Omega

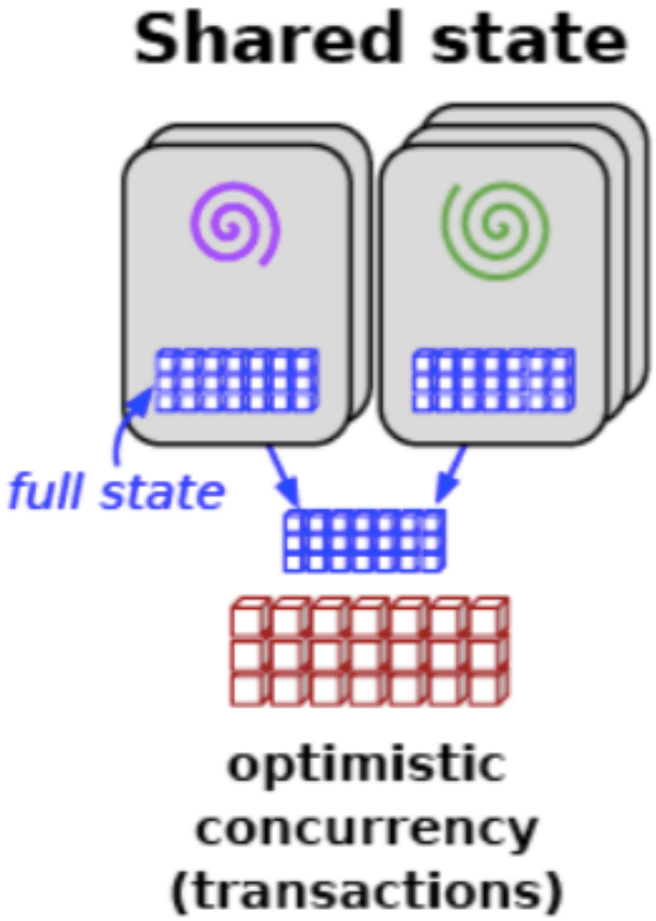
shared state scheduler

Workload Heterogeneity

- Service jobs
  - long-time running
  - e.g. end-user operations or internal infrastructure
  - stringent availability and performance targets
  - require placement to avoid failures
- Batch jobs
  - computation then finish
  - e.g. batch log analysis
  - short, fast turnaround is important
  - require lightweight, low quality approach

Existing Problems

- Monolithic system
  - complex calculation of priority
  - multiple code paths for different types of jobs, difficult to support in a single code base
- Two level-system
  - assume job sizes are small compared to the size of the cluster
  - no global view of resources, no preemption
  - hoarding for gang scheduling, potentially deadlock



1. master maintain “cell state”, a copy of the resource allocation

2. each scheduler maintain a local copy of “cell state”

3. each scheduler could claim any available cluster resources

4. master would only allow one claim to be succeed in case of conflict

5. scheduler may resync local copy of cell state and rerun scheduling algorithm

Data sources

- A: medium-sized, fairly busy one
- B: larger clusters
- C: scheduler workload trace in [1][2]

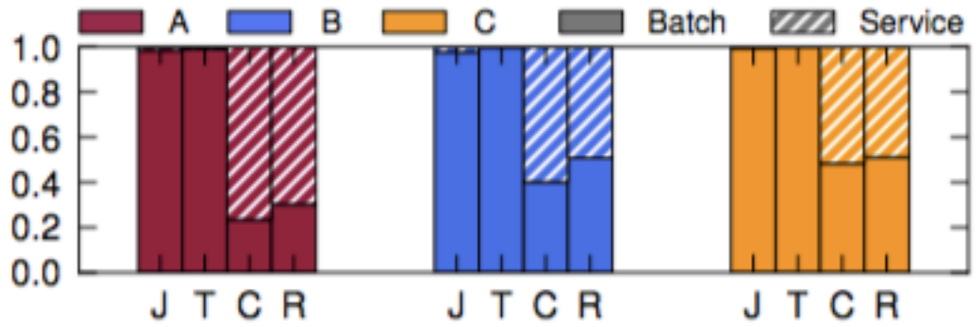


Figure 2: Batch and service workloads for the clusters A, B, and C: normalized numbers of jobs (**J**) and tasks (**T**), and aggregate requests for CPU-core-seconds (**C**) and RAM GB-seconds (**R**). The striped portion is the service jobs; the rest is batch jobs.

[1] REISS, C., TUMANOV, A., GANGER, G. R., KATZ, R. H., AND KOZUCH, M. A. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of SoCC* (2012).
[2] WILKES, J. More Google cluster data. Google research blog, Nov. 2011. Posted at http://goo.gl/9B7PA.

Simulation

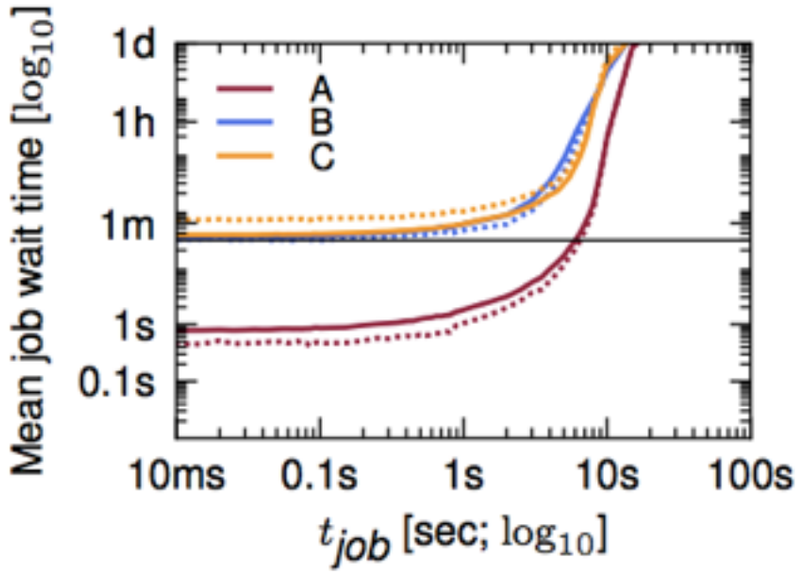
- Lightweight simulator: obtain matrices derived from real workload

- High-fidelity simulator: driven by the actual workload traces

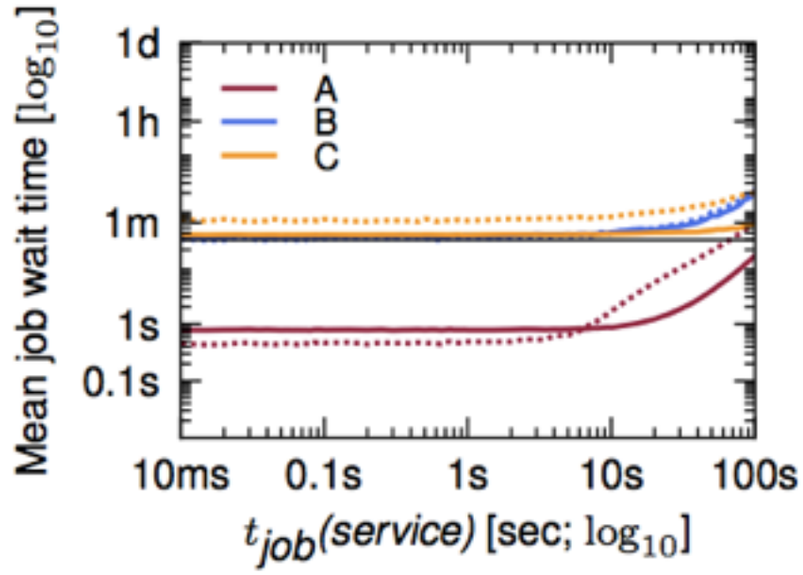
	Lightweight (§4)	*High-fidelity (§5)*
Machines	homogeneous	actual data
Resource req. size	sampled	actual data
Initial cell state	sampled	actual data
tasks per job	sampled	actual data
λ_{jobs}	sampled	actual data
Task duration	sampled	actual data
Sched. constraints	ignored	obeyed
Sched. algorithm	randomized first fit	Google algorithm
Runtime	fast (24h ≋ 5 min.)	slow (24h ≋ 2h)

Table 2: Comparison of the two simulators; "actual data" refers to use of information found in a detailed workload-execution trace taken from a production cluster.

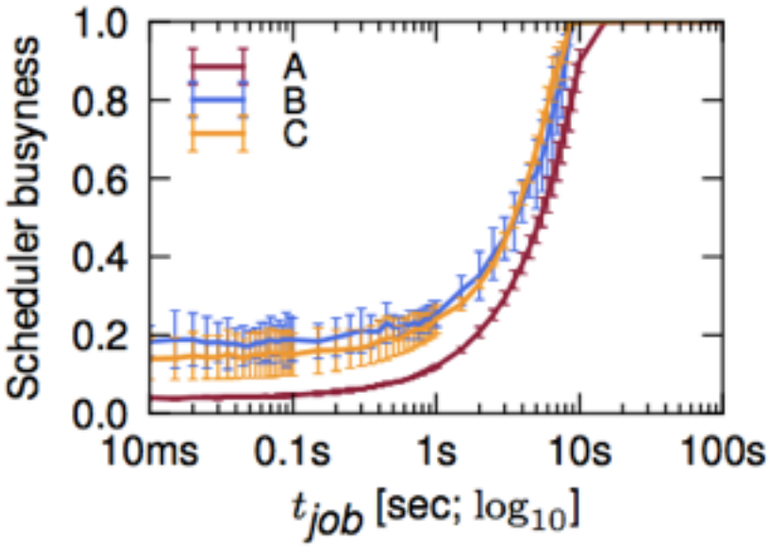
Monolithic scheduler



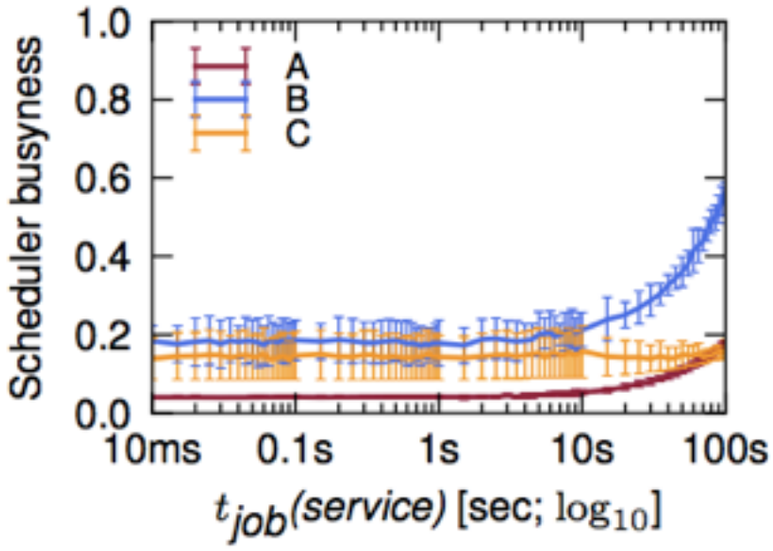
(a) Single-path.



(b) Multi-path.

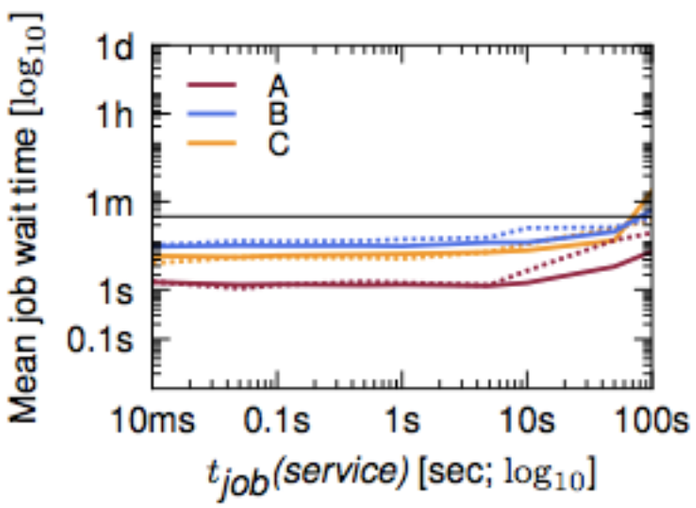


(a) Single-path.

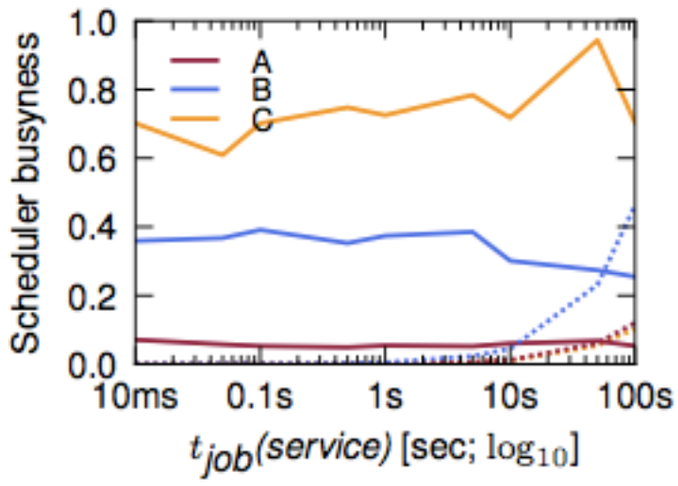


(b) Multi-path.

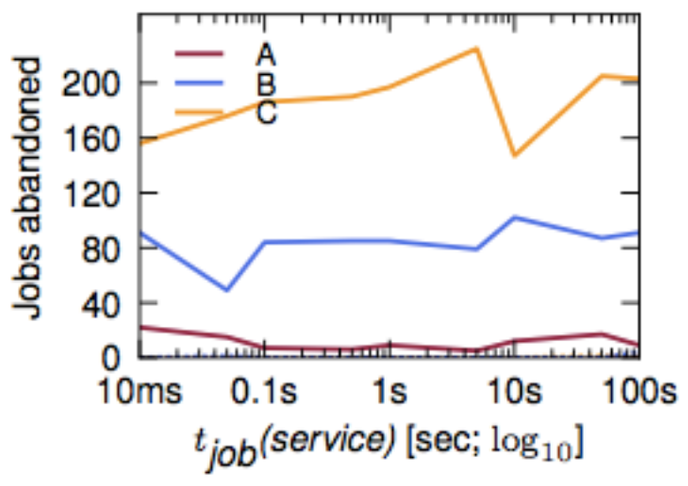
Two-level scheduler



(a) Job wait time.

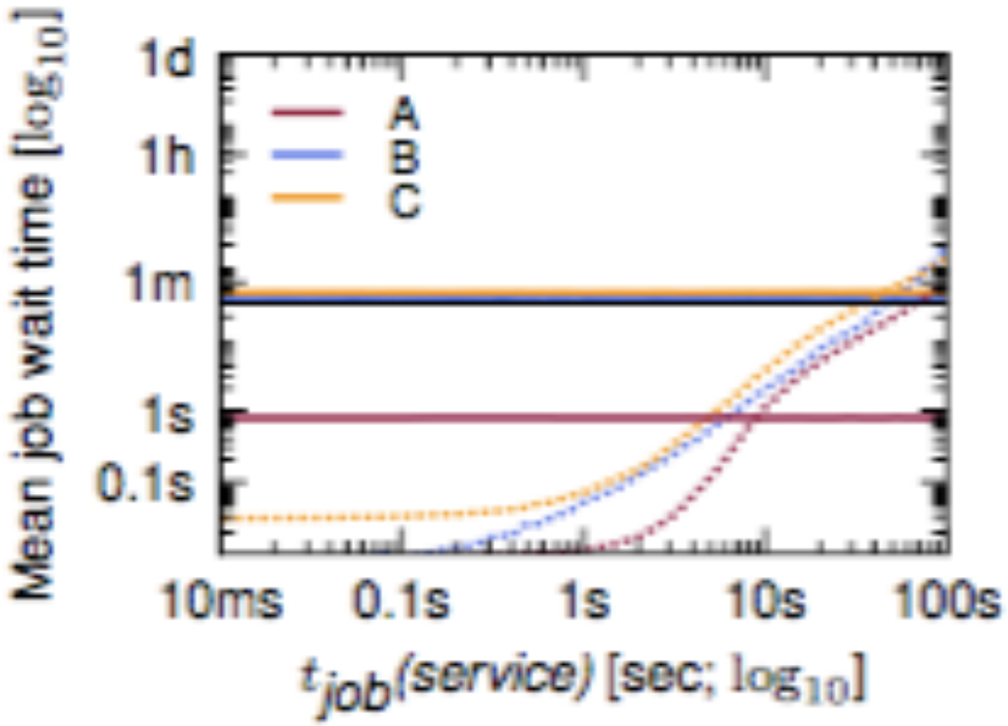


(b) Scheduler busyness.

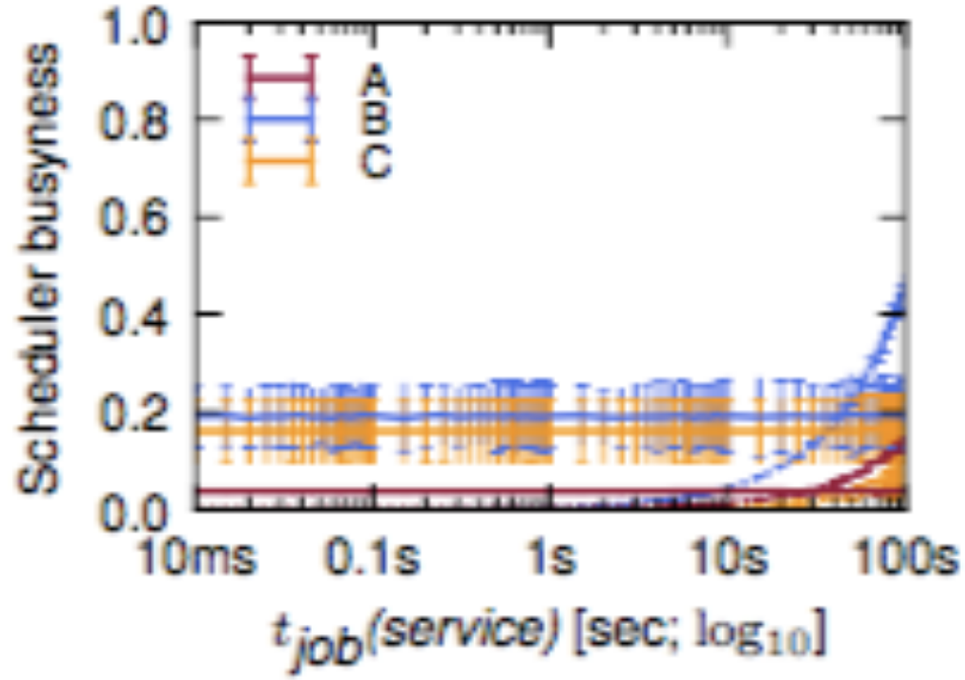


(c) Unscheduled jobs.

Omega

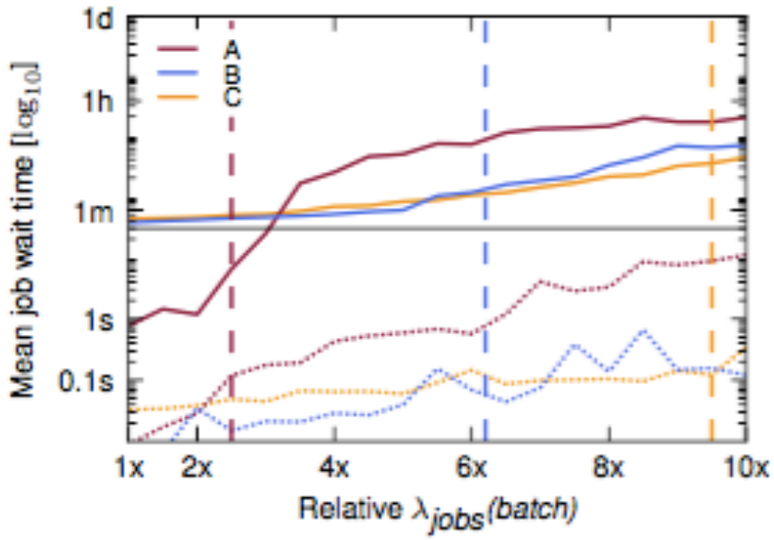


(c) Shared state.

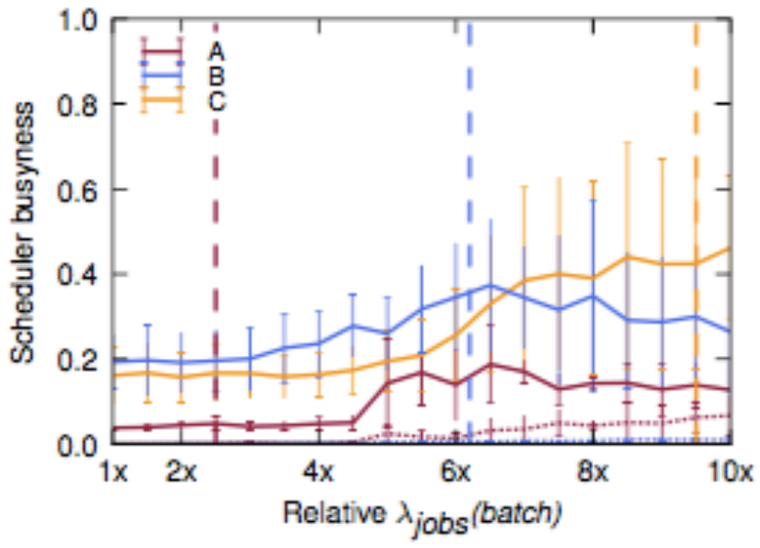


(c) Shared state.

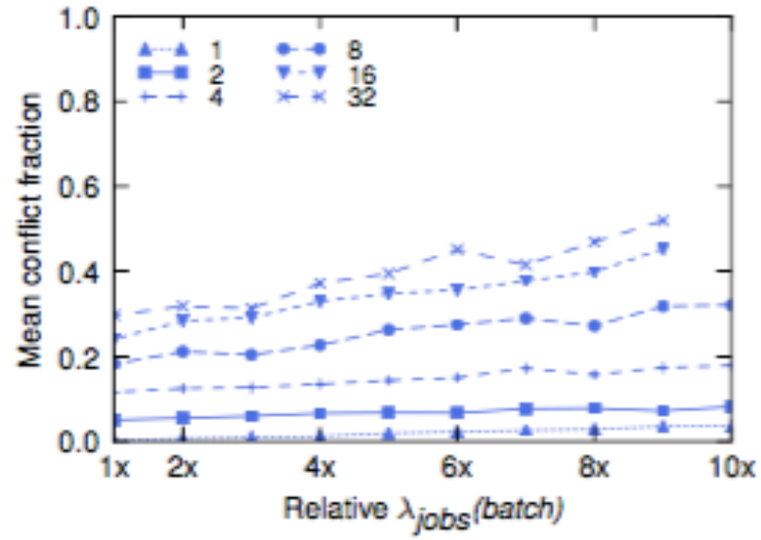
Omega Scalability



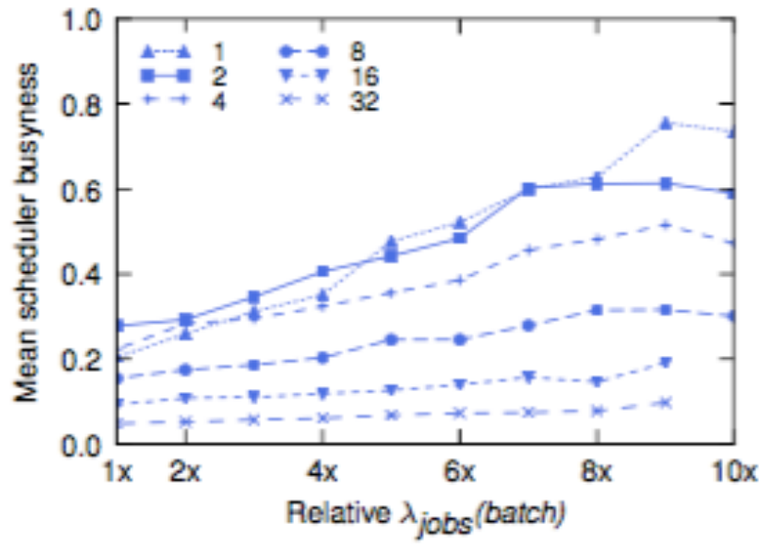
(a) Job wait time.



(b) Scheduler busyness.



(a) Mean conflict fraction.



(b) Mean sched. busyness.

Simulation

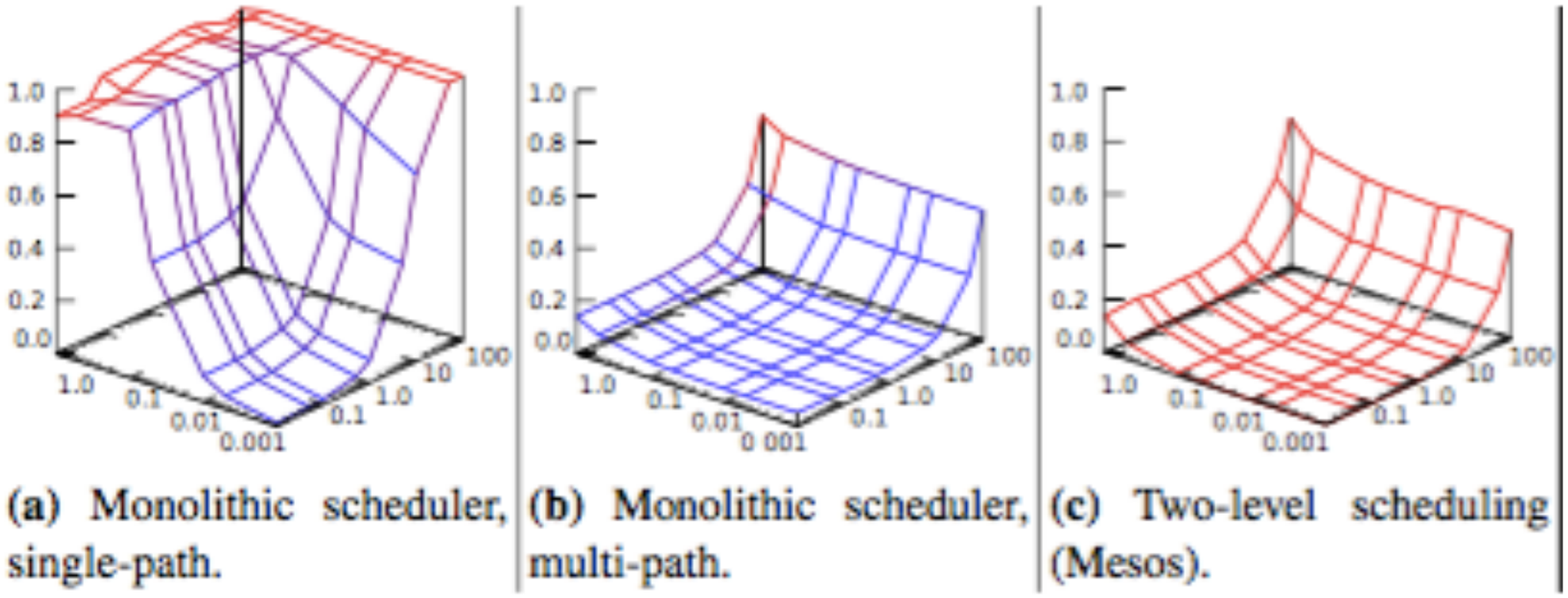
- Lightweight simulator: obtain matrices derived from real workload

- High-fidelity simulator: driven by the actual workload traces

	Lightweight (§4)	*High-fidelity (§5)*
Machines	homogeneous	actual data
Resource req. size	sampled	actual data
Initial cell state	sampled	actual data
tasks per job	sampled	actual data
λ_{jobs}	sampled	actual data
Task duration	sampled	actual data
Sched. constraints	ignored	obeyed
Sched. algorithm	randomized first fit	Google algorithm
Runtime	fast (24h ≋ 5 min.)	slow (24h ≋ 2h)

Table 2: Comparison of the two simulators; "actual data" refers to use of information found in a detailed workload-execution trace taken from a production cluster.

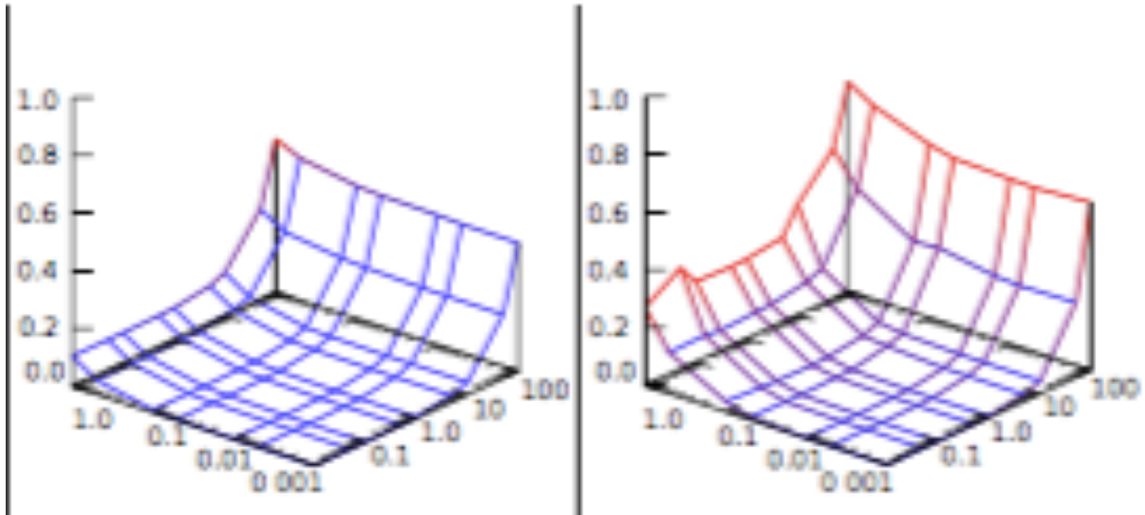
Omega Performance



(a) Monolithic scheduler, single-path.

(b) Monolithic scheduler, multi-path.

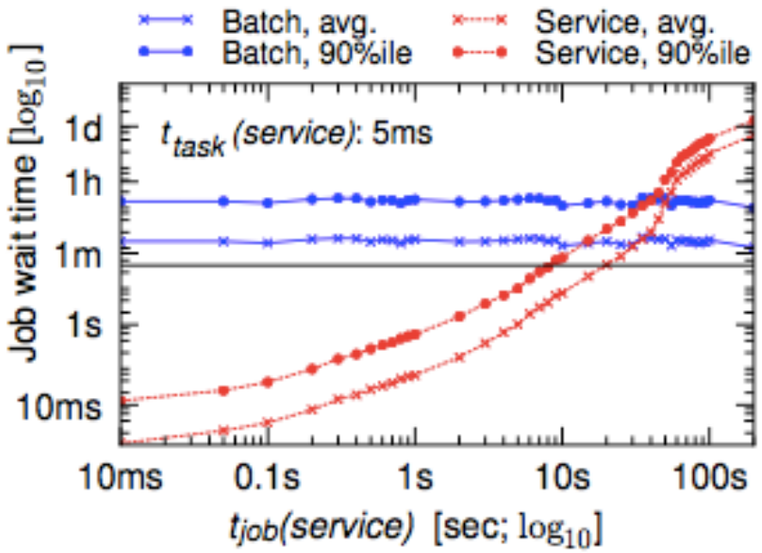
(c) Two-level scheduling (Mesos).



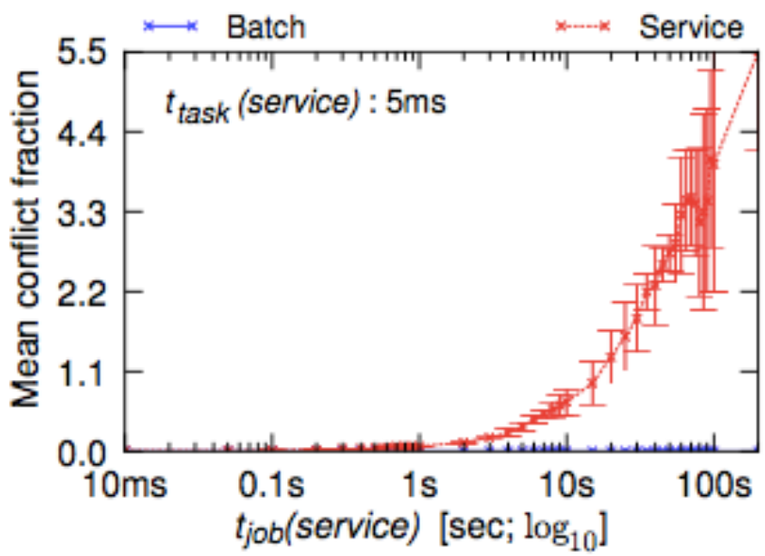
(d) Shared-state (Omega).

(e) Shared-state, coarse, gang scheduling.

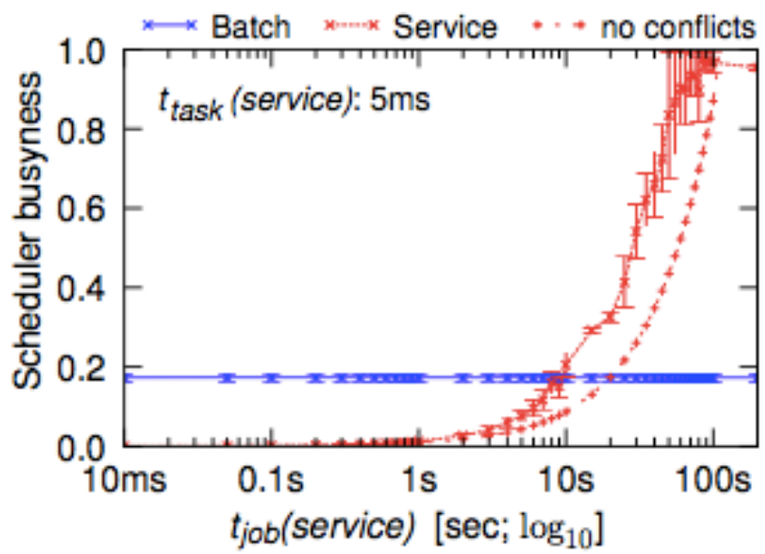
Omega Performance



(a) Job wait time.

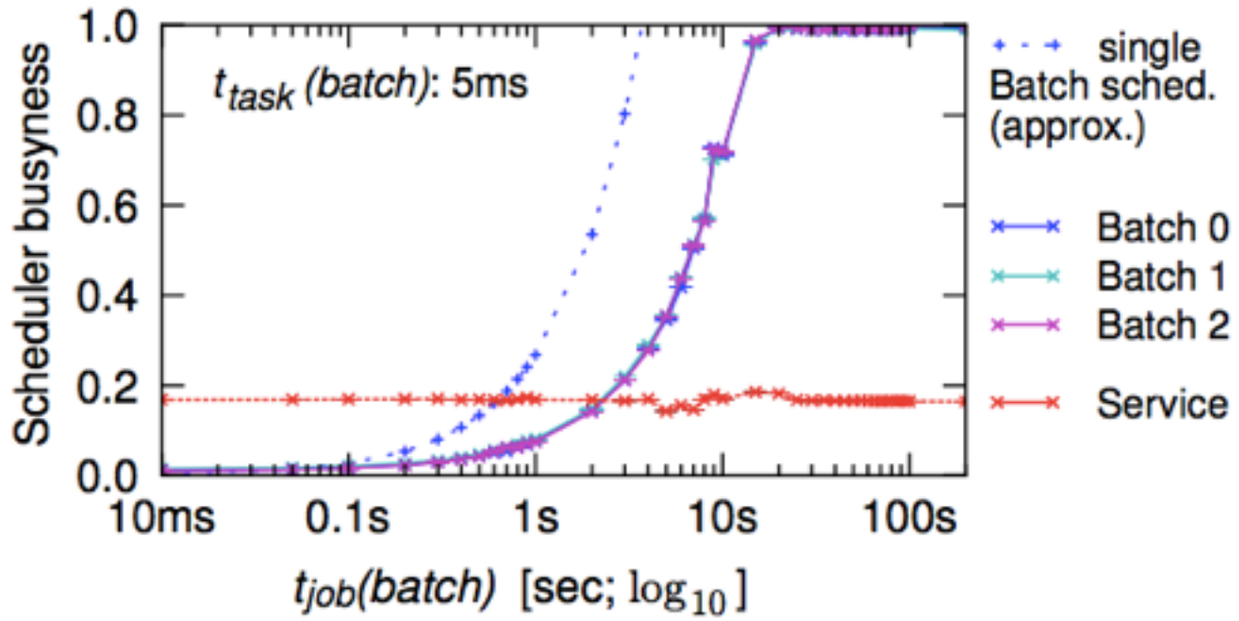


(b) Conflict fraction.

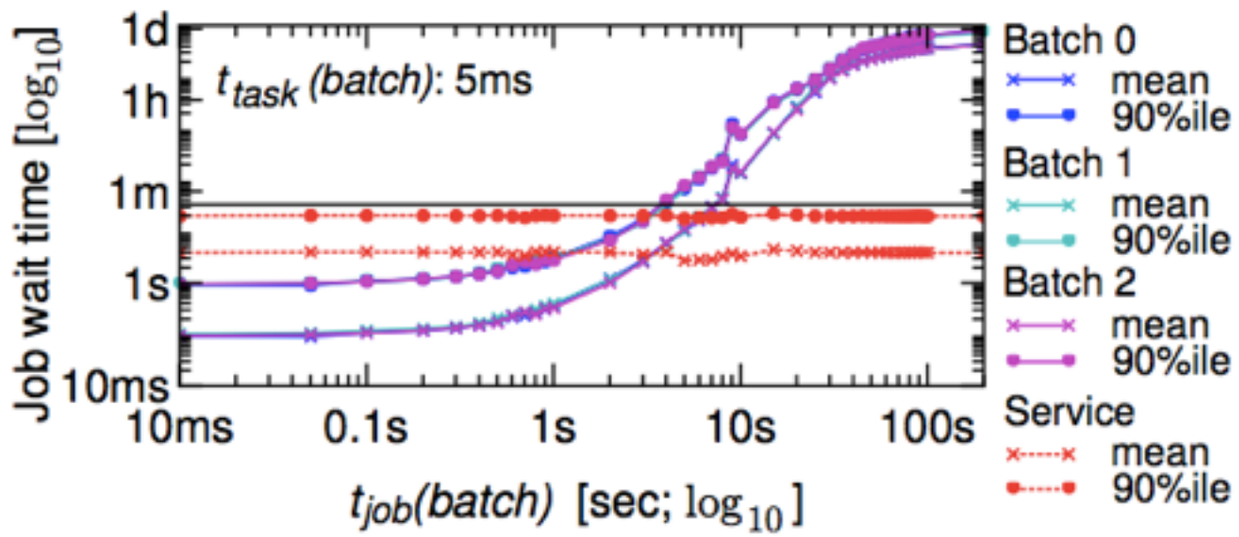


(c) Scheduler busyness.

Omega Performance



(a) Scheduler busyness.



(b) Job wait time.