Distributed Global Scheduling in Datacenters

UNIVERSITY OF CAMBRIDGE

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

Smita Vijayakumar
First Year PhD Student

Evangelia Kalyvianaki
PhD Supervisor

Anil Madhavapeddy
PhD Supervisor

firstname.lastname@cl.cam.ac.uk

Underutilised Datacenter resources

Azure*
- 60% VMs have <= **20%** CPU usage!

Alibaba** -
- Average server CPU **50%**
- Memory <= **60%**

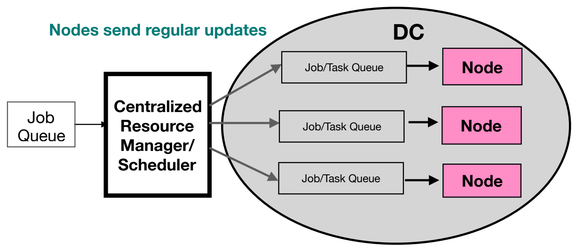
100MW DC*
1% compute cycles = Small City Energy-Saving

Datacenter resources can be better utilised!

**Resource Central, SOSP'17*
***https://github.com/alibaba/clusterdata*
****Scalable system scheduling for HPC and big data, JPDC'17*

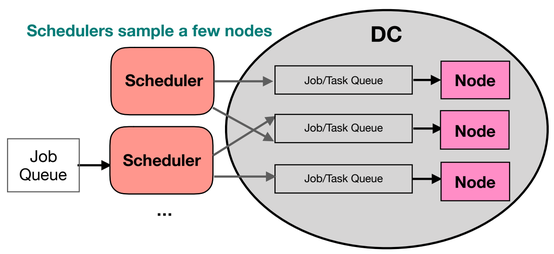
Schedulers In Datacenter

Centralized



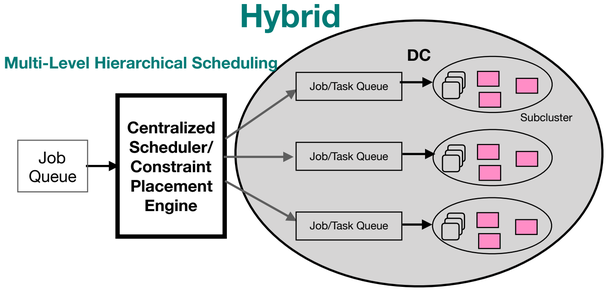
Examples - Mesos *[NSDI'11]*, **Yarn, Apollo** *[OSDI'14]*
- ☑ **Global resource view**
- ✖ **Scheduler bottleneck**
- ✖ **Large node status traffic**

Decentralised



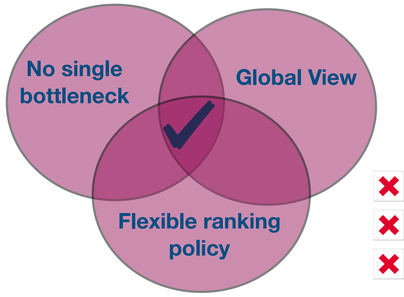
Example - Sparrow *[NSDI'14]*
- ☑ **Fast and simple**
- ✖ **Unsuited for Long Running Applications**
- ✖ **Not globally optimal**

Hybrid



Example - Hydra *[NSDI'19]*, **Medea** *[EuroSys'18]*, **Borg** [EuroSys'15]
- ☑ **Better job/task placement**
- ☑ **Lesser node information traffic**
- ✖ **Central components**

Node Level Global Scheduling Intelligence

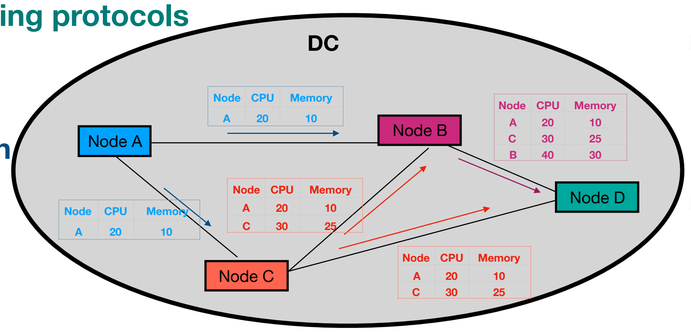


- ✖ Unsuited for short jobs
- ✖ Large node status traffic
- ✖ Non-trivial convergence time

Timely Current Global View At Each Node

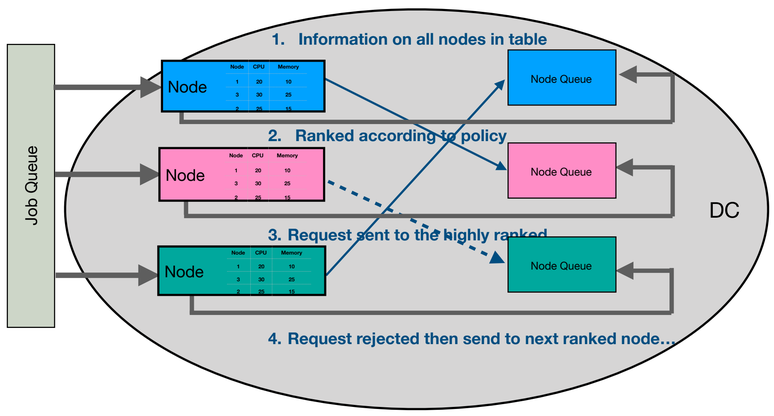
Proposed solution inspired by routing protocols

- ☑ **BGP, OSPF, ...**
- ☑ **Resource information propagation**
- ☑ **Global view convergence**
- ☑ **Identical ranking policy**



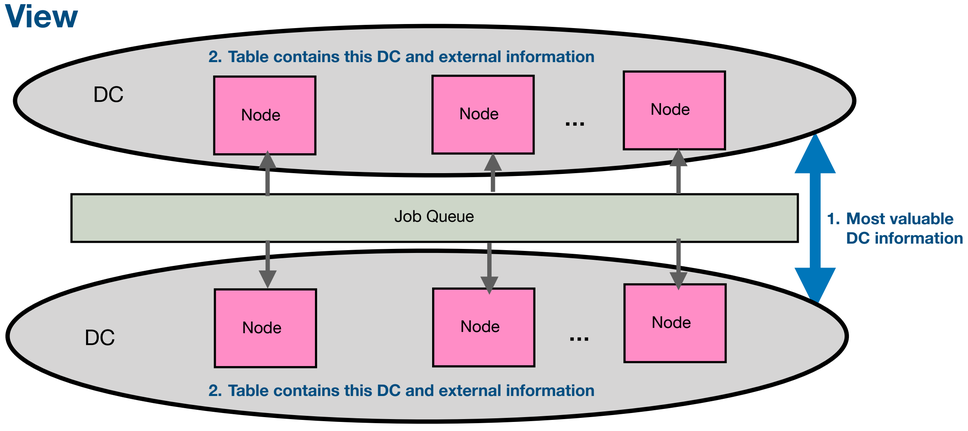
- ☑ **Resource Information**
  - ❖ **Current resource utilisation**
  - ❖ **Predicted future utilisation**
- ☑ **Ranking**
  - ❖ **Better load balancing**
  - ❖ **Higher utilisation**
  - ❖ **Best fit, worst fit, ...**

Scheduling Using Timely Current Global View



Intra-DC Load Balanced Scheduling

Various Design Approaches
- **Multiple job requests sent**
- **Resource pattern learning**
- **Suggestions?**



Inter-DC Load Balanced Scheduling