UNIVERSITY OF CAMBRIDGE

Systems Research Group (SRG)
Department of Computer Science
University of Cambridge

Distributed Global Scheduling in Datacenters

Smita Vijayakumar
First Year PhD Student
sv440@cst.cam.ac.uk

Evangelia Kalyvianaki
PhD Supervisor
ek264@cst.cam.ac.uk

Anil Madhavapeddy
PhD Supervisor
avsm2@cst.cam.ac.uk

Underutilised Datacenter resources

Azure[1]

❖ 60% VMs have <= **20%** CPU usage!

Alibaba[2] -

❖ Average server CPU **50%**

❖ Memory <= **60%**

Underutilisation is Expensive![3]

Datacenter resources can be better utilised!

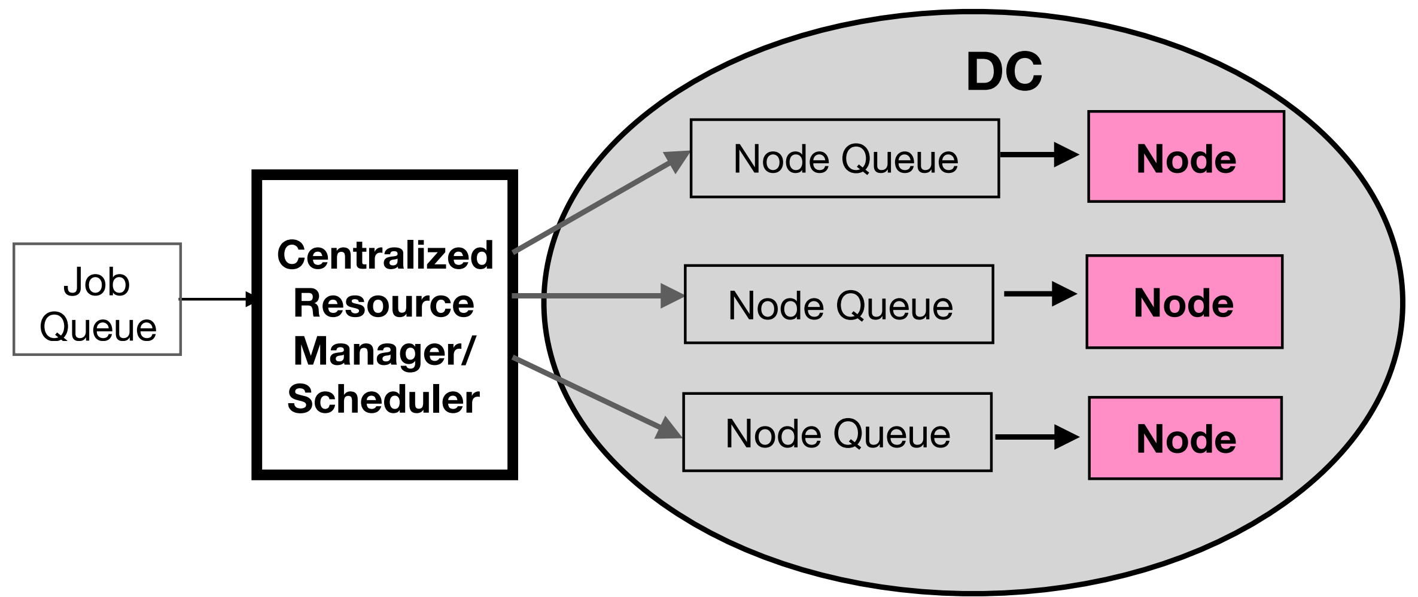
[1] *[Resource Central, SOSP,'17]*

[2] *[https://github.com/alibaba/clusterdata]*

[3] *[Scalable system scheduling for HPC and big data, JPDC,17]*

Schedulers In Datacenter

Centralized



Regular updates by nodes

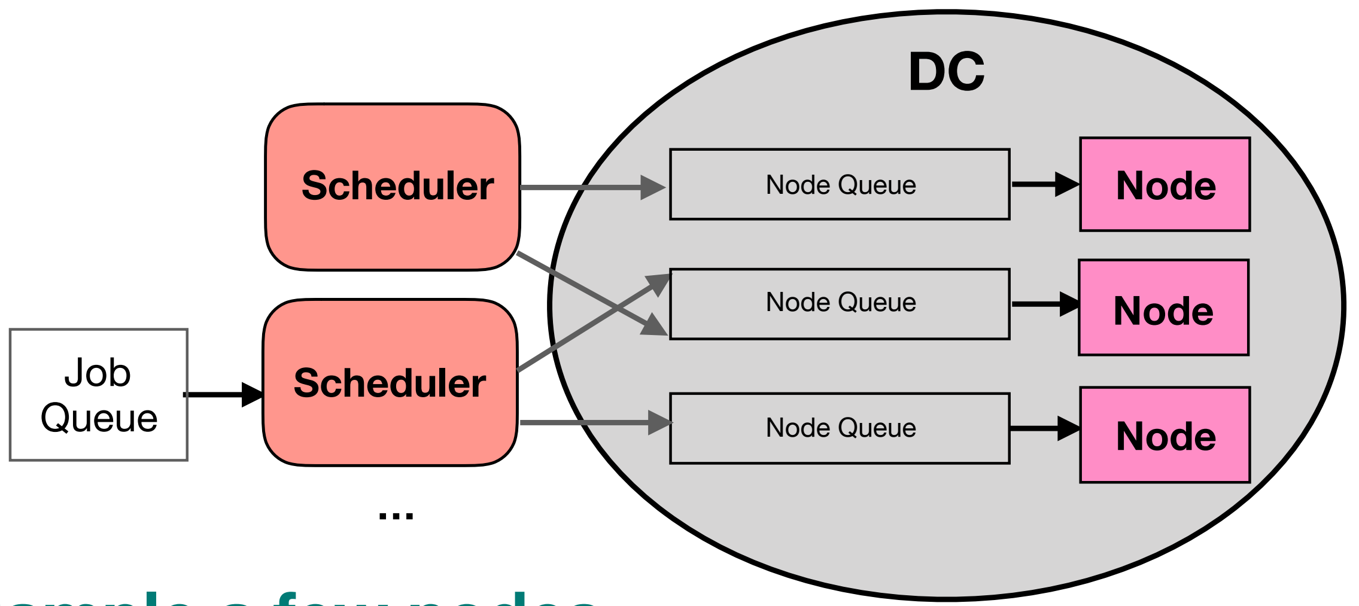
Examples - Mesos *[NSDI'11]*, **Yarn, Apollo** *[OSDI'14]*

☑ Global resource view

☒ Scheduler can be a bottleneck

☒ Delayed and high volumes of node updates

Decentralised



Sample a few nodes

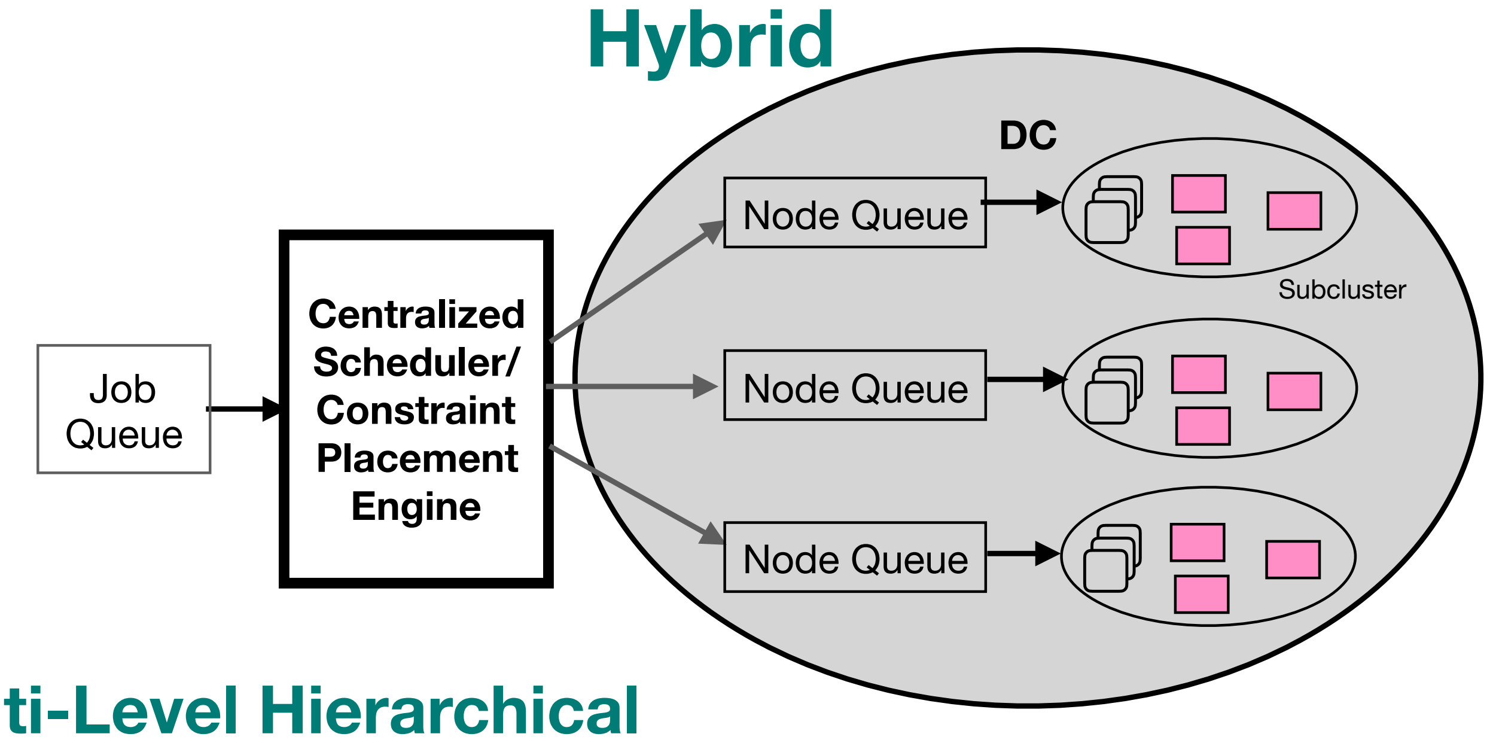
Example - Sparrow *[NSDI'14]*

☑ Fast and simple

☒ Unsuited for Long Running Applications

☒ Not globally optimal

Hybrid



Multi-Level Hierarchical

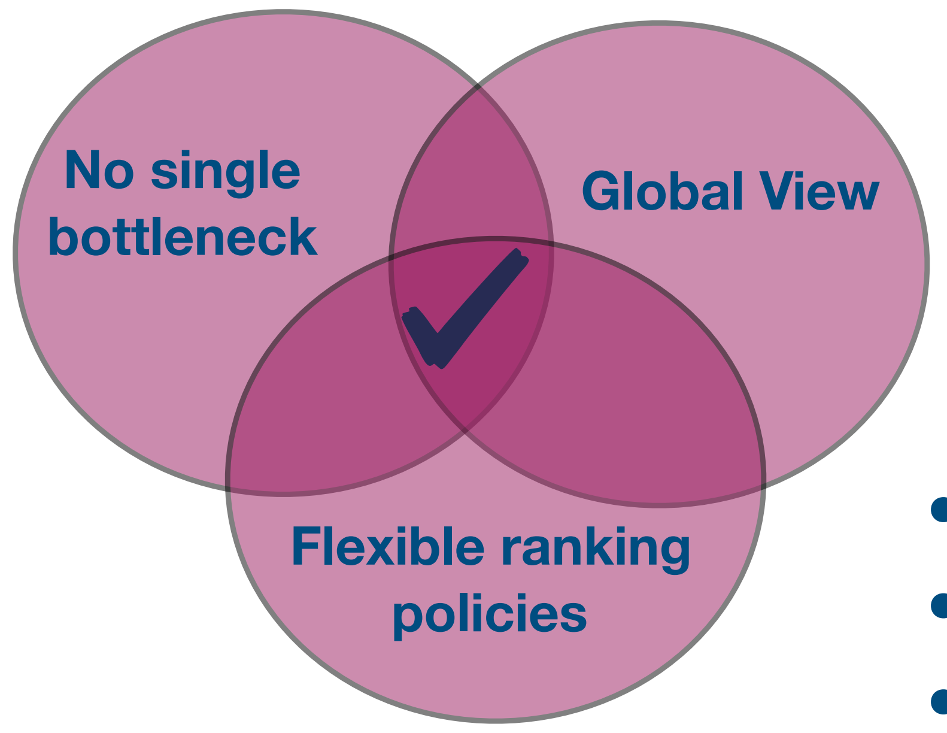
Example - Hydra *[NSDI'19]*, **Medea** *[EuroSys'18]*, **Borg** *[EuroSys'15]*

☑ Policy-driven job/task placement

☑ Less node information traffic

☒ Central components

Global Scheduling at Node Level



Challenges

- Unsuited for short jobs
- Large node status traffic
- Non-trivial convergence time

Up-to-Date Global View At Each Node

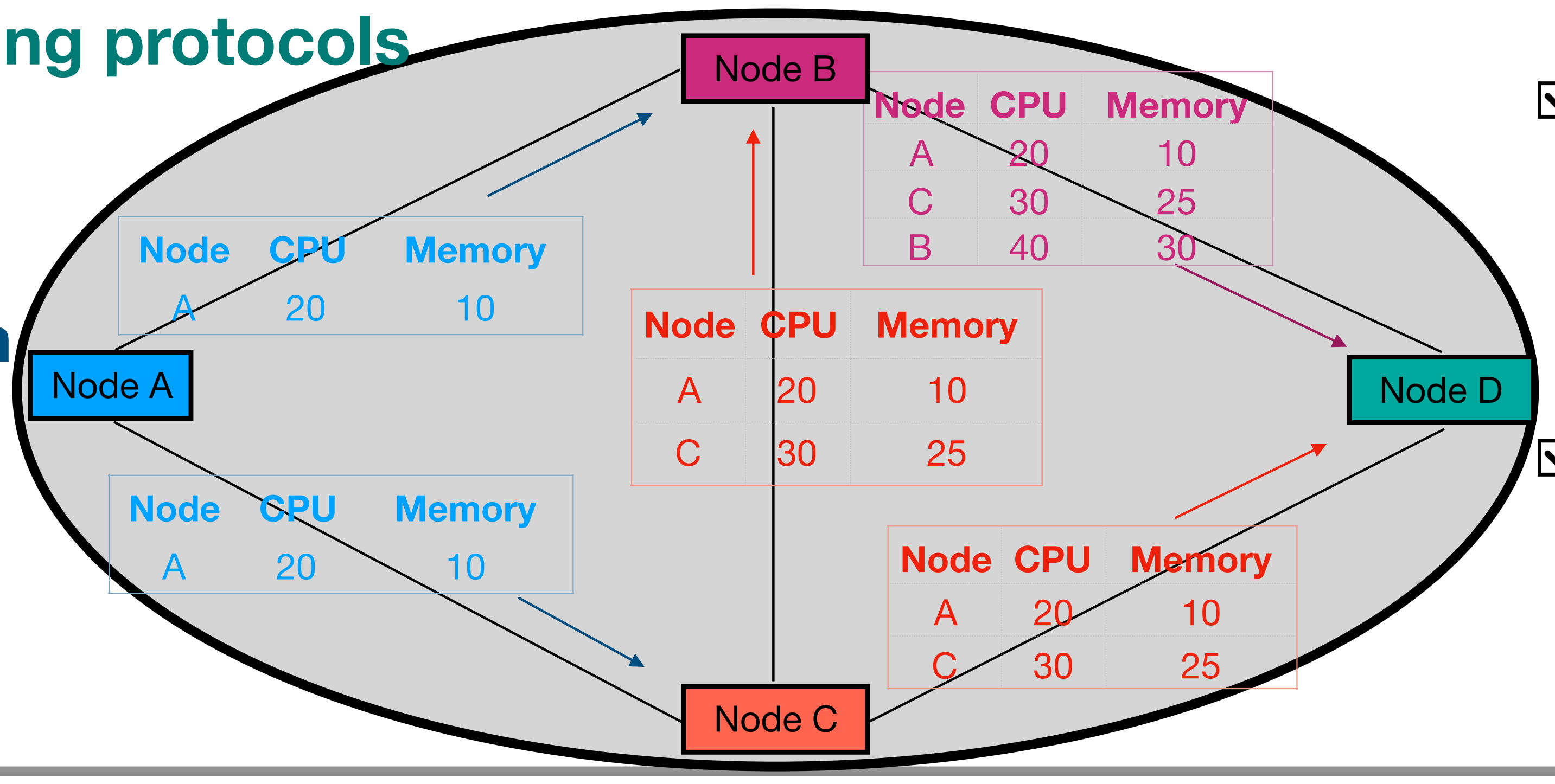
Proposed solution inspired by routing protocols

☑ **BGP, OSPF, ...**

☑ **Resource information propagation**

☑ **Global view convergence**

☑ **Identical ranking policy**



☑ **Resource Information**

❖ Current resource utilisation

❖ Predicted future utilisation

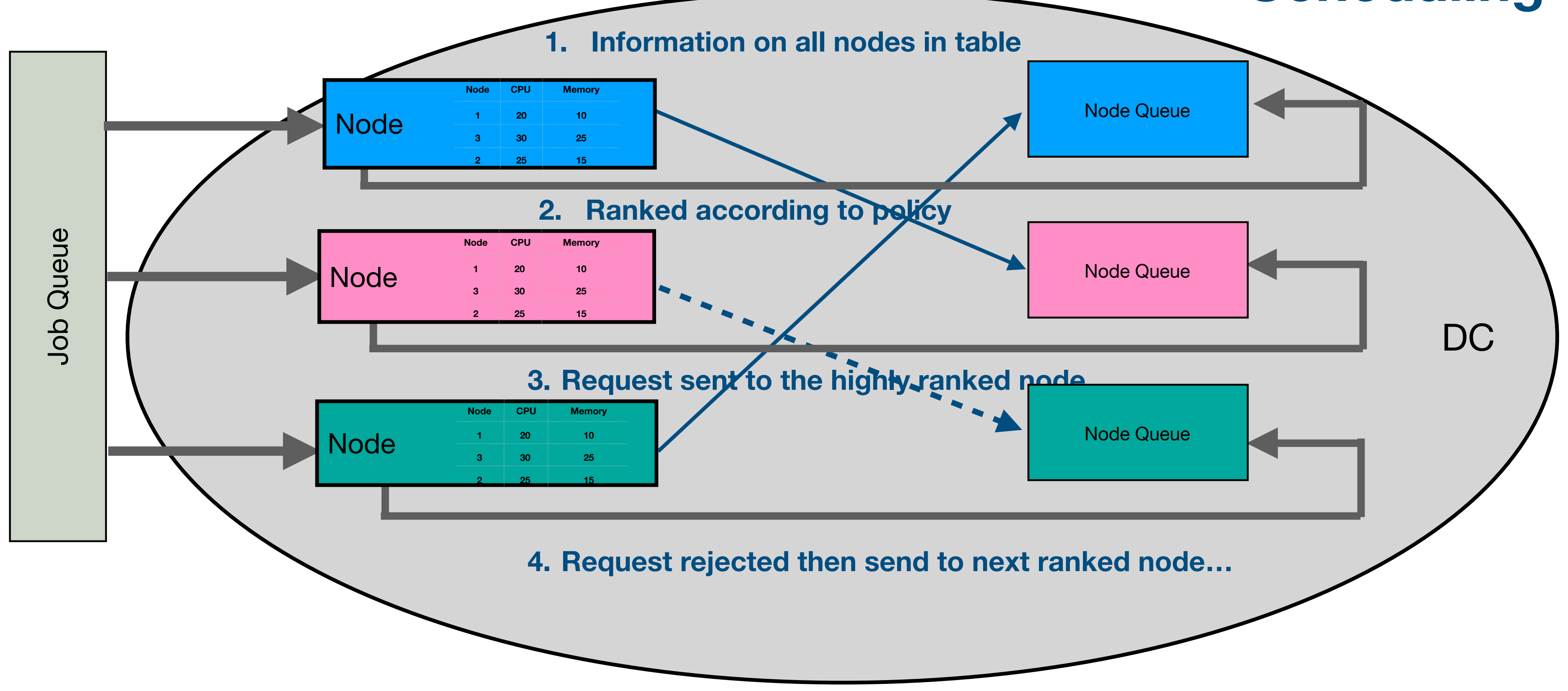
☑ **Ranking**

❖ Better load balancing

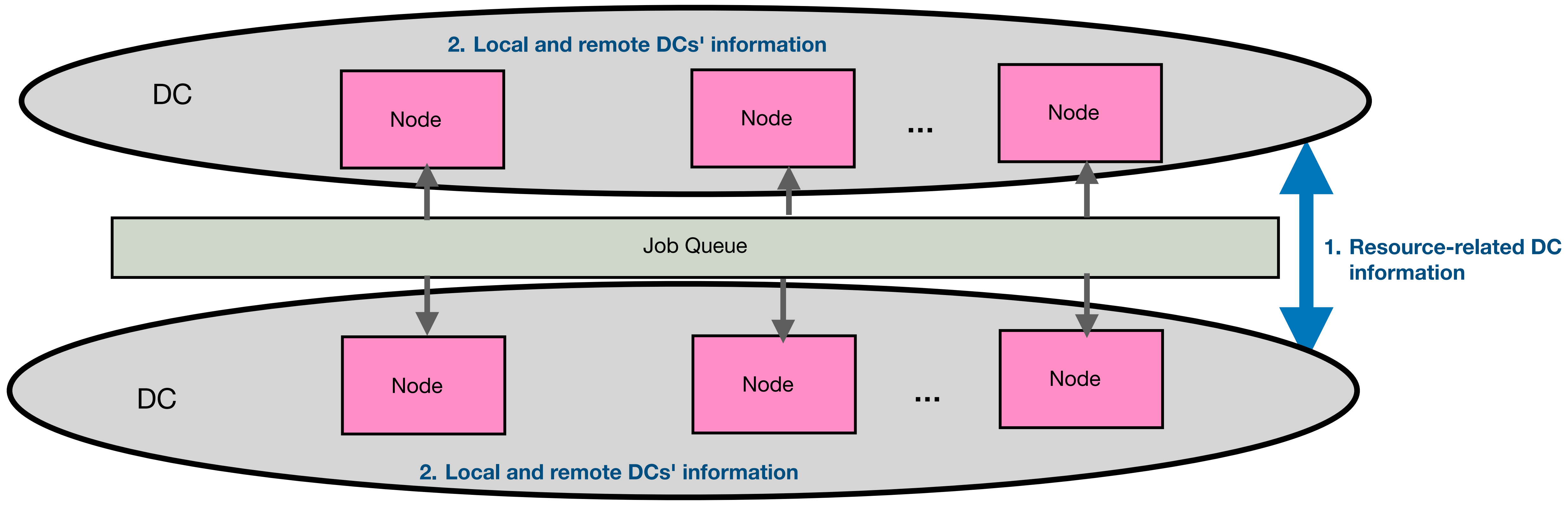
❖ Higher utilisation

❖ Best fit, worst fit, ...

Scheduling Using Up-to-Date Current Global View


Job Queue
DC
Node
Node
Node
Node CPU Memory
1 20 10
3 30 25
2 25 15
Node Queue
Node Queue
Node Queue
1. Information on all nodes in table
2. Ranked according to policy
3. Request sent to the highly ranked node
4. Request rejected then send to next ranked node...

Intra-DC Load Balanced Scheduling


DC
DC
Node
Node
...
Node
Job Queue
Node
Node
...
Node
2. Local and remote DCs' information
2. Local and remote DCs' information
1. Resource-related DC information

Inter-DC Load Balanced Scheduling