Porting Research Pipelines into Clouds

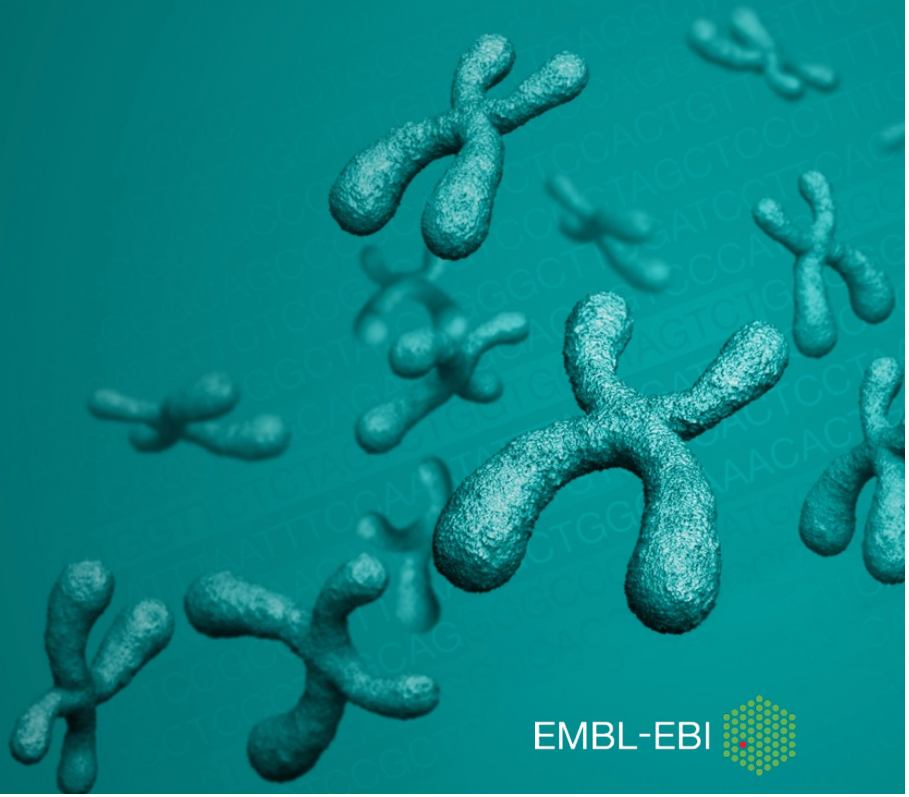
Architectural considerations

David Yuan, Ph.D.

Cloud Bioinformatics Application Architect

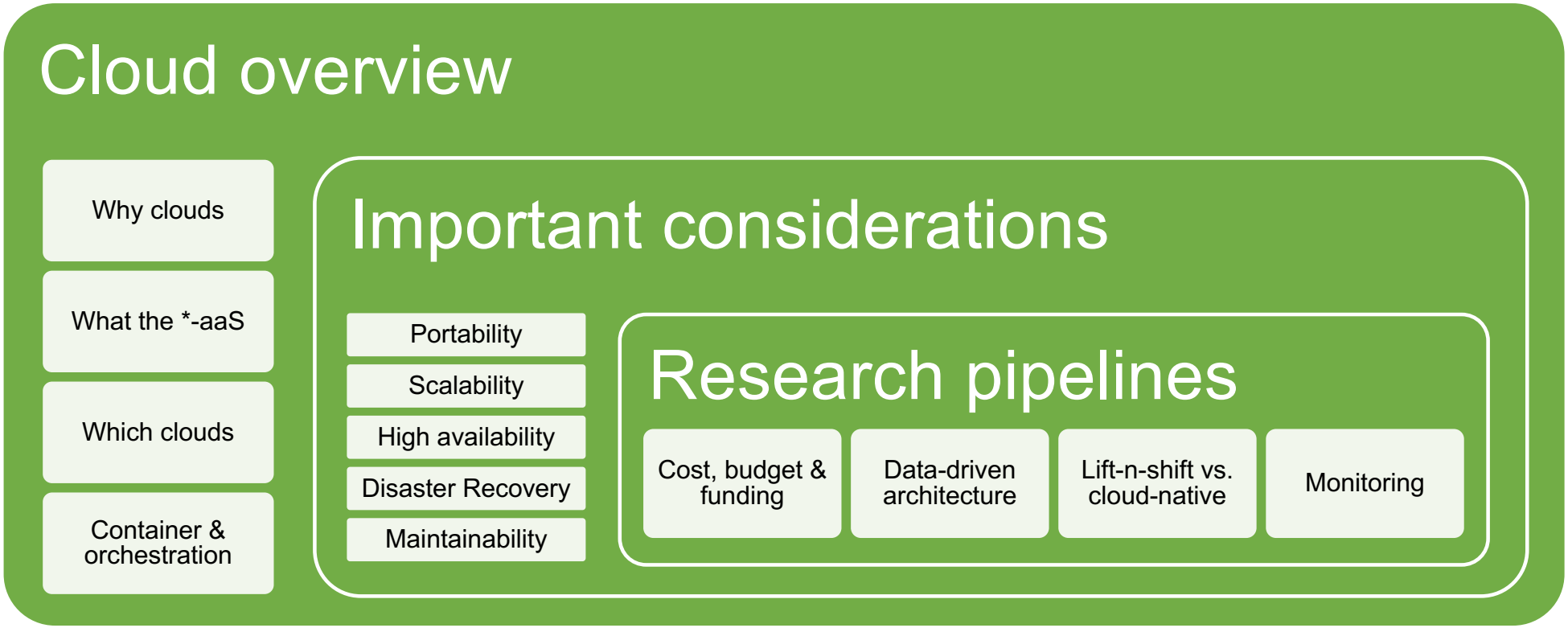
Technology and Science Integration

European Bioinformatics Institute, EMBL



EMBL-EBI

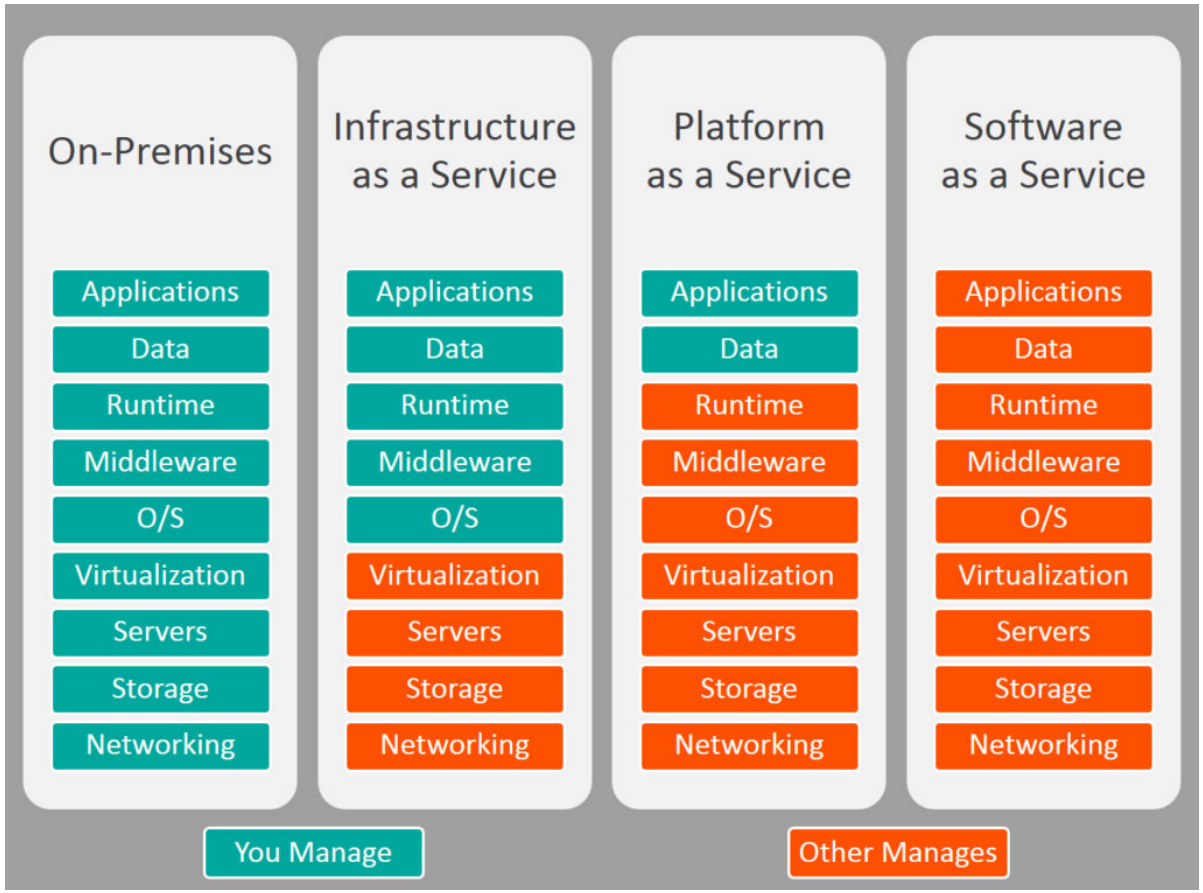
Porting into clouds



Cloud overview – why?

Research pipelines	Good candidates for the cloud!	Pros	Cons
• Archive of sequence data, images, publications or ontology information • Pipelines to analyse data • Services to aggregate other research tools or databases	• You know your pinch-points. • Cloud is mature and fast-evolving. • Lift-n-shift is possible. • Being cloud-native provides benefit way over cost.	• Stable infrastructure • Global collaboration by default • Flexible resource management • Potential cost reduction • Latest and greatest technology stack	• Accounting model is very different. • The whole field is still growing. • Beginners often face steep learning curves.

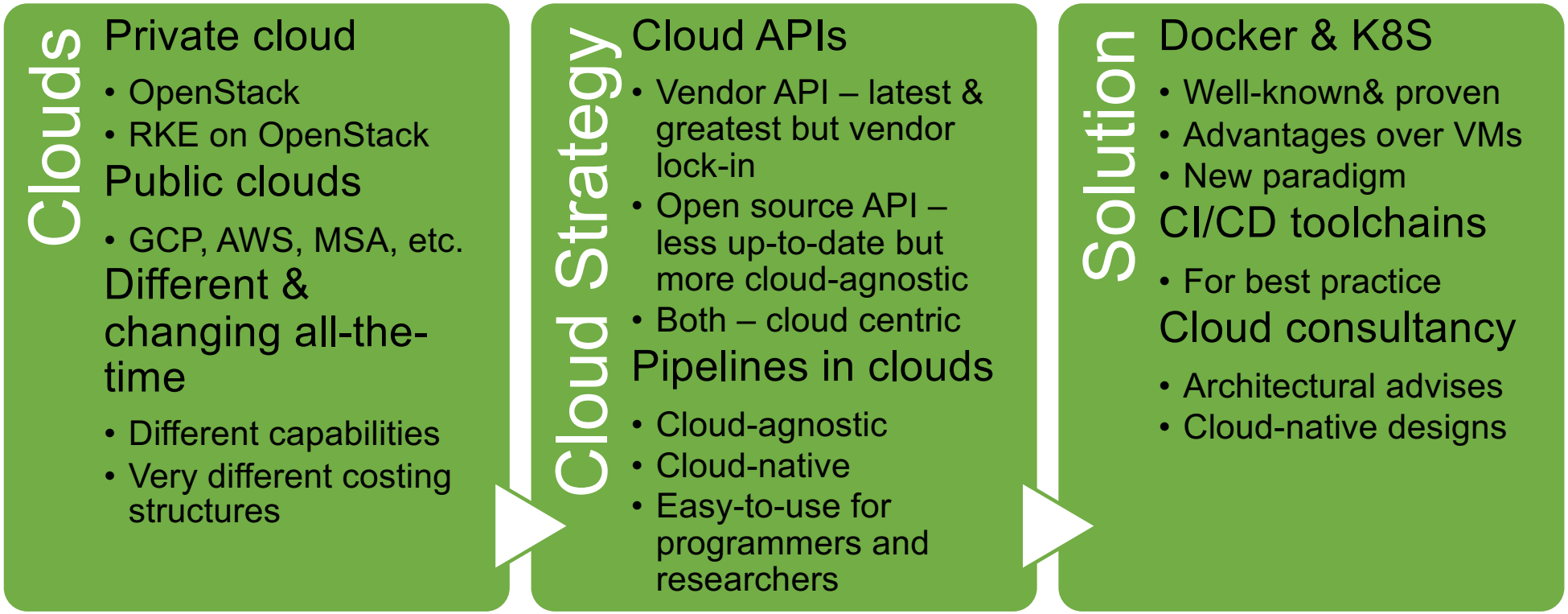
Cloud overview – what?



*** - as a Service**

- Infrastructure as a Service (IaaS)
  - OpenStack
    - GCP, AWS, MSA
  - RKE on OpenStack
    - GKE, EKS, AKS
- Platform as a Service (PaaS)
  - AWS Lambda
  - Azure App Service
- Software as a Service (SaaS)
  - AWS Route53
  - Oracle Autonomous Data Warehouse Cloud

Cloud overview – which?



Cloud overview – container & orchestration

Docker & Kubernetes

- De-facto standards of runtime & orchestration
- Docker
  - Runtime architecture
  - Packaging tool
- Kubernetes
  - Orchestration engine

Benefit over VMs

- Light-weight
- Very high portability
- Seamless integration with CI/CD
- Across hardware boundaries
- Portability, scalability, high availability, disaster recovery & maintainability

Growing pains

- More difficult to use
- Dependent on VMs in some clouds
- Tricky integration with POSIX filesystems

Best practices

- KISS principle
- Security
  - Official Docker images
  - Non-root ID
- Compute, data & configuration
  - Stateless container
  - Data on storage
  - StatefulSet for configuration

Important considerations

Portability

- Poor portability between clouds
- Docker & K8S: De-facto standards
- Major decision to be made as early as possible

Scalability

- Cloud scaling up and scaling down limited by hardware
- Docker & K8S: vertical scaling, horizontal scaling, autoscaling across hardware boundary
- Storage IO often being the bottleneck

High availability

- Cloud better than traditional DCs
- K8S: ReplicaSet & StatefulSet across hardware boundary
- Shared POSIX filesystems: single point of failure

Disaster Recovery

- Double or triple redundancy: resilient to disaster
- Infrastructure-as-code: faster recovery
- K8S: clear separation of compute, configuration and data
- Shared POSIX filesystems: single point of failure

Maintainability

- Cloud usually no scheduled downtime
- K8S: eliminating scheduled downtime
  - Rolling up upgrade K8S nodes, underlying hardware, application
  - Auto-recovery built in

Cost, budget & funding

Current situation

- Genomic 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.

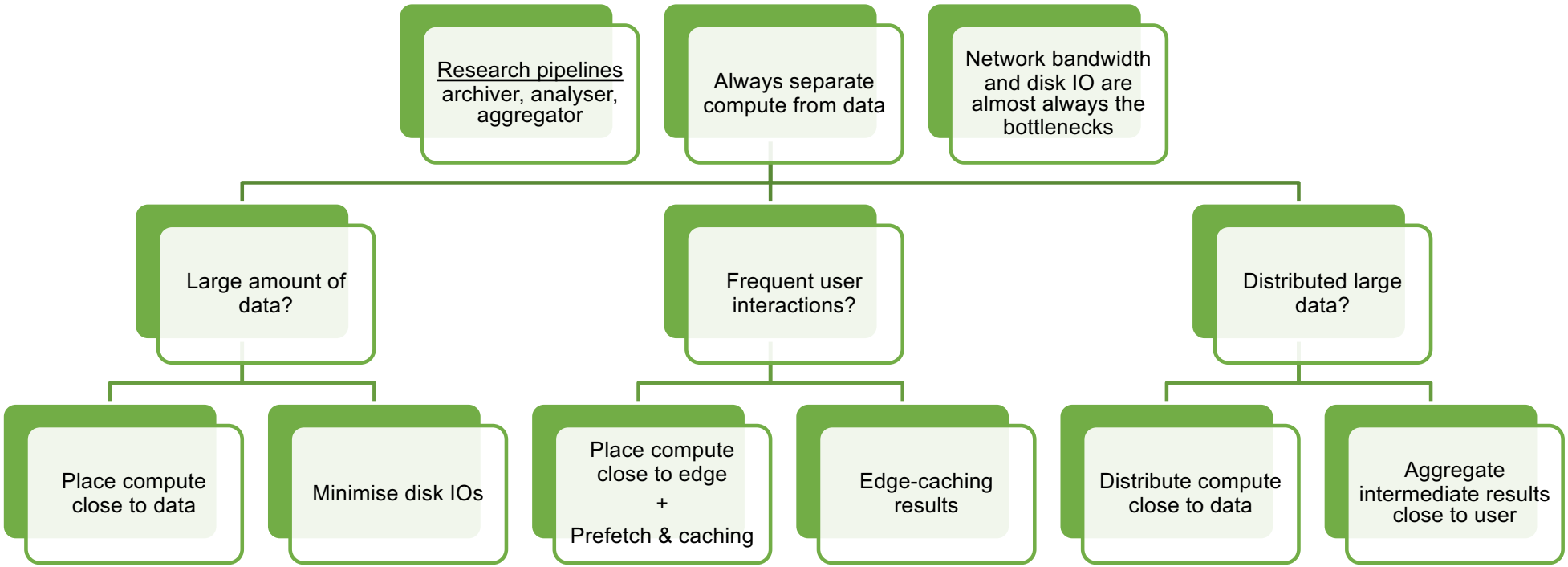
Cloud requirement

- 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

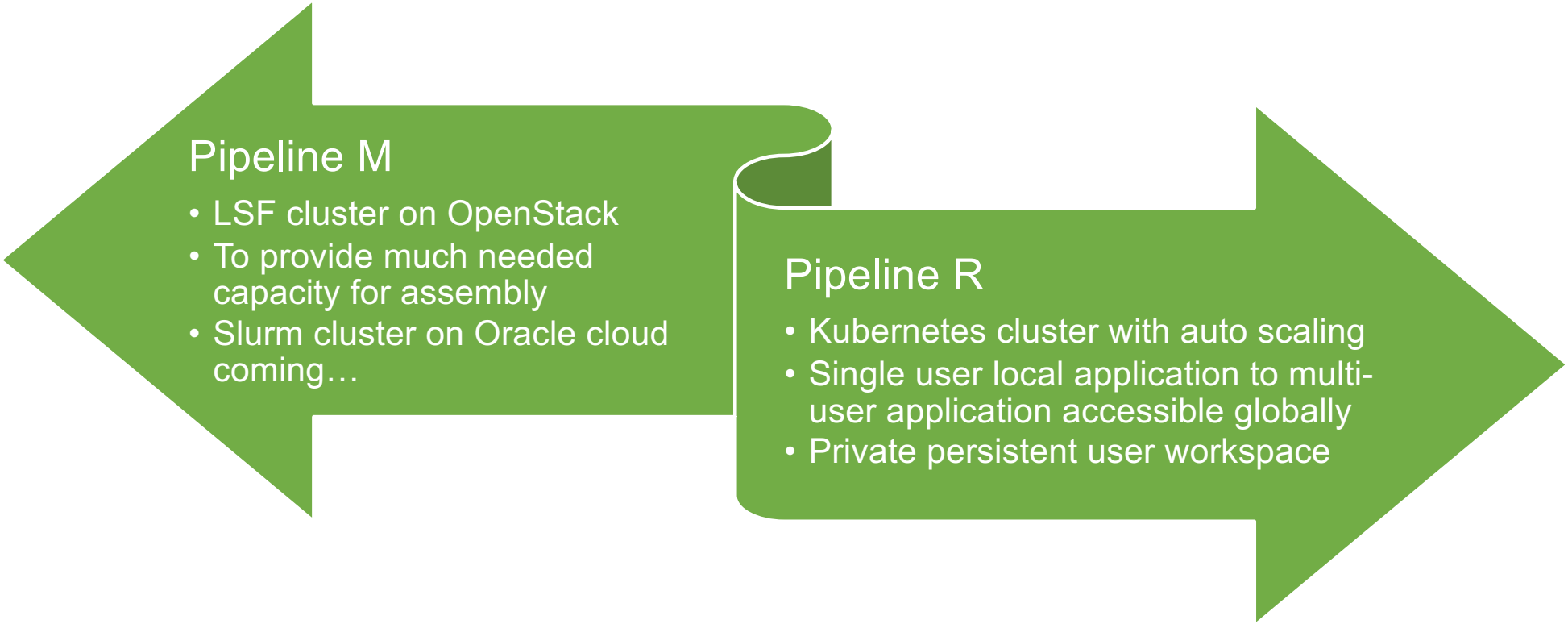
Advises

- 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



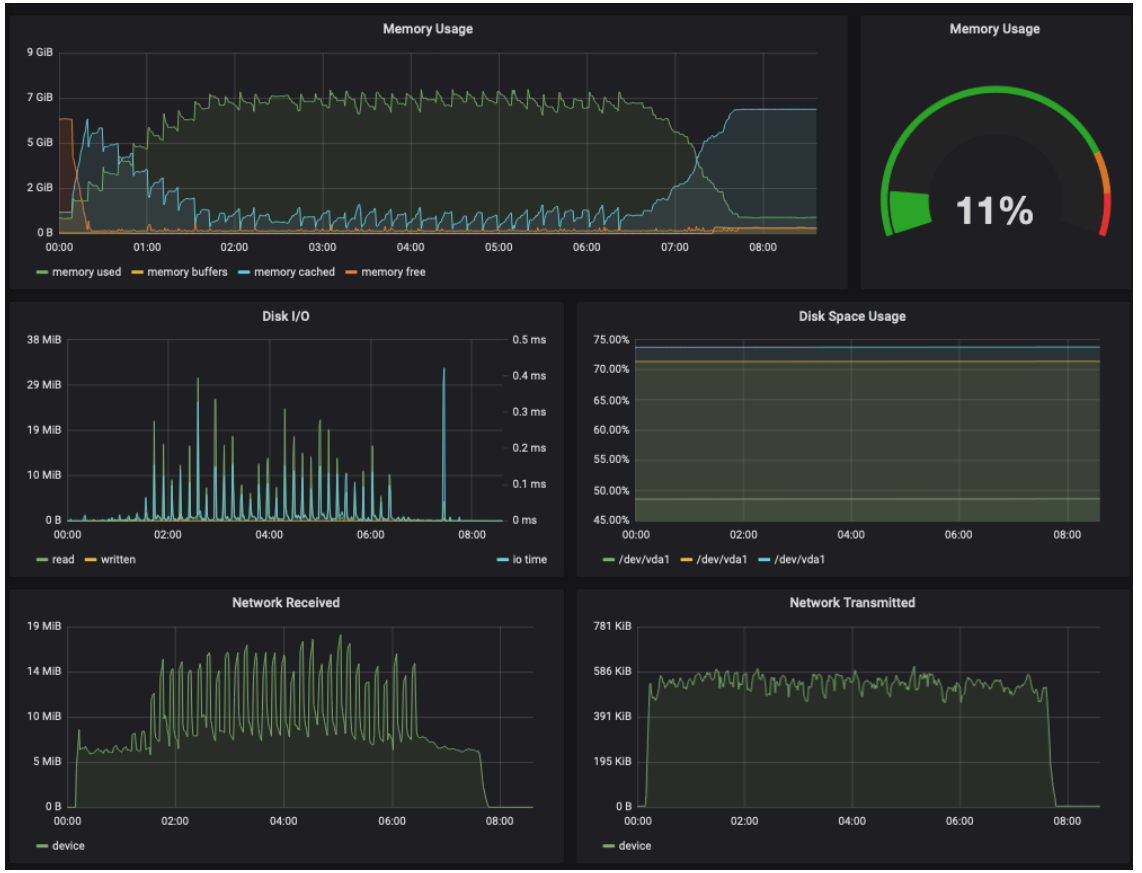
Lift-n-shift vs. cloud-native



Monitoring

Never flying blind

- Monitoring on pipelines is generally lacking
- K8S can be monitored with Prometheus + Grafana
- Kubernetes Dashboard is highly recommended for private K8S
- Monitoring for K8S on public clouds is poor in general



Summary

- 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