Building Scalable and Flexible Cluster Managers Using Declarative Programming

Lalith Suresh, Joao Loff[1], Faria Kalim[2], Sangeetha Abdu Jyothi[3], Nina Narodytska, Leonid Ryzhyk, Sahan Gamage, Brian Oki, Pranshu Jain, Michael Gasch

VMware, [1]IST (ULisboa) / INESC-ID, [2]UIUC, [3]UC Irvine and VMware

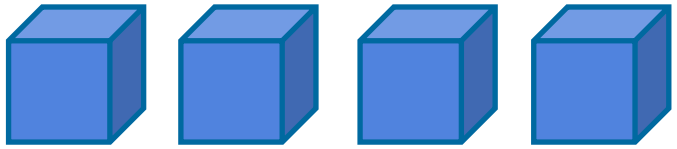
Hard to Build ☹

Cluster Managers

DCM ☺

Code-generate implementations from high-level specifications

Pods

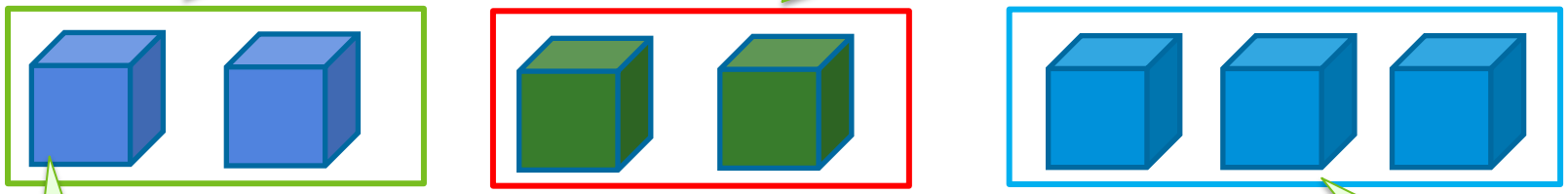


Kubernetes Scheduler

Nodes

Place us on the same rack!

Do **NOT** place us on the same rack!



POD
- 2GB RAM
- 16GB disk
- 1 core

Kubernetes Scheduler

Distribute us evenly!

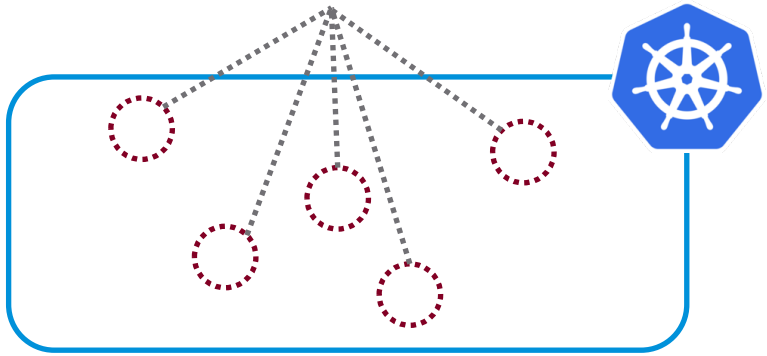
30 hard and soft constraints

NP-Hard
Multi-dimensional bin-packing with constraints

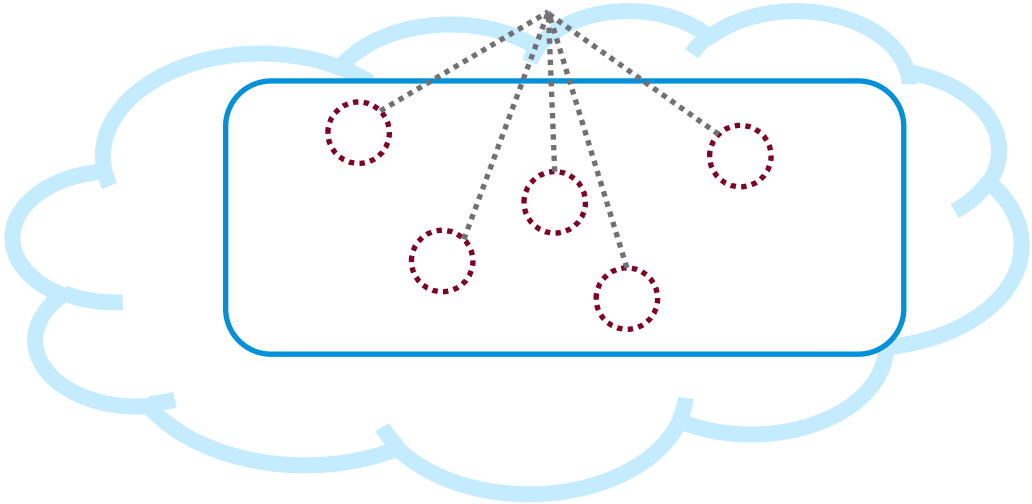
Nodes

Kubernetes Scheduler

Custom Best-effort Heuristics



Custom Best-effort Heuristics



Scalability?

Challenging with complex constraints

Decision quality?

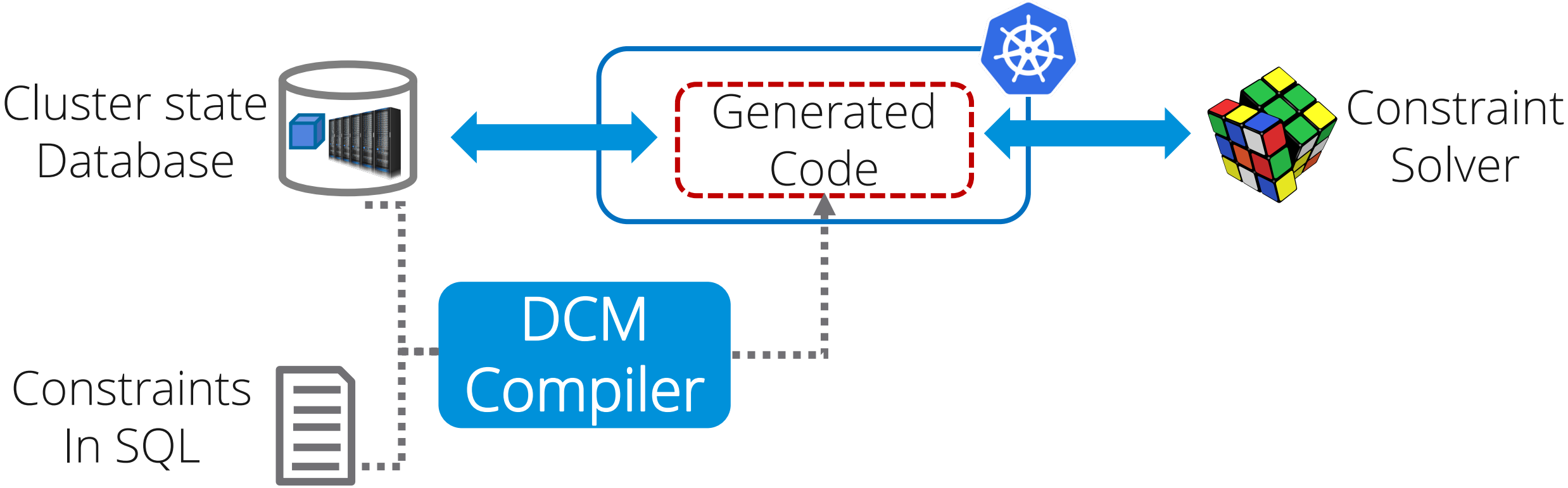
Can miss feasible solutions

Extensibility?

Hard to add new policies and features

Our approach

Declarative Cluster Managers (DCM)



Our approach

Declarative Cluster Managers (DCM)

Use cases

Kubernetes Scheduler	VM Load Balancing Tool	Distributed Transactional Datastore

Scalability

Decision quality

Extensibility

Our approach

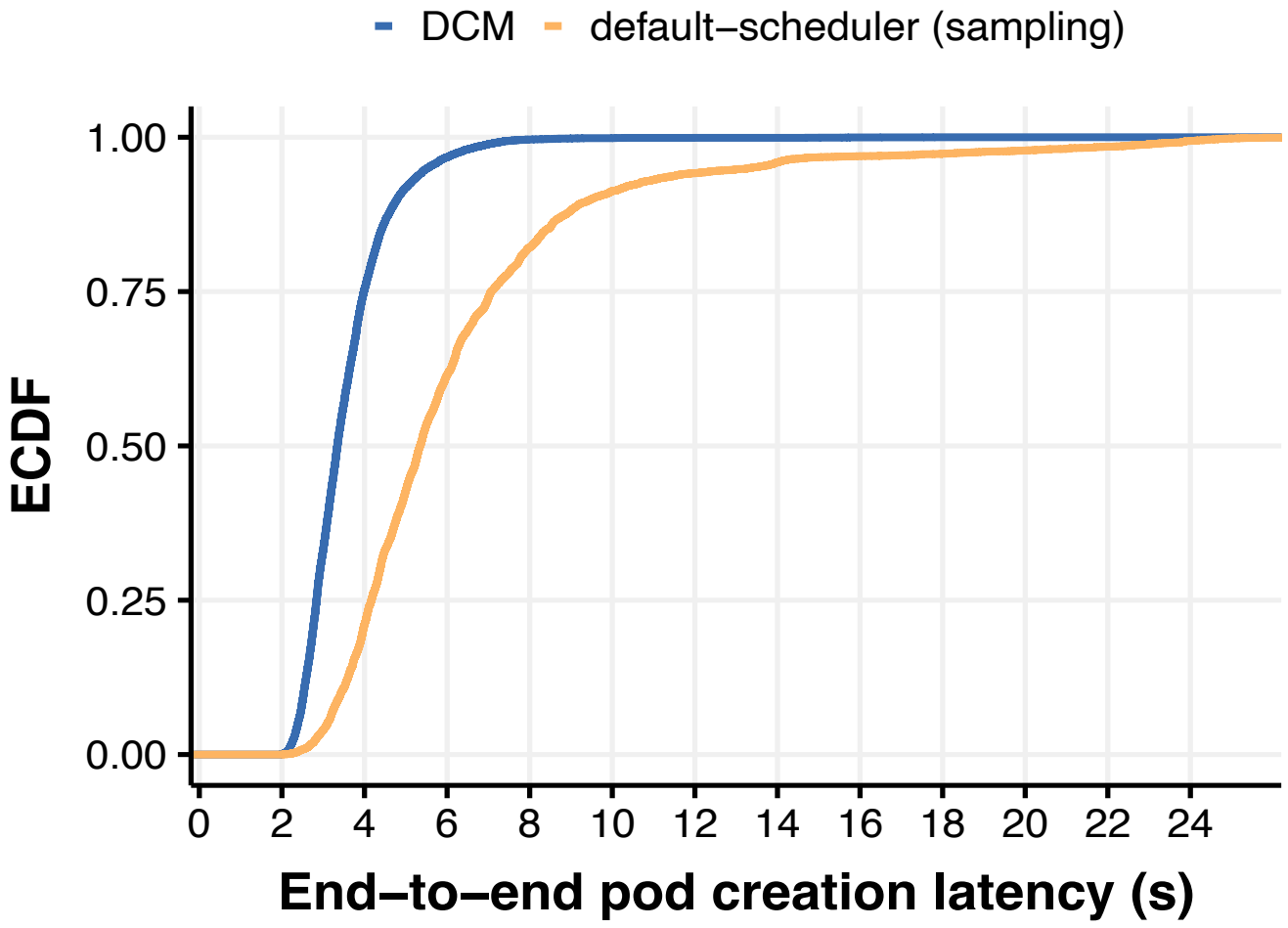
Declarative Cluster Managers (DCM)

Kubernetes Scheduler

Scalability

Up to 2x faster (p95) pod placement than Kubernetes Scheduler (500 node scale)

tig



Our approach

Declarative Cluster Managers (DCM)

Use cases

Kubernetes Scheduler	VM Load Balancing Tool	Distributed Transactional Datastore

Scalability

Up to 2x faster (p95) pod placement than Kubernetes Scheduler (500 node scale)

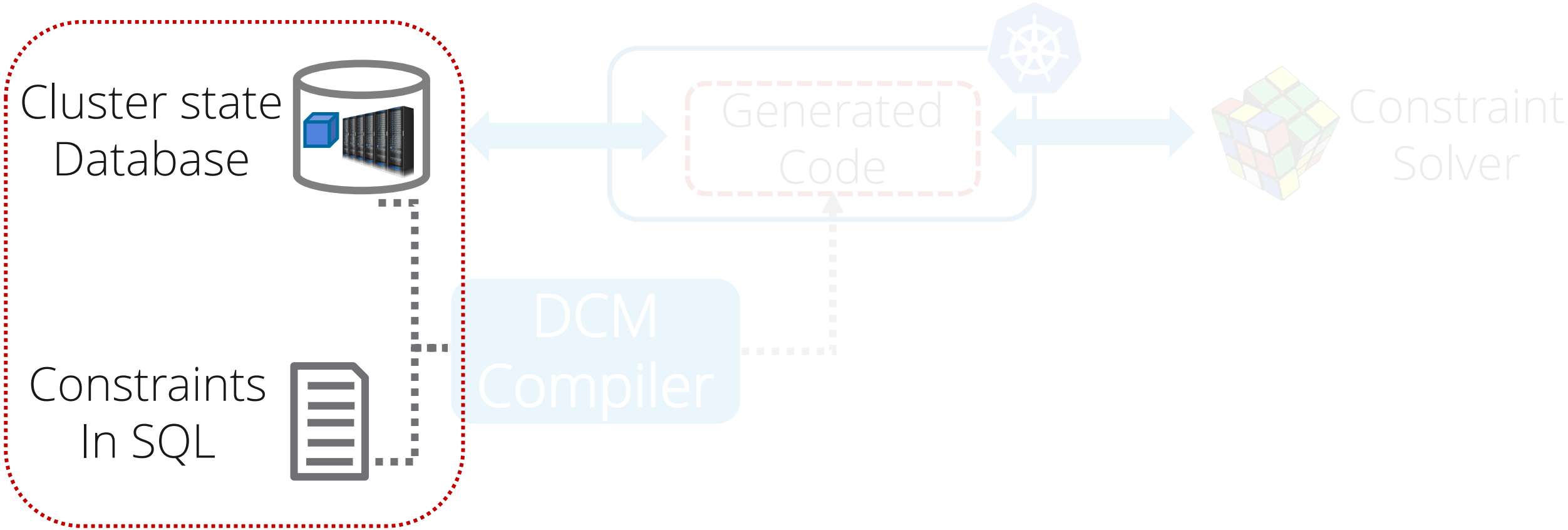
Decision quality

4x better load balancing
2x faster pre-emption
tightly constrained scenarios

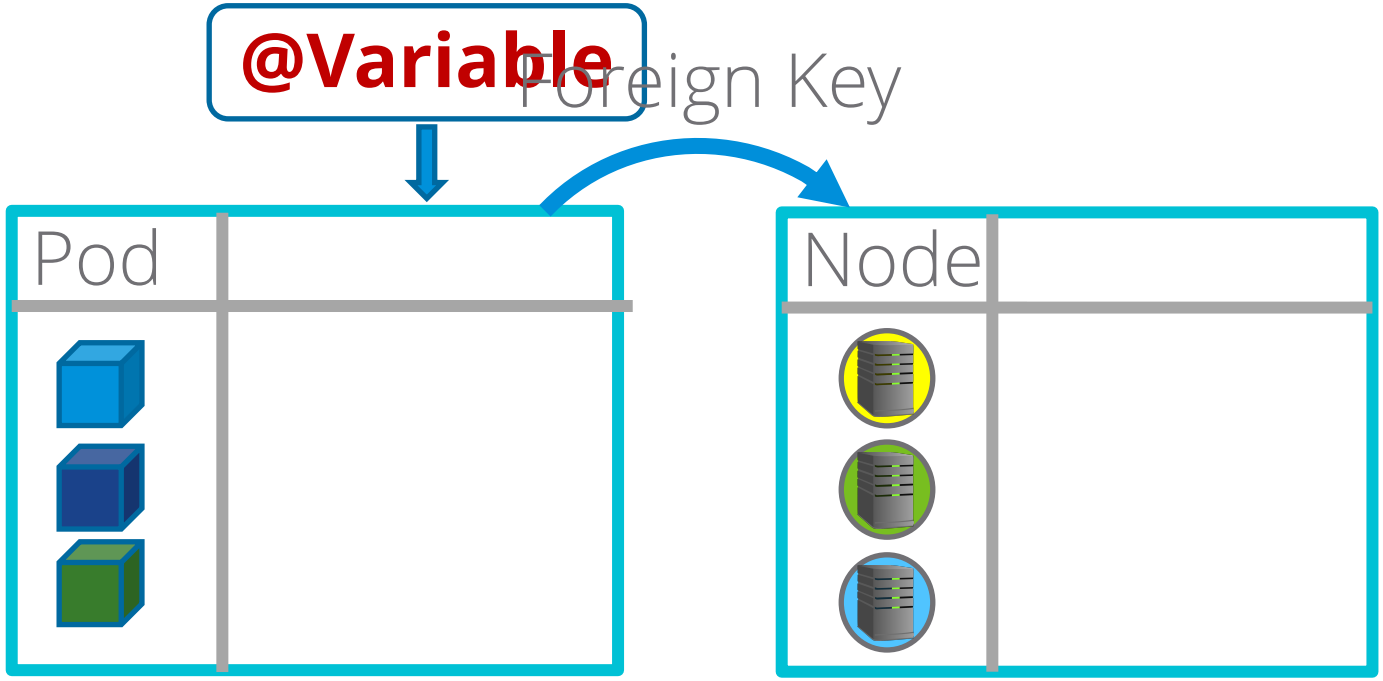
Extensibility

Policies in <20 lines of SQL
Non-trivial features (Unified Pod/VM scheduling)

Programming Model



Variable Columns



Hard Constraints

@Variable

Pod	Node
	?
	?
	?

Node	Mem Overload
	False
	True
	False

@ hard constraint

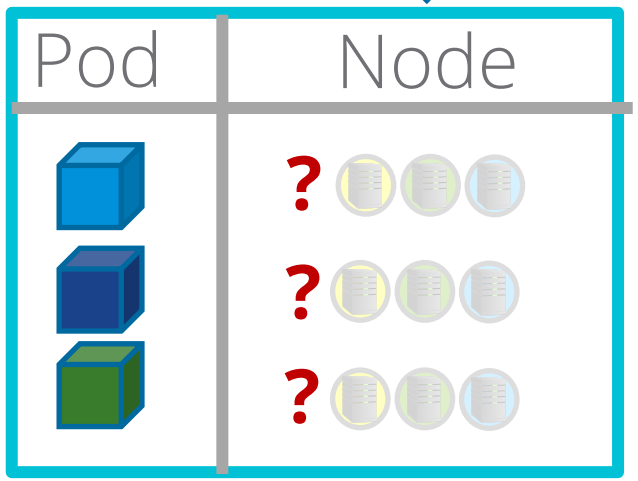
```
CREATE VIEW avoid_mem_overload AS
```

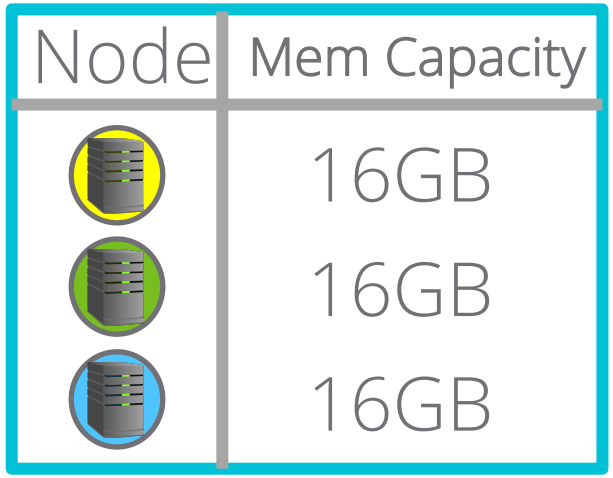
Select some rows

Predicate

Soft Constraints

@Variable





@ soft constraint

```
CREATE VIEW load_balance AS
```

Scalar expression to maximize

Programming Model

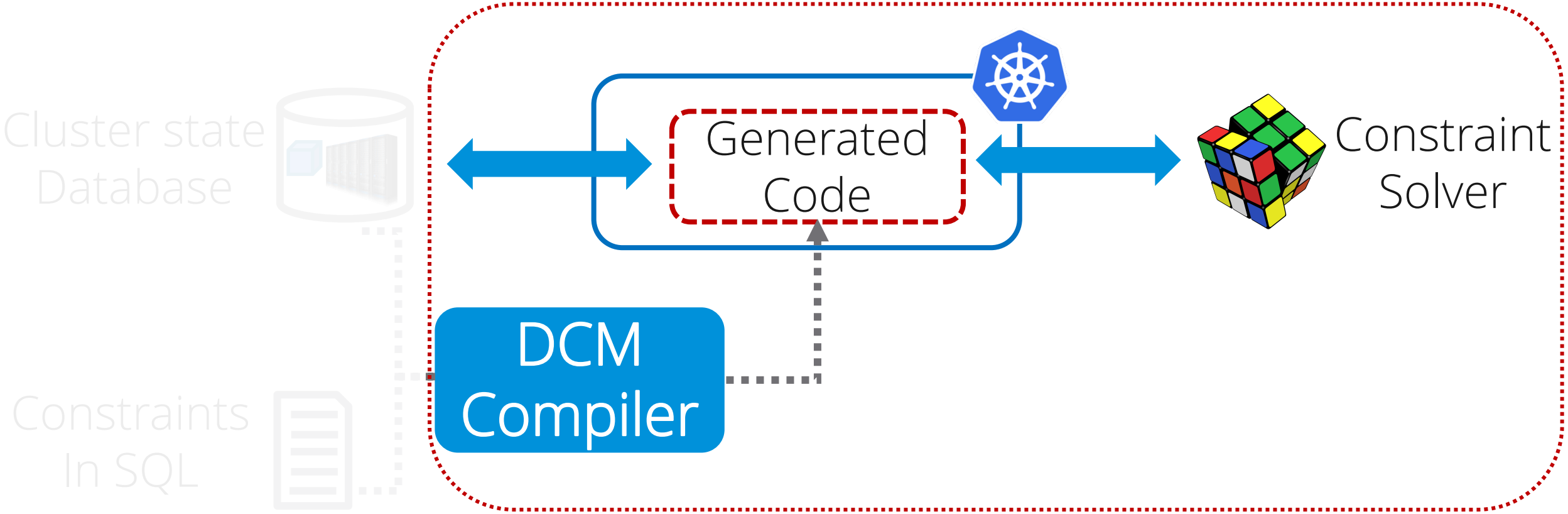
Express policies concisely using joins, aggregates, group bys, sub-queries, correlated sub-queries, arrays...

```
model = Model.create(dbConnection,
                     constraints.sql);

model.solve();
```

Instantiate different models for different tasks and timescales

DCM Compiler



```
create view constraint_1 as
select * from t1 join t2 on t1.b = t2.b
where t2.e == 10
check(t1.c * t2.d = t2.c)
```

DCM Compiler

Flagship backend

Generates Java code that interfaces with Google OR-Tools CP-SAT solver

```
create view constraint_1 as
select * from t1 join t2 on t1.b = t2.b
where t2.e == 10
check(t1.c * t2.d = t2.c)
```

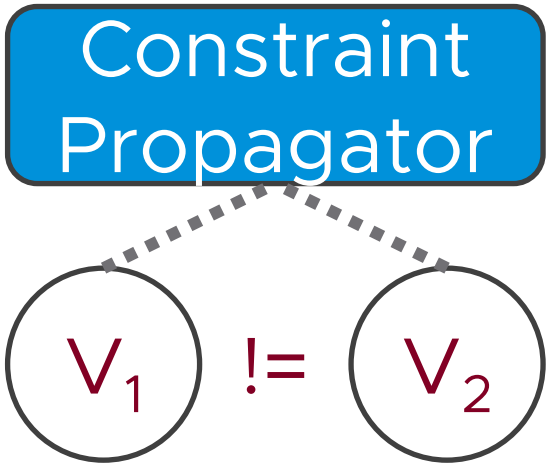
DCM Compiler

Iterate efficiently over tables

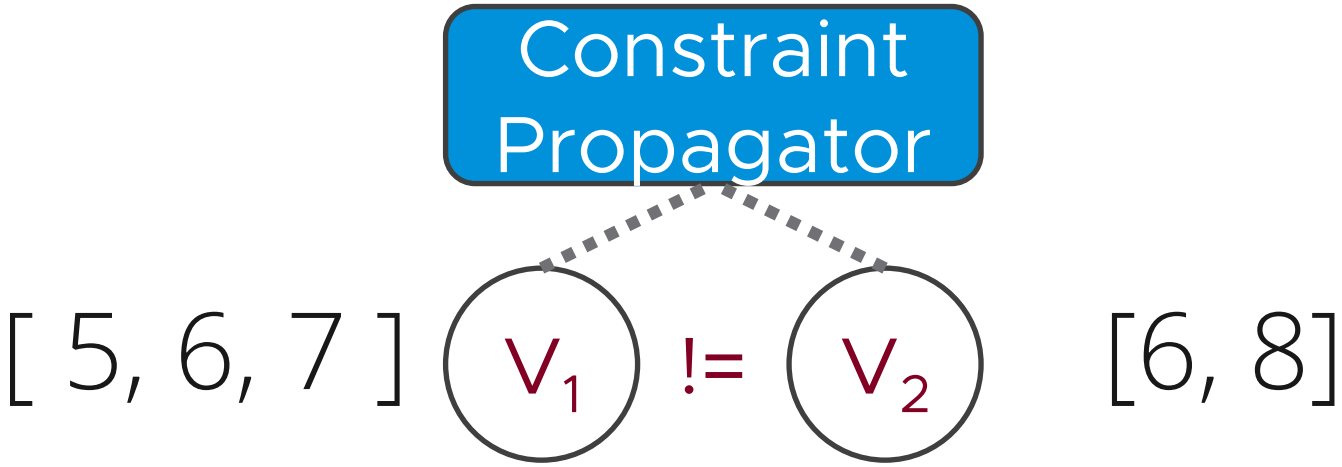
Filter out rows

Encode constraints

Solver performance is highly sensitive to the encoding



Solver performance is highly sensitive to the encoding



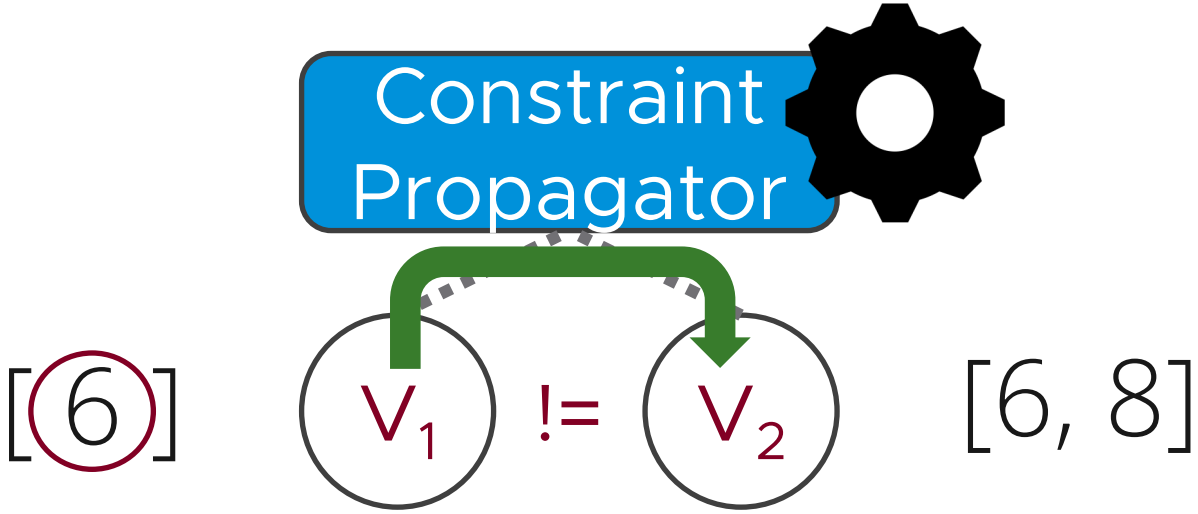
Solver performance is highly sensitive to the encoding

Solver fixes to 6

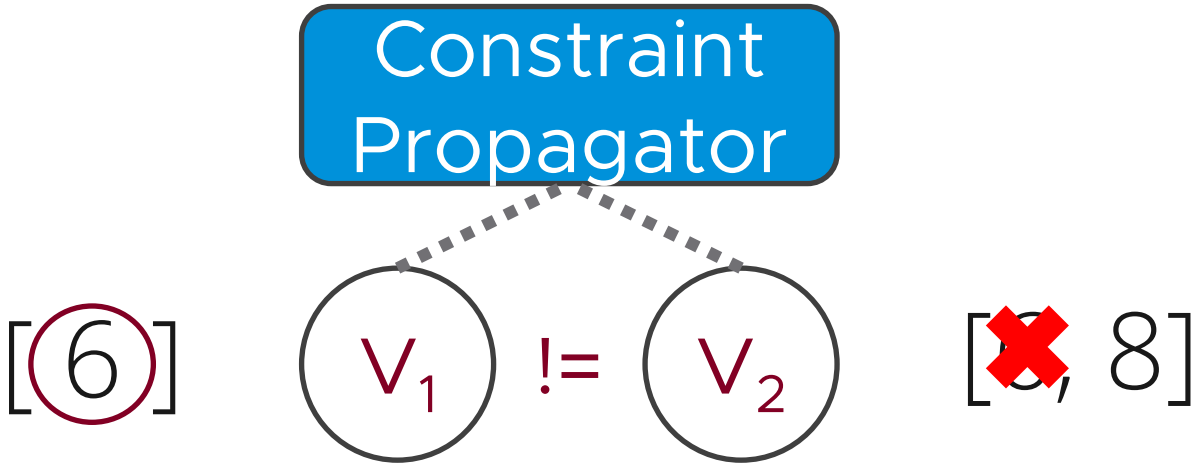
[6]

v_1

Solver performance is highly sensitive to the encoding



Solver performance is highly sensitive to the encoding



Solver performance is highly sensitive to the encoding

- Reduce number of introduced variables and constraints
- Leverage specialized algorithms (i.e., *global constraints*)

Benchmark
Assign 50 tasks to 1000 workers

Naïve: 25 seconds

With optimizations: 85 ms!

Evaluation

Use cases

Kubernetes Scheduler	VM Load Balancing Tool	Distributed Transactional Datastore

Scalability

Decision quality

Extensibility

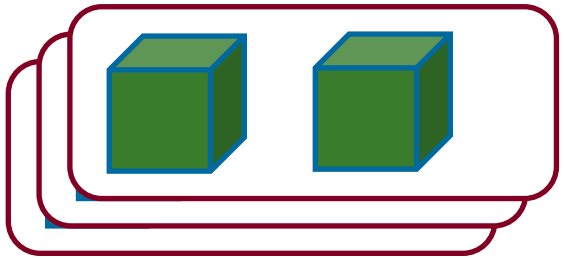
Evaluation

Kubernetes Scheduler

Scalability

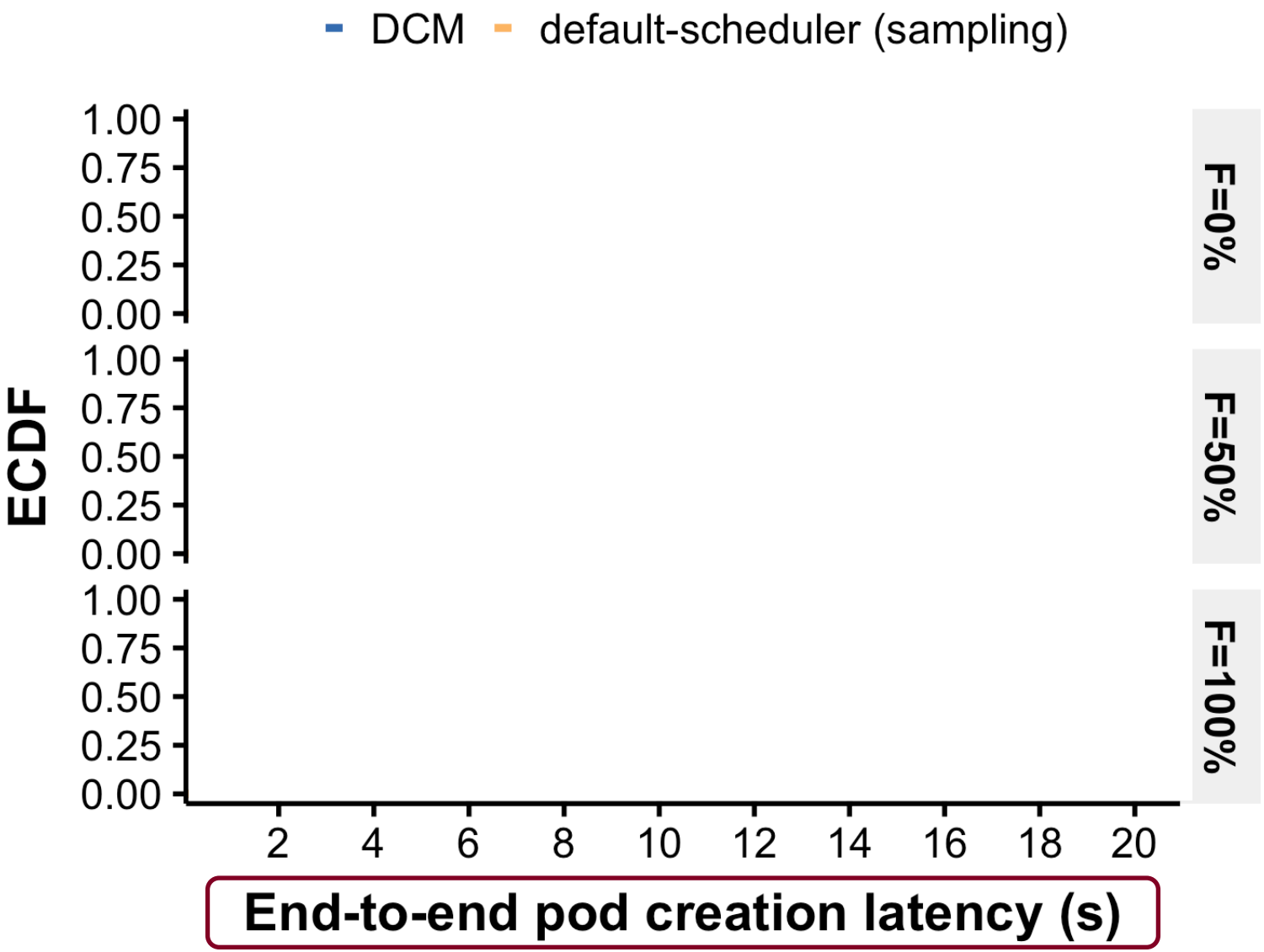
- 500 node Kubernetes cluster
- Deploy a series of apps in an open-loop
- Azure 2019 trace
- Inter-pod anti-affinity constraint

Do **NOT** place us on the same node!

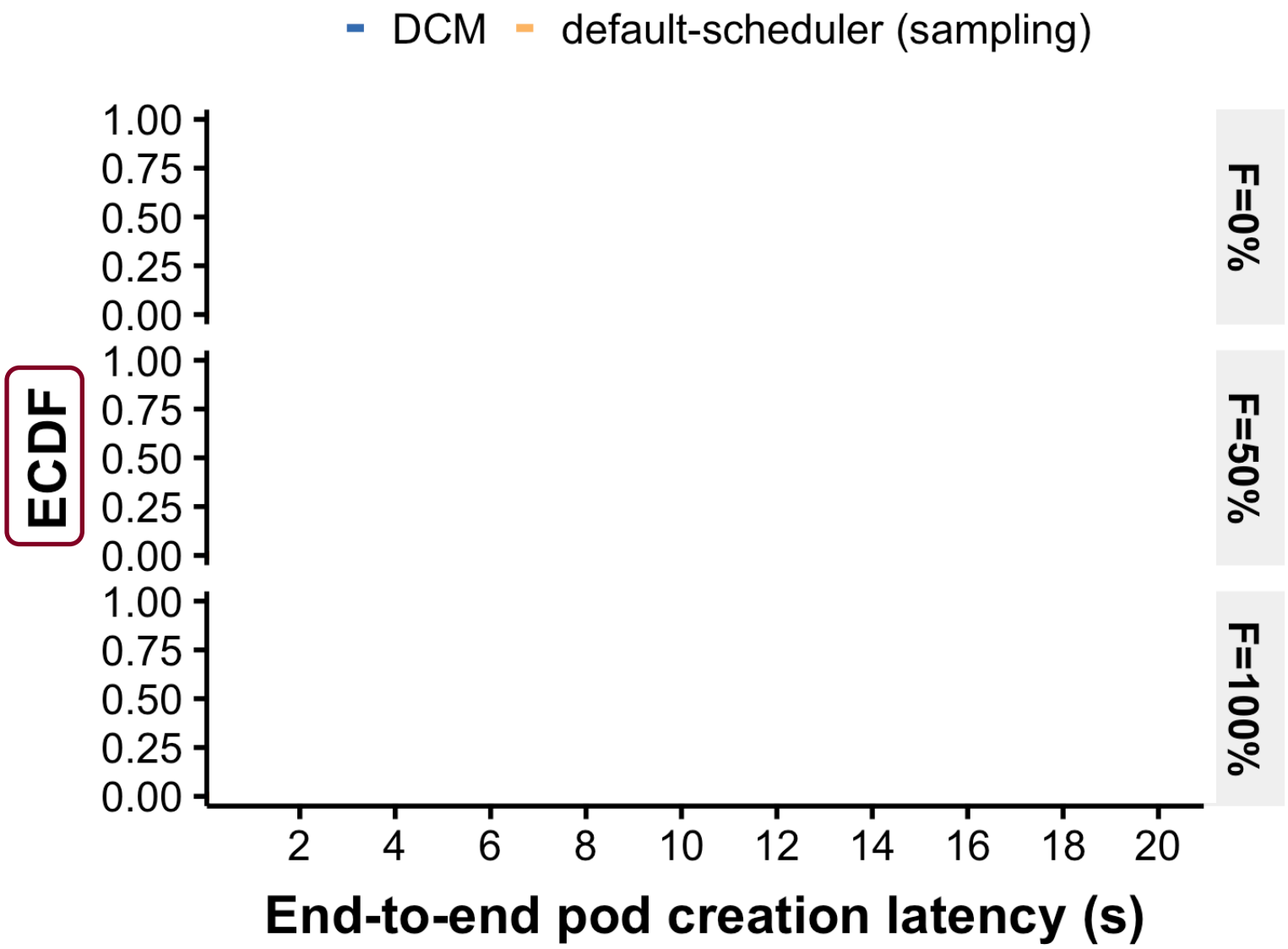


Recommended best practice, but a challenging constraint

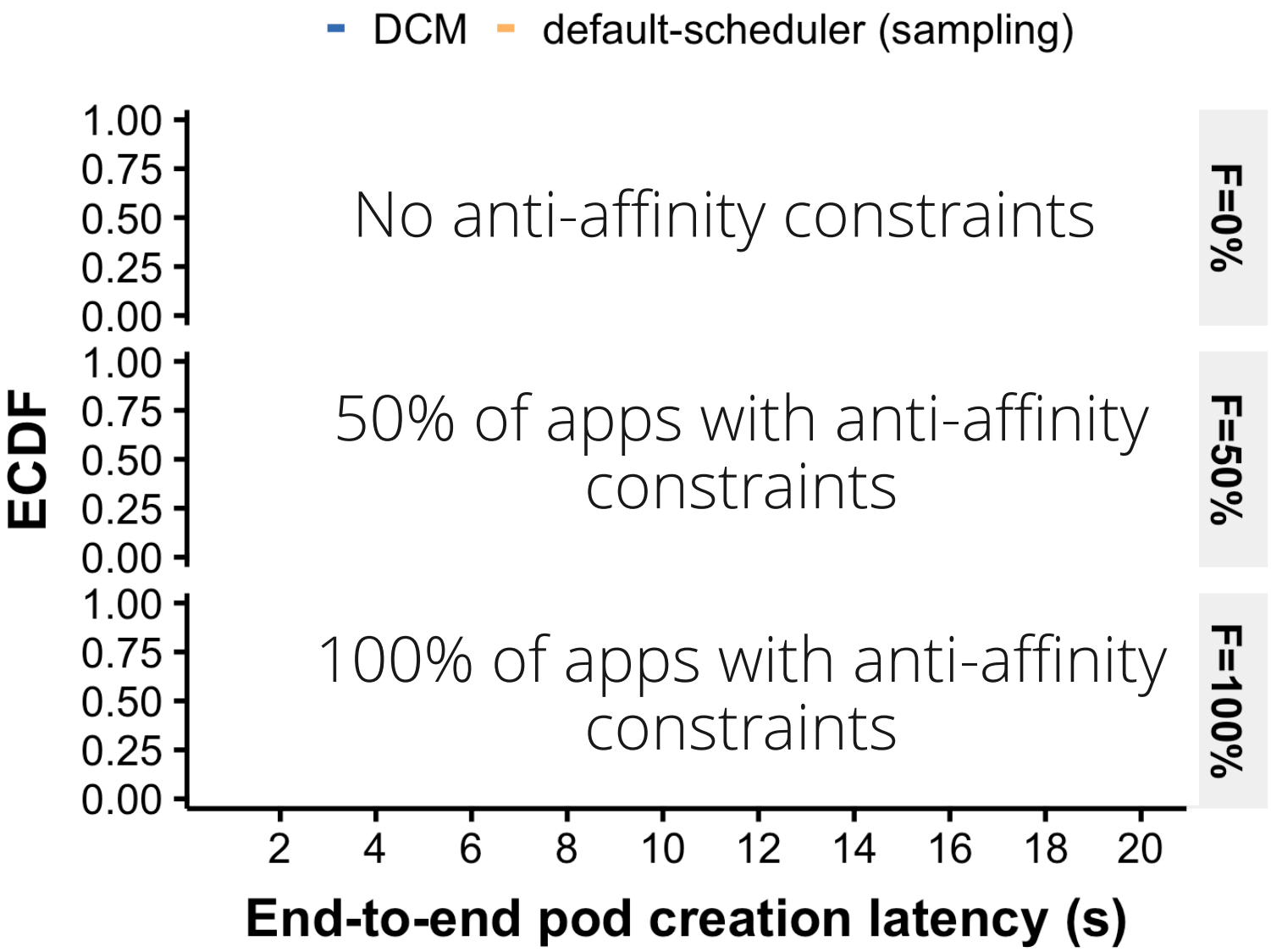
Kubernetes Scalability Evaluation



Kubernetes Scalability Evaluation



Kubernetes Scalability Evaluation

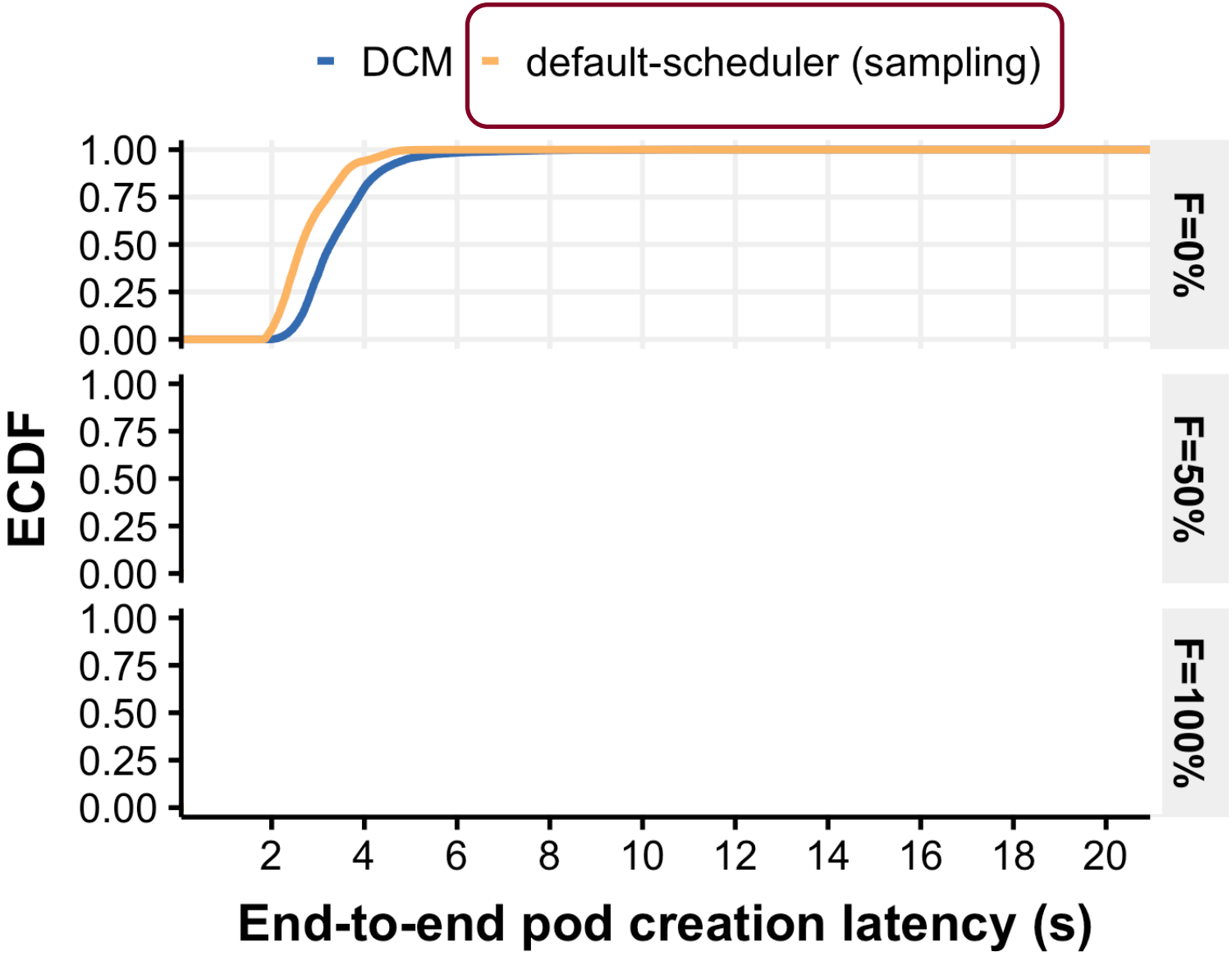


Kubernetes Scalability Evaluation

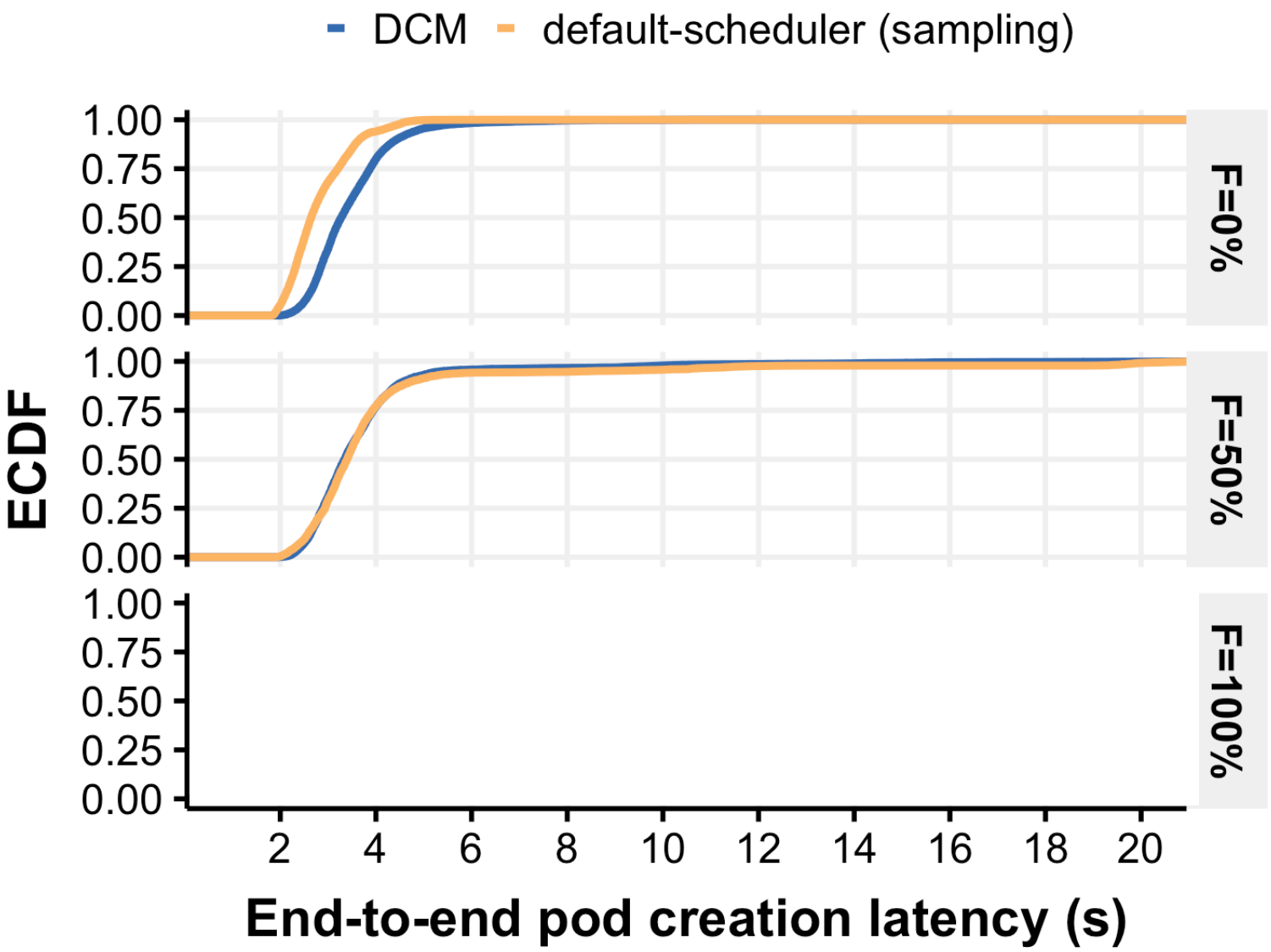
Baseline samples only 50% of nodes

p95 latency

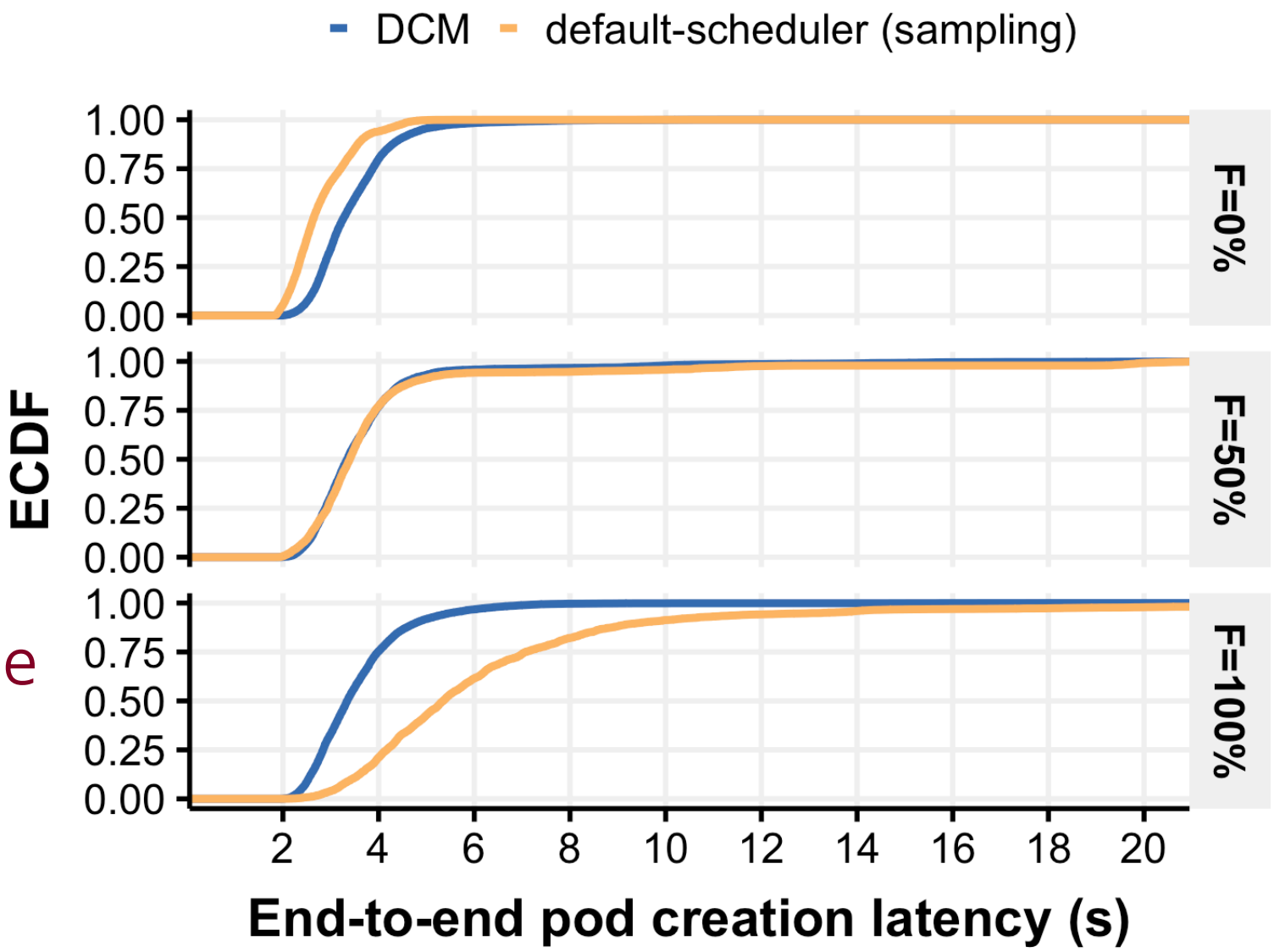
DCM = 5.33s
Baseline = 4.13s



Kubernetes Scalability Evaluation



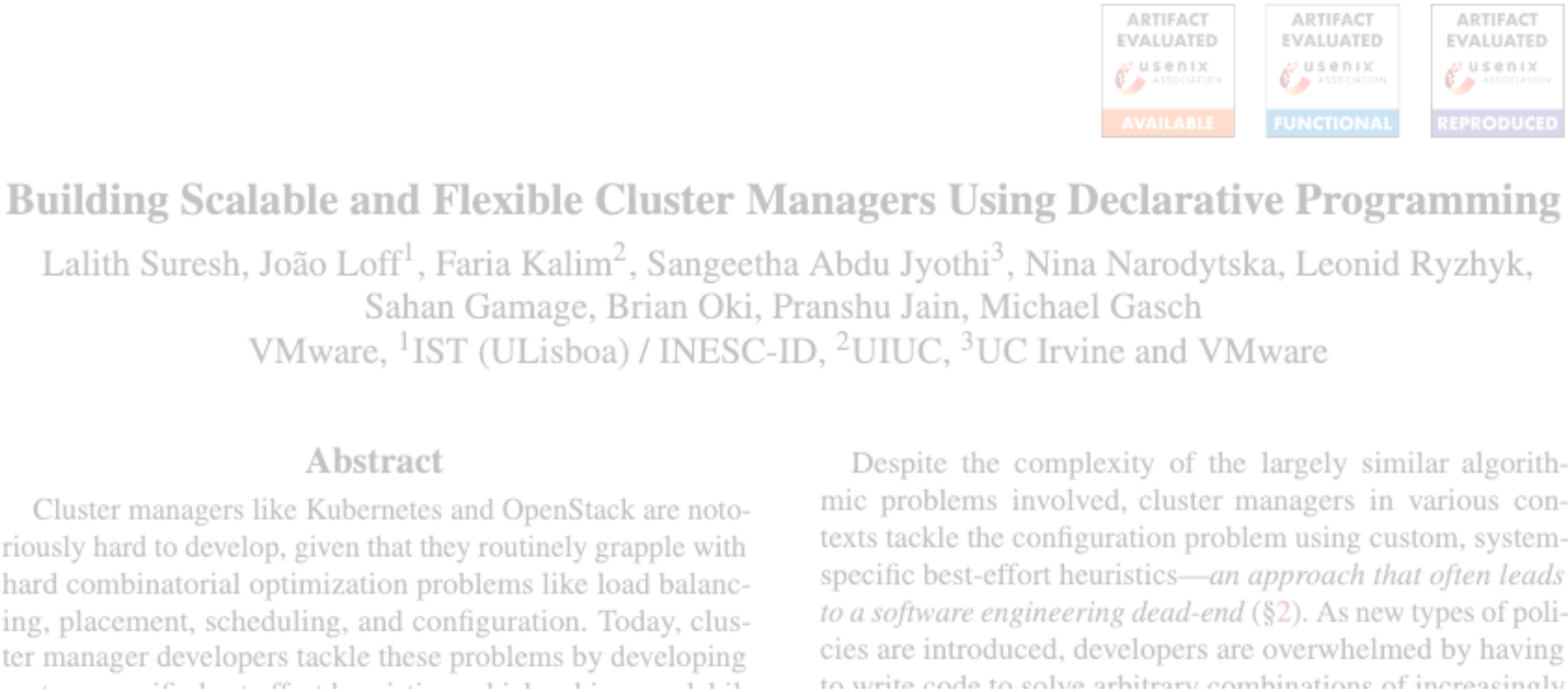
Kubernetes Scalability Evaluation



DCM cuts 95th percentile latency in half

More details in the paper!

Compiler internals, debugging, lessons learnt,
DCM's generality and limitations…



Kubernetes Scheduler

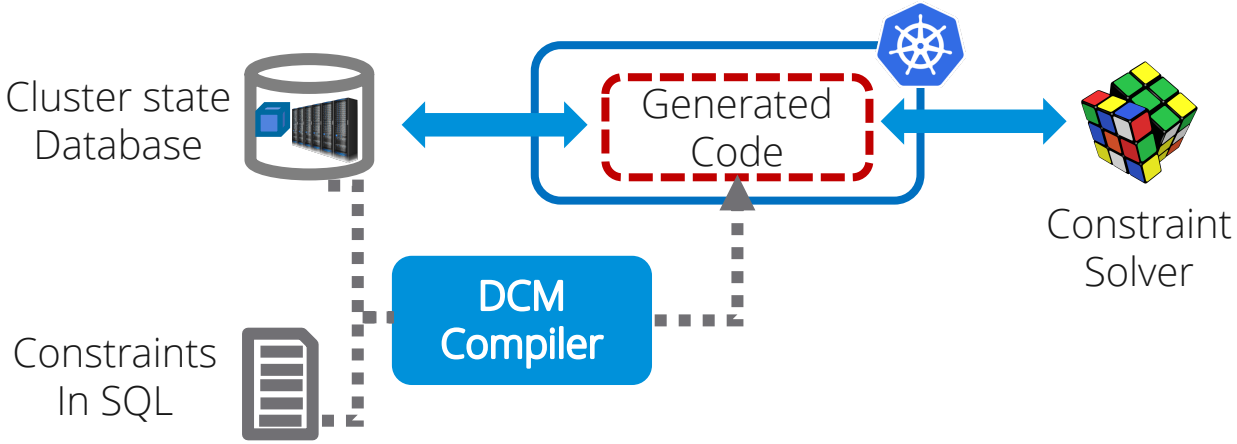
VM Load Balancing Tool

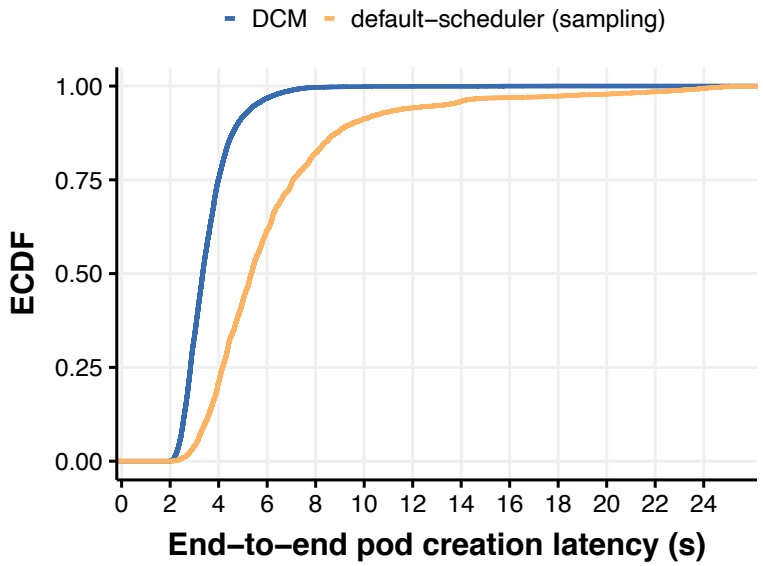
Distributed Transactional Datastore

Scalability

Decision quality

Extensibility





Thank you!

lsuresh@vmware.com

Code: https://github.com/vmware/declarative-cluster-management/