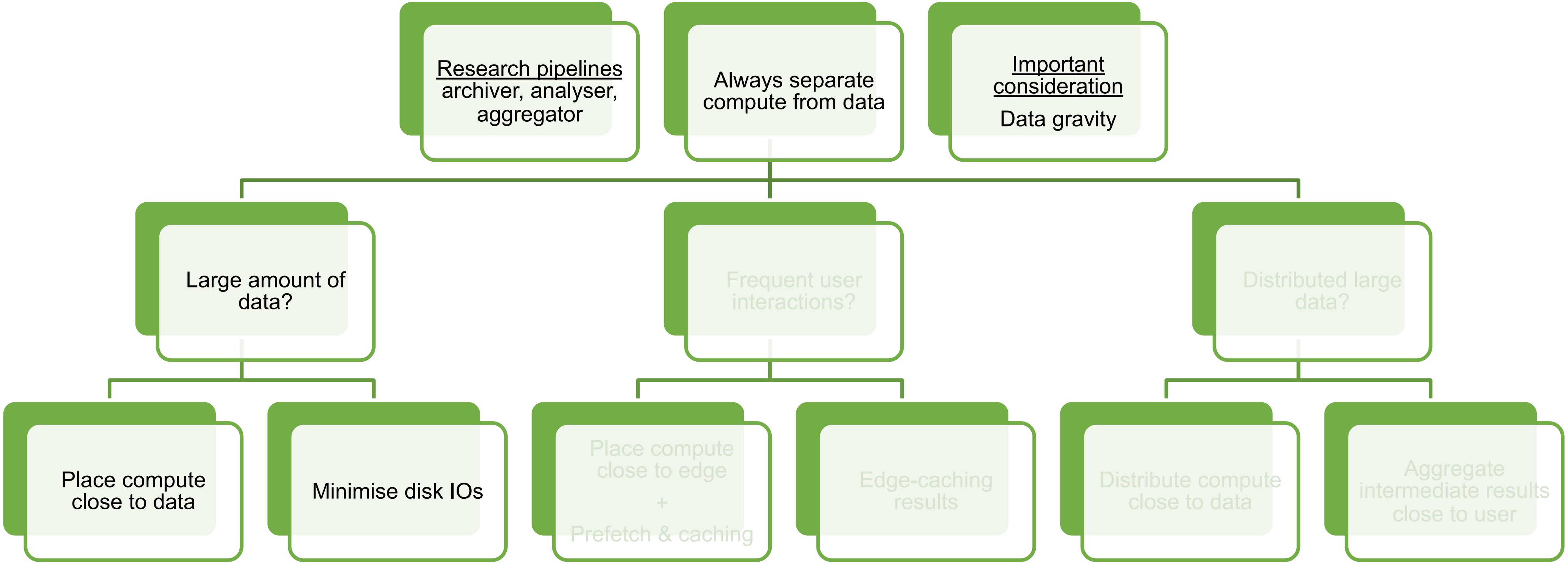
Place compute close to edge
+
Prefetch & caching

Edge-caching results

Distribute compute close to data

Aggregate intermediate results close to user

Data-driven architecture for research pipelines



Research pipelines archiver, analyser, aggregator

Always separate compute from data

Important consideration

Data gravity

Large amount of data?

Frequent user interactions?

Distributed large data?

Place compute close to data

Minimise disk IOs

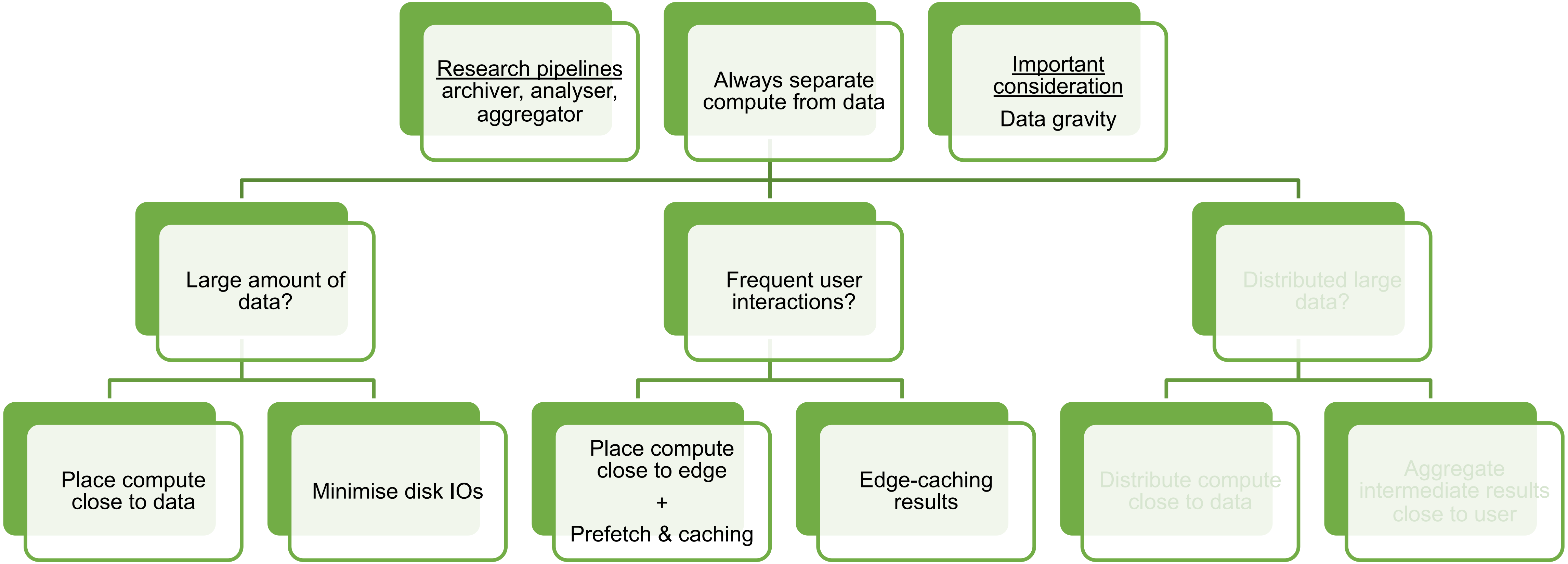
Place compute close to edge + Prefetch & caching

Edge-caching results

Distribute compute close to data

Aggregate intermediate results close to user

Data-driven architecture for research pipelines



Research pipelines
archiver, analyser, aggregator

Always separate compute from data

Important consideration
Data gravity

Large amount of data?

Place compute close to data

Minimise disk IOs

Frequent user interactions?

Place compute close to edge
+
Prefetch & caching

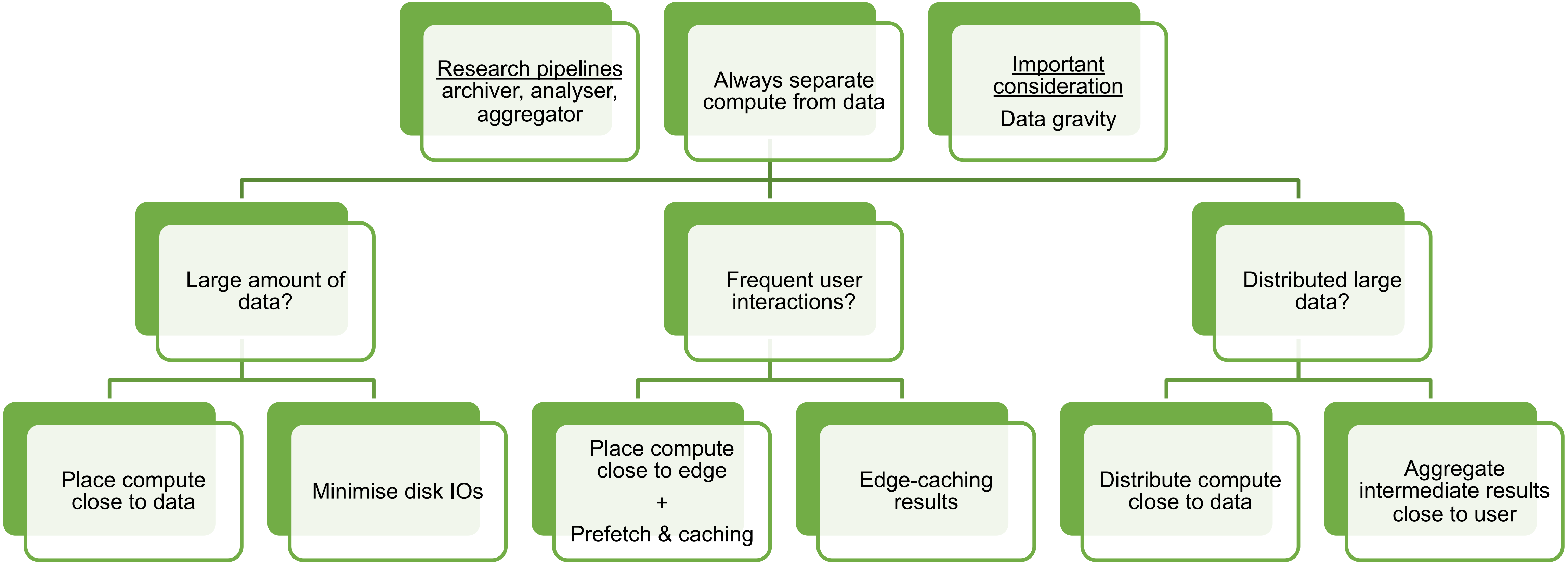
Edge-caching results

Distributed large data?

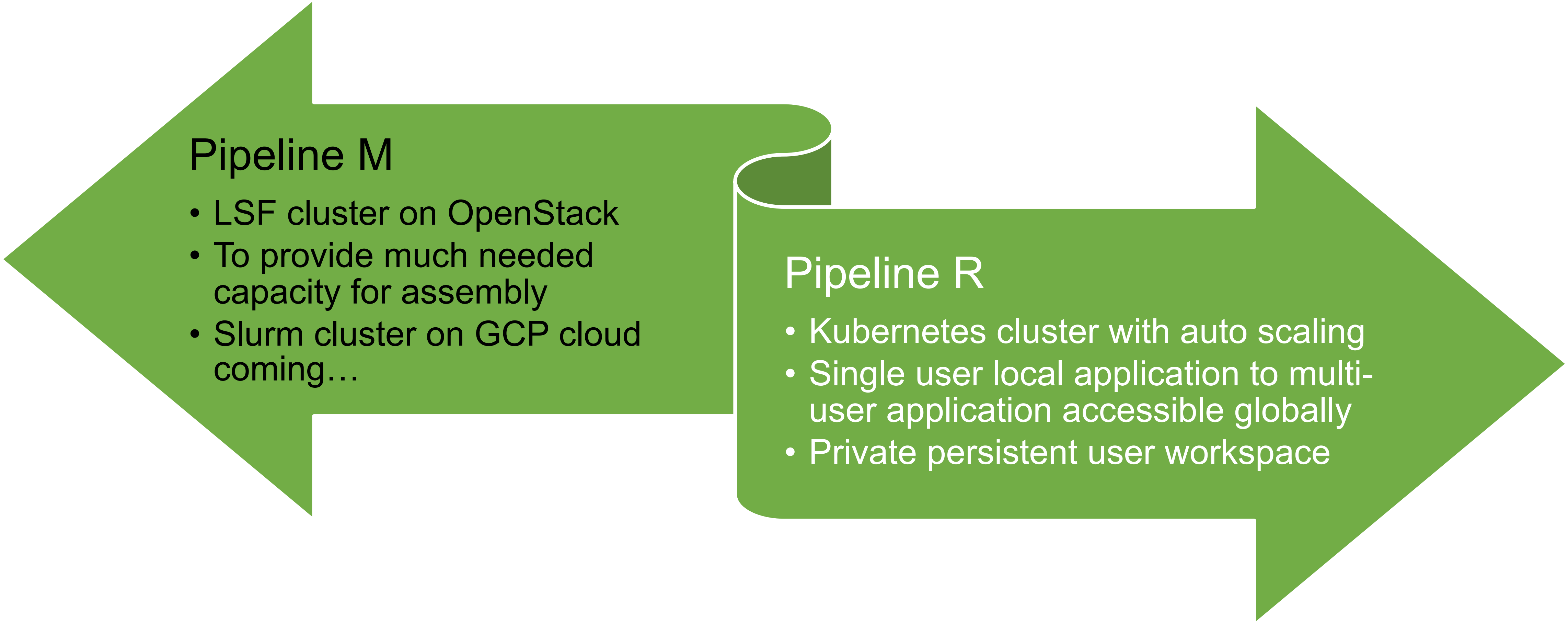
Distribute compute close to data

Aggregate intermediate results close to user

Data-driven architecture for research pipelines


Research pipelines
archiver, analyser, aggregator
Always separate compute from data
Important consideration
Data gravity
Large amount of data?
Frequent user interactions?
Distributed large data?
Place compute close to data
Minimise disk IOs
Place compute close to edge
+
Prefetch & caching
Edge-caching results
Distribute compute close to data
Aggregate intermediate results close to user

Lift-n-shift vs. cloud-native



Pipeline M

- LSF cluster on OpenStack
- To provide much needed capacity for assembly
- Slurm cluster on GCP cloud coming…

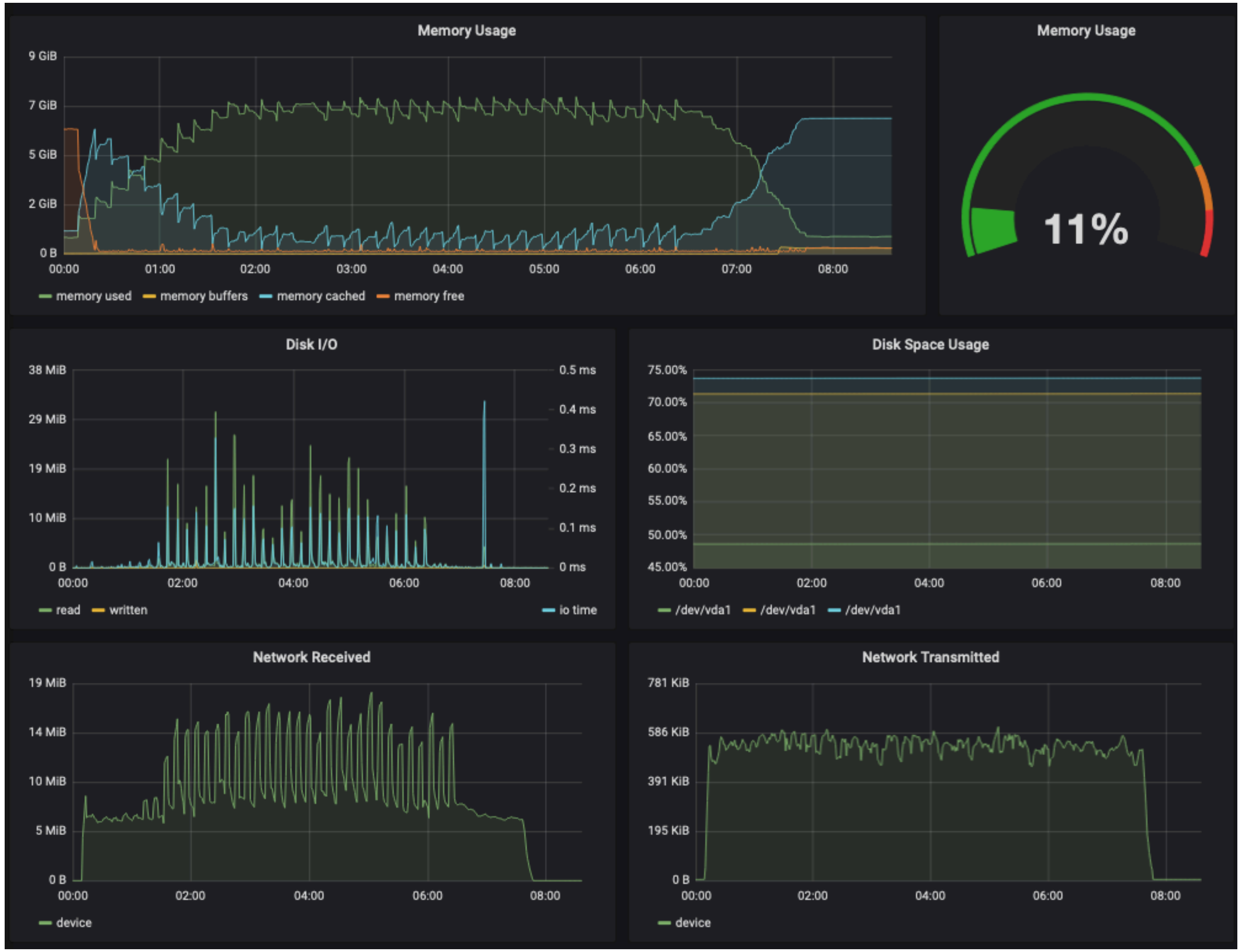
Pipeline R

- Kubernetes cluster with auto scaling
- Single user local application to multi-user application accessible globally
- Private persistent user workspace

Monitoring

Never flying blind

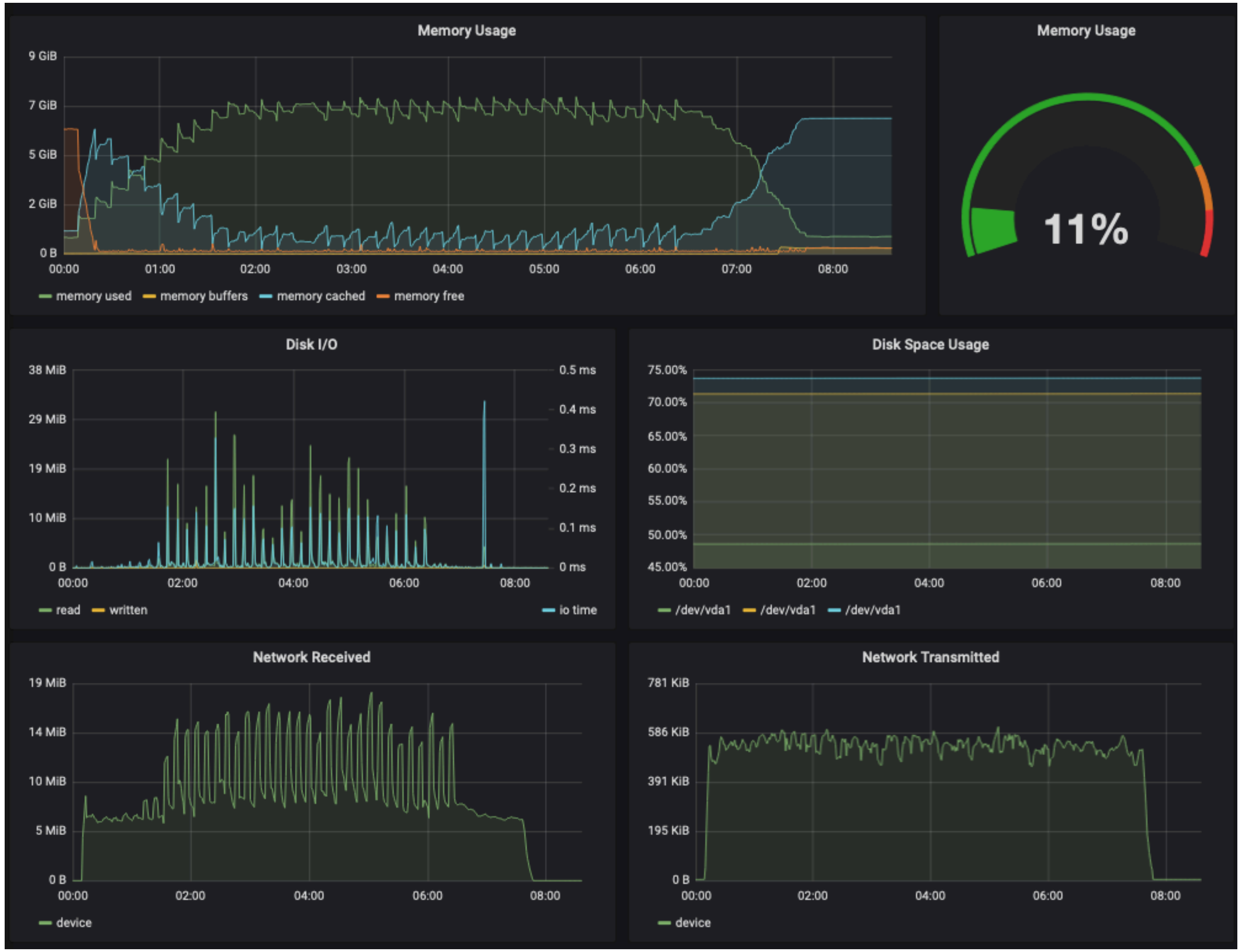
- Monitoring on pipelines is generally lacking
- K8S can be monitored with Prometheus + Grafana
- Kubernetes Dashboard is highly recommended
- Elasticsearch can be considered for K8S in multiple clouds



Monitoring

Never flying blind

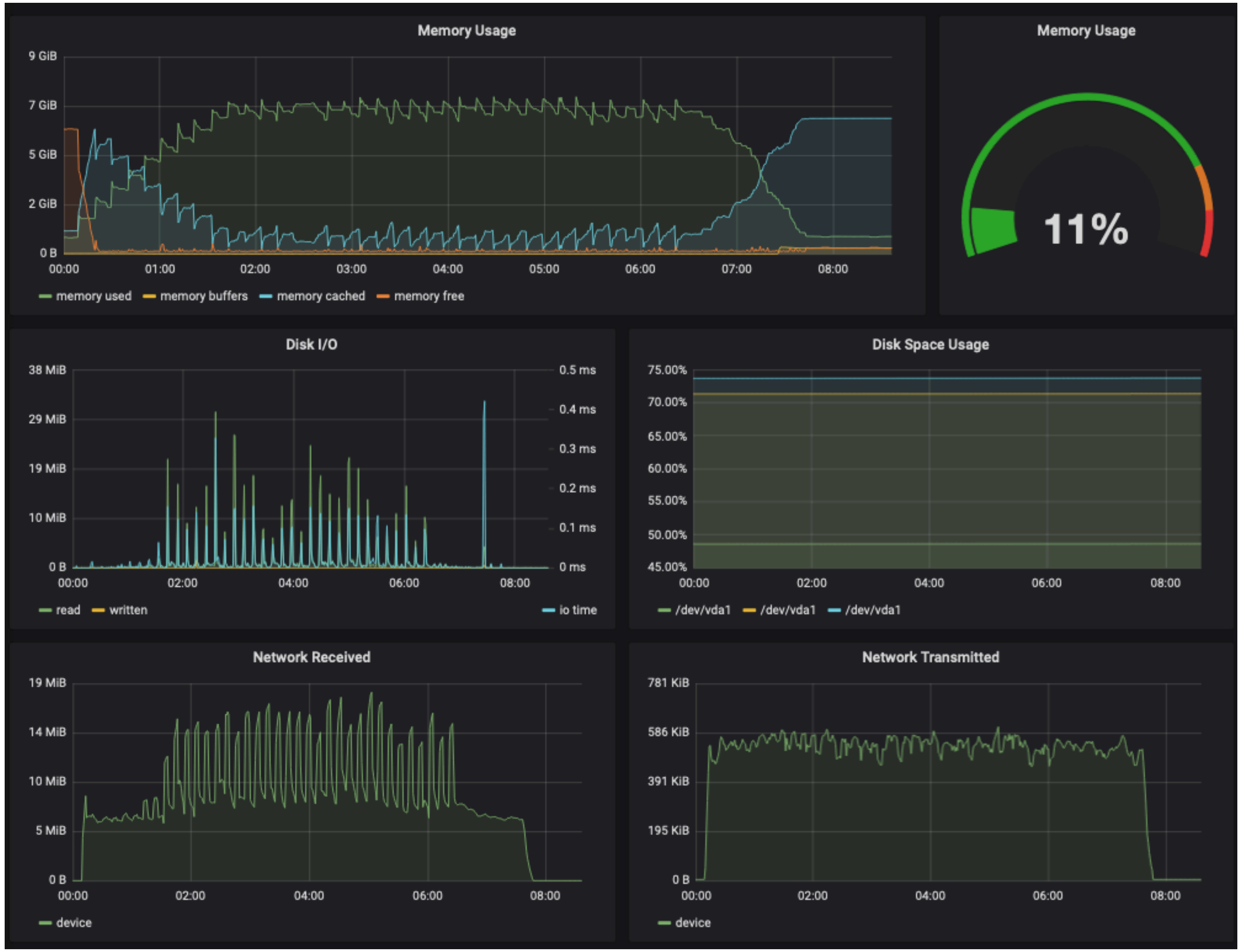
- Monitoring on pipelines is generally lacking
- K8S can be monitored with Prometheus + Grafana
- Kubernetes Dashboard is highly recommended
- Elasticsearch can be considered for K8S in multiple clouds



Monitoring

Never flying blind

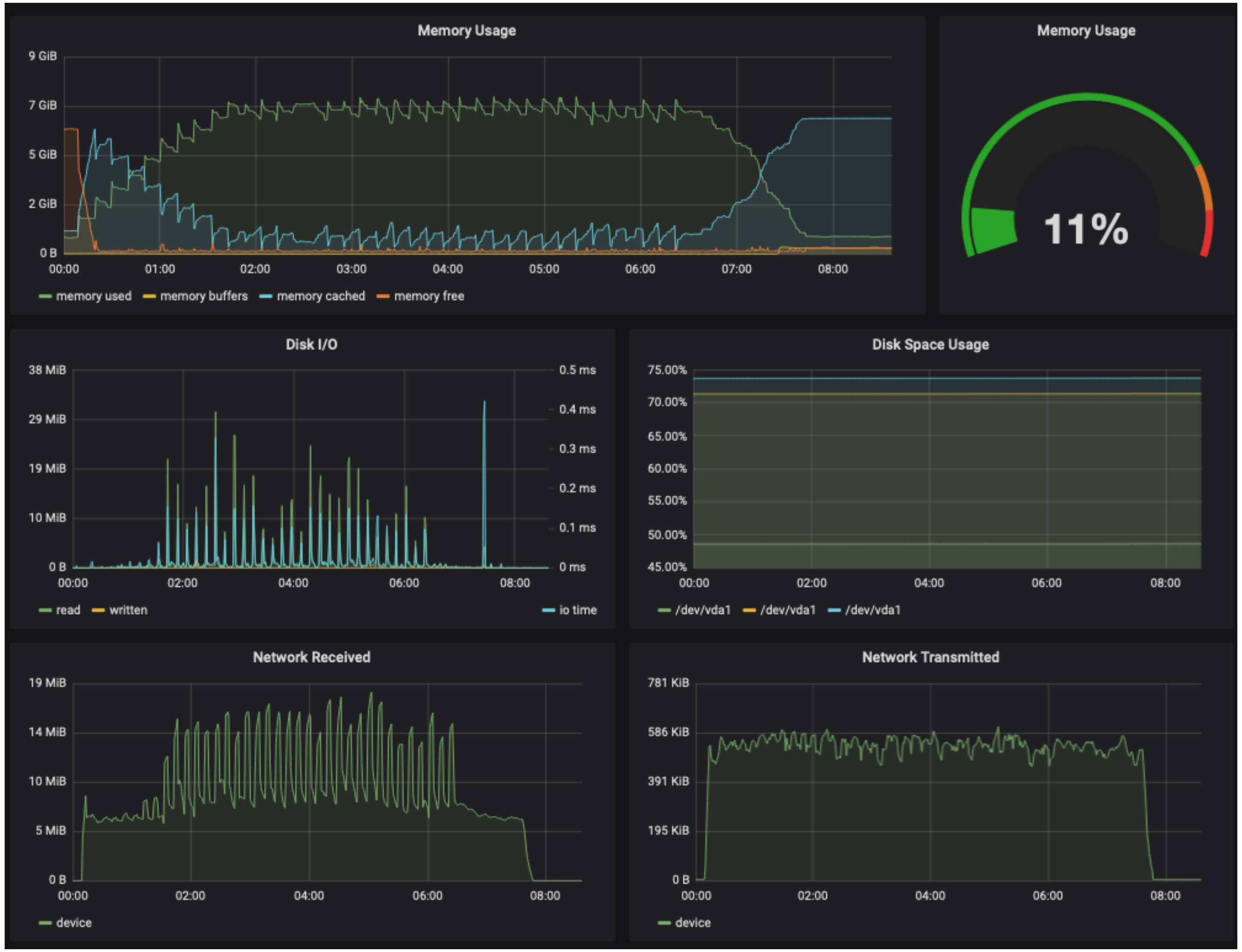
- Monitoring on pipelines is generally lacking
- K8S can be monitored with Prometheus + Grafana
- Kubernetes Dashboard is highly recommended
- Elasticsearch can be considered for K8S in multiple clouds



Monitoring

Never flying blind

- Monitoring on pipelines is generally lacking
- K8S can be monitored with Prometheus + Grafana
- Kubernetes Dashboard is highly recommended
- Elasticsearch can be considered for K8S in multiple clouds



Summary

- Overview of porting into clouds
  - Why, what, which & how – particularly container & K8S
- Important considerations and why Kubernetes
  - Portability, scalability, high availability, disaster recovery & maintainability
- Special considerations for research pipelines
  - Cost budget & funding, data-driven architecture, lift-n-shift vs. cloud-native & monitoring
- Contact us
  - https://bit.ly/cc-doc
  - cloud-consultants@ebi.ac.uk