Response Time-Optimized Distributed Cloud Resource Allocation

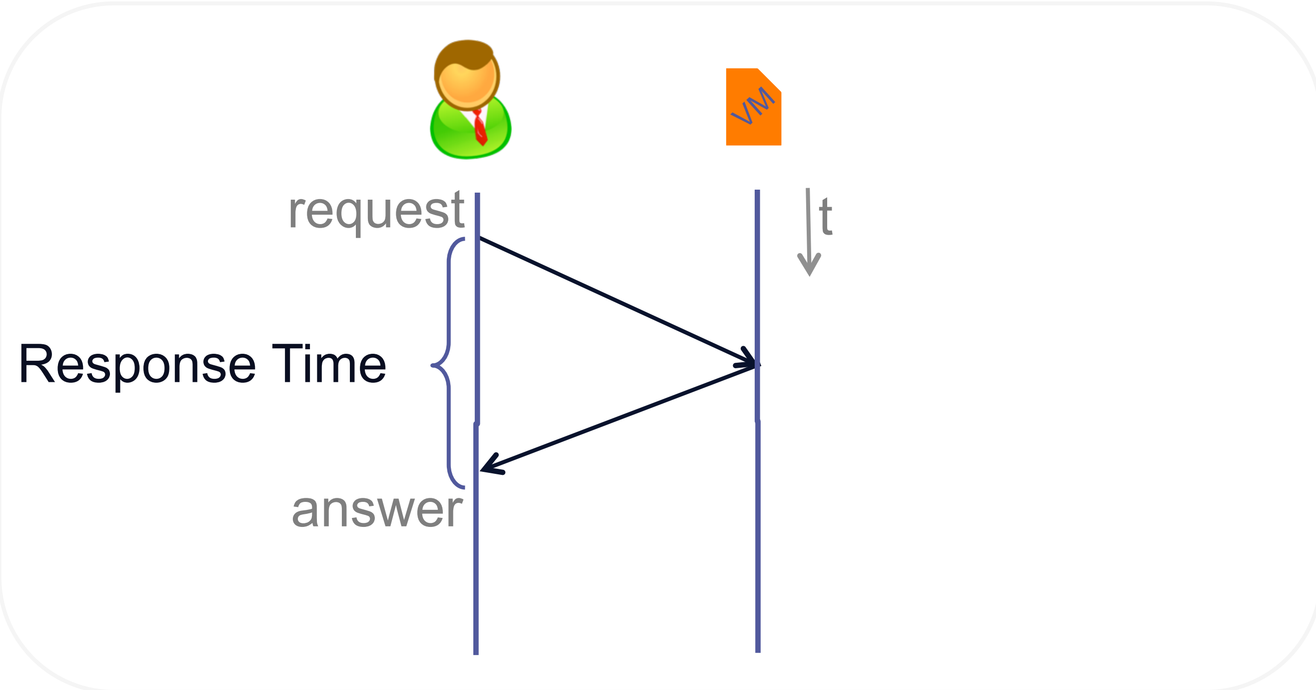
Matthias Keller

Holger Karl

Computer Networks Group
Universität Paderborn

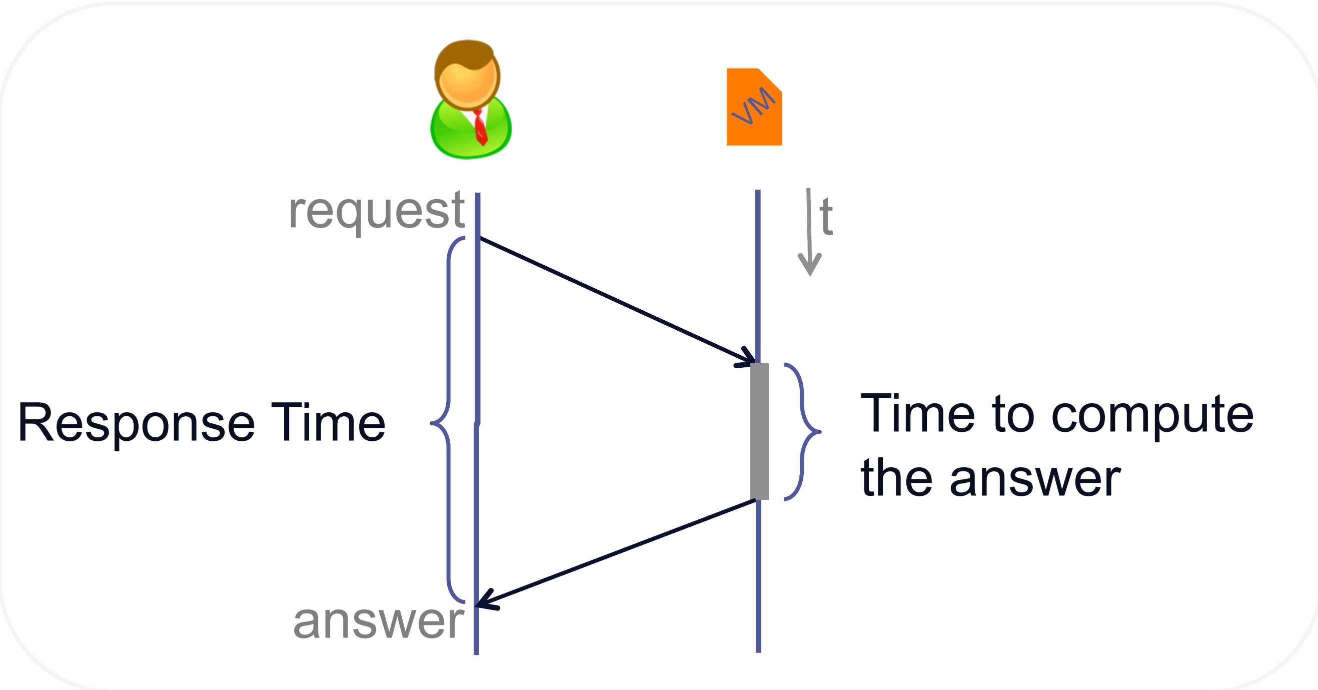
Minimizing response times

- Latency-critical service
  - Interactive, emergency service

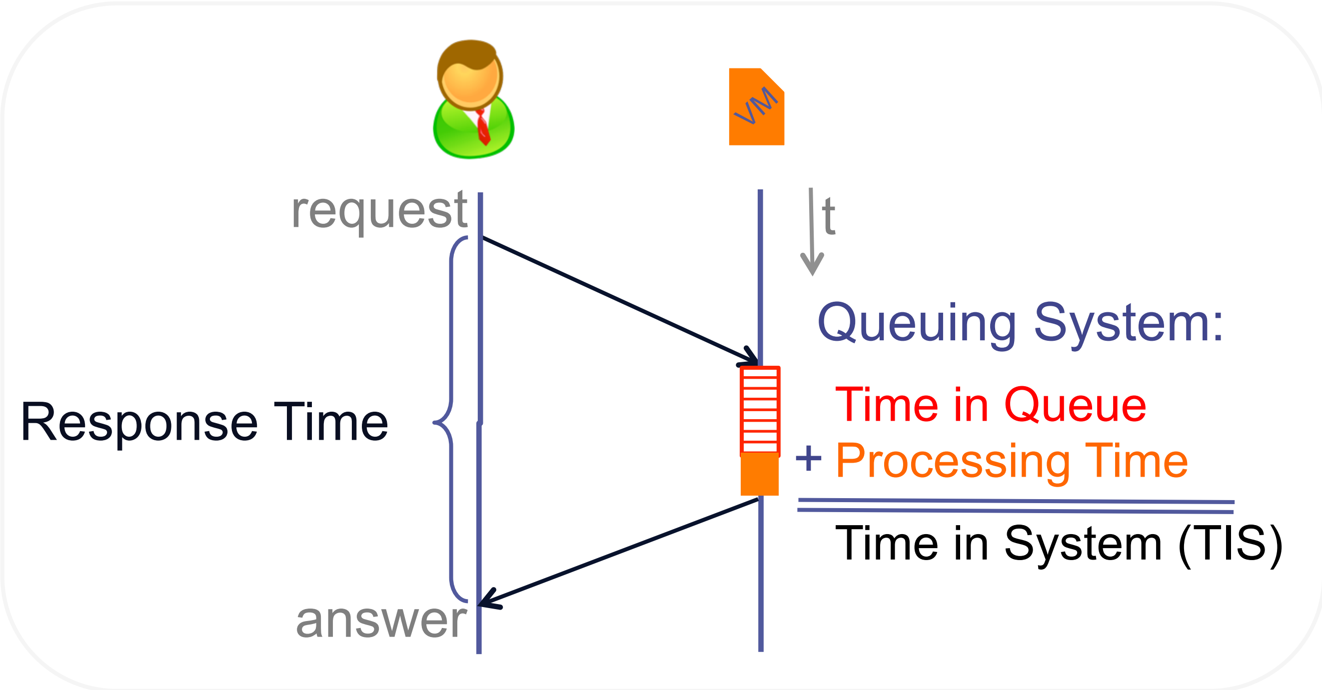


Minimizing response times

- Latency-critical service
  - Interactive, emergency service

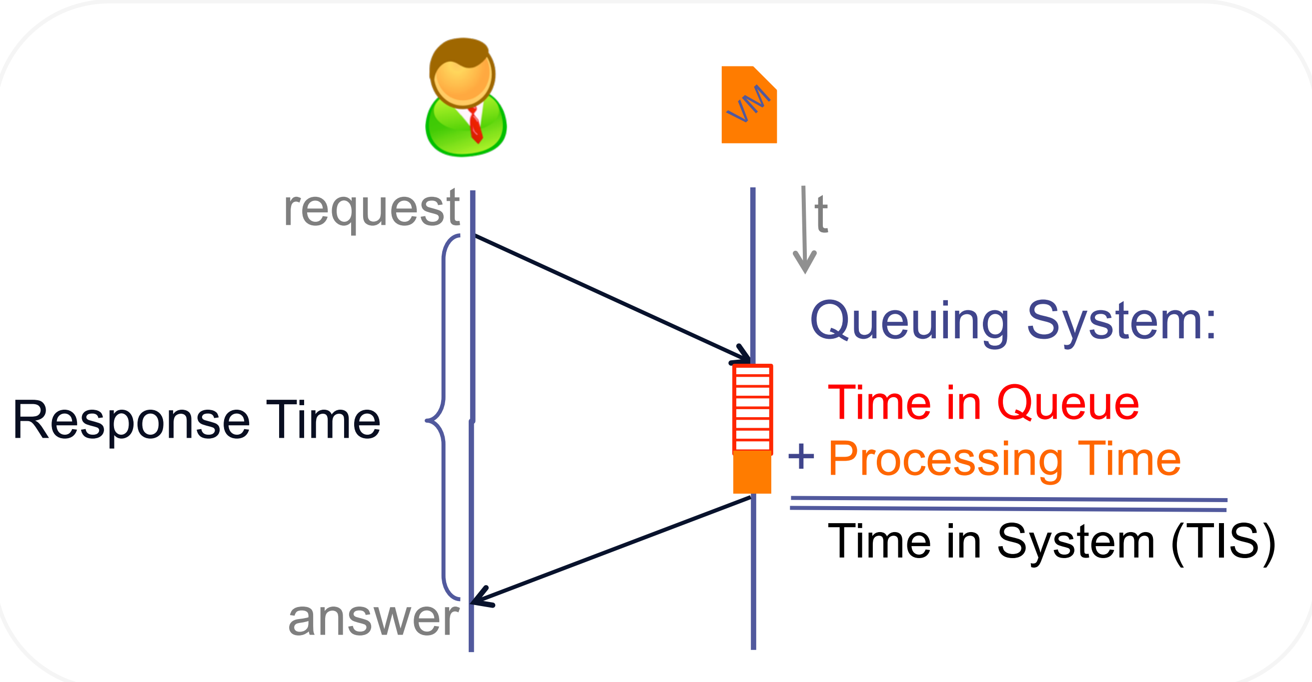


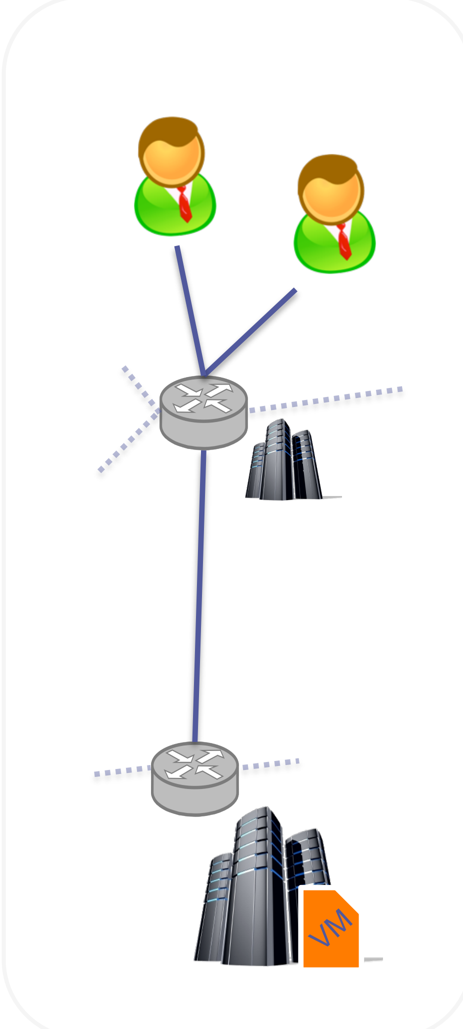
Minimizing response times



Minimizing response times

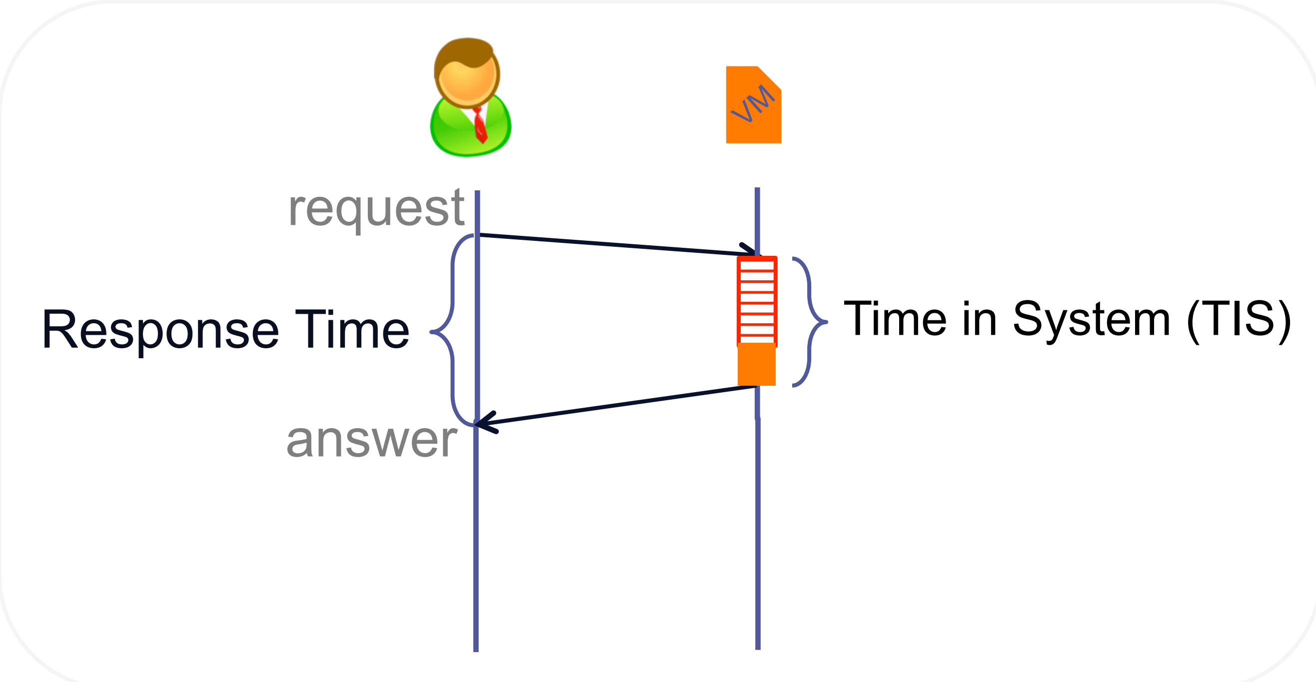
- Latency-critical service
  - Interactive, emergency service
- Decision: Spend time on RTT or TIS

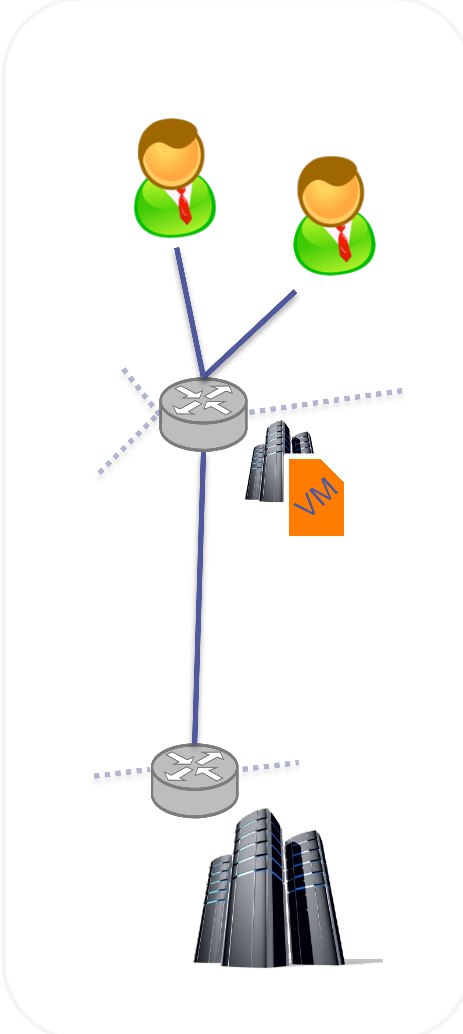




Minimizing response times

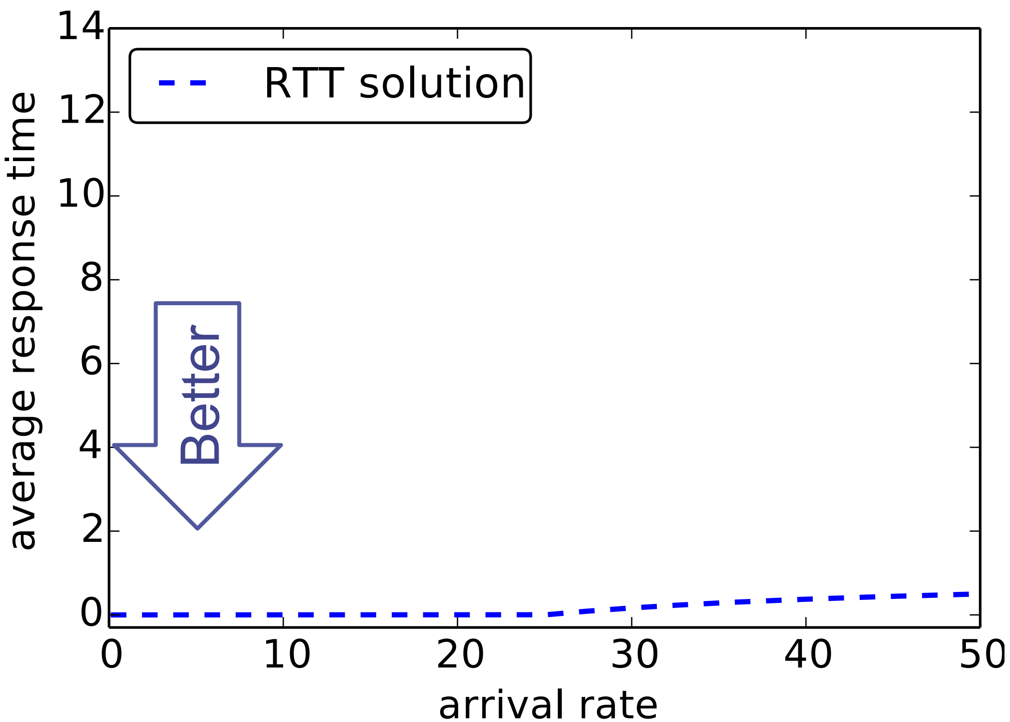
- Latency-critical service
  - Interactive, emergency service
- Decision: Spend time on RTT or TIS

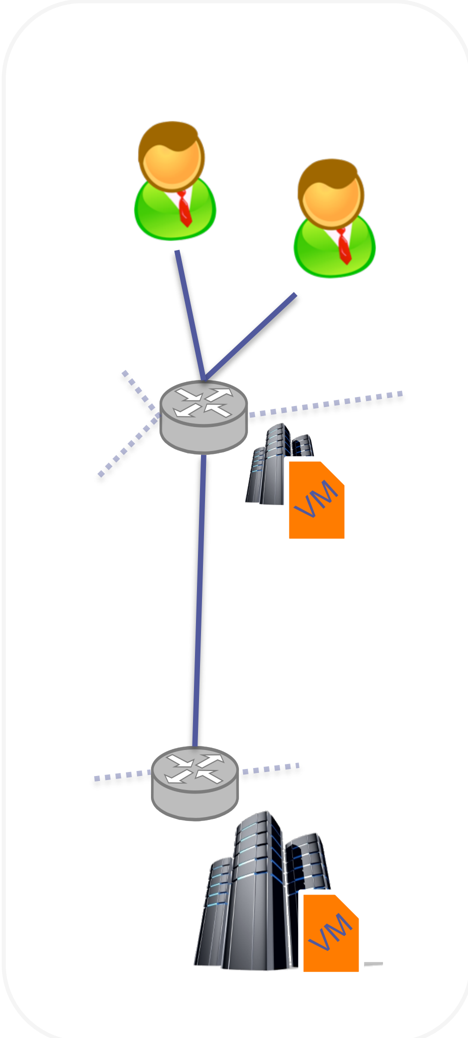




Example: RTT + TIS

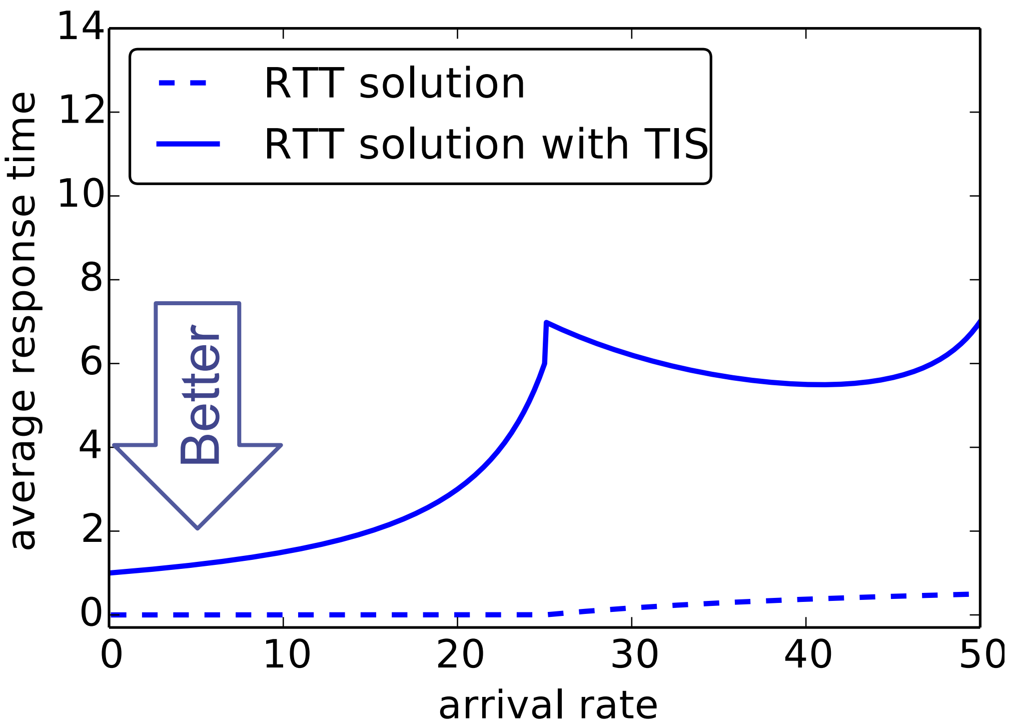
- Demand assignment
  - Facility Location Solution with RTT only

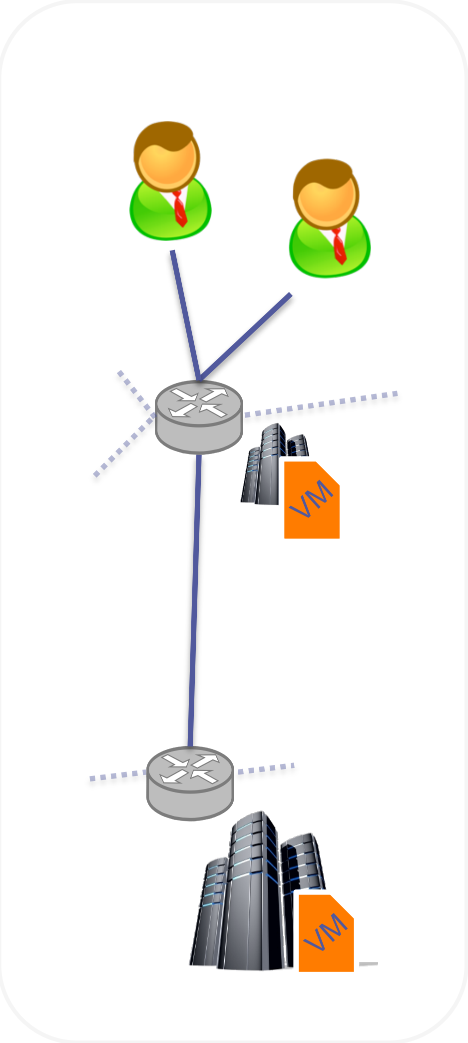




Example: RTT + TIS

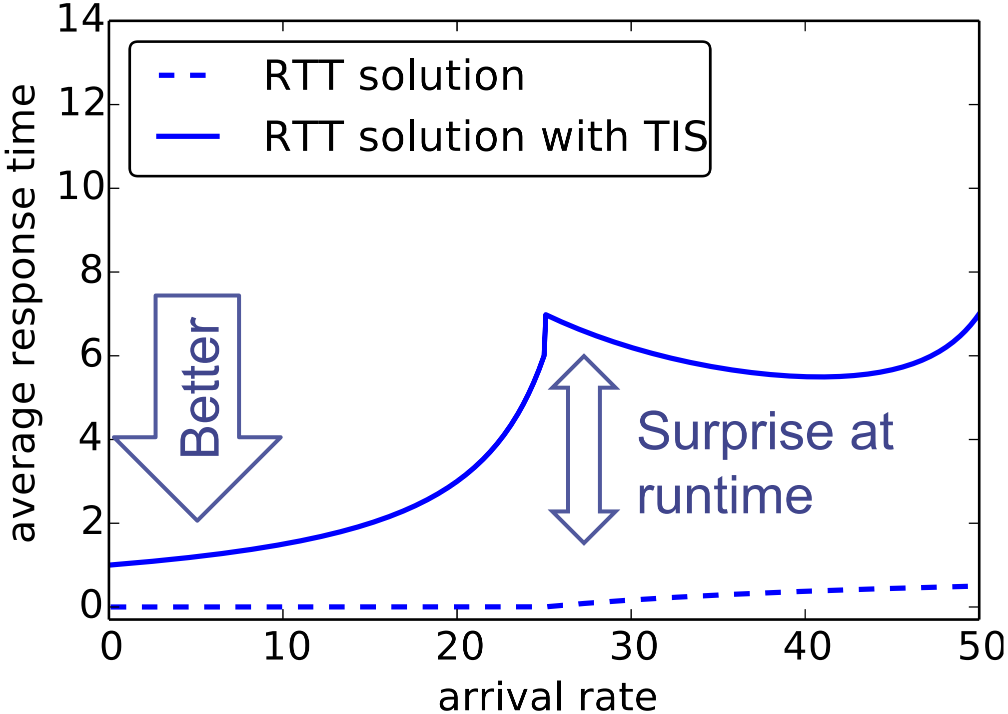
- Demand assignment
  - Facility Location Solution with RTT only

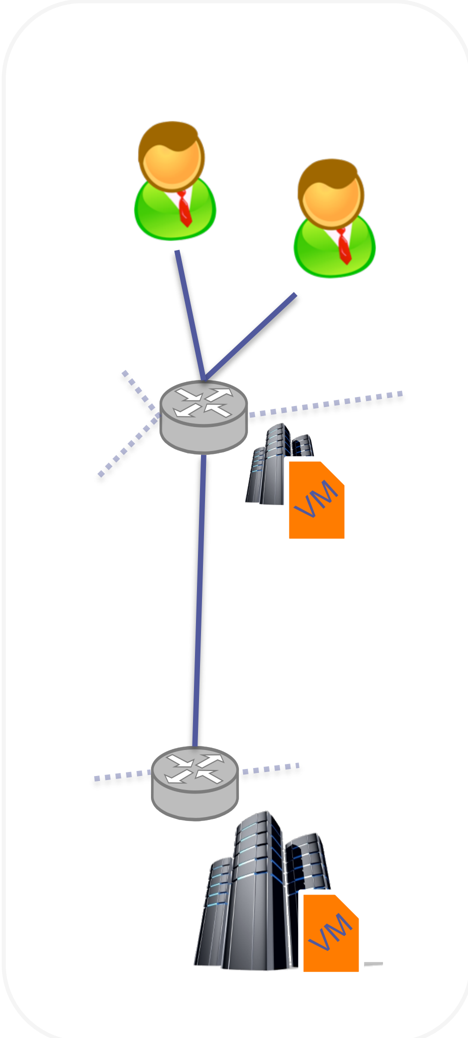




Example: RTT + TIS

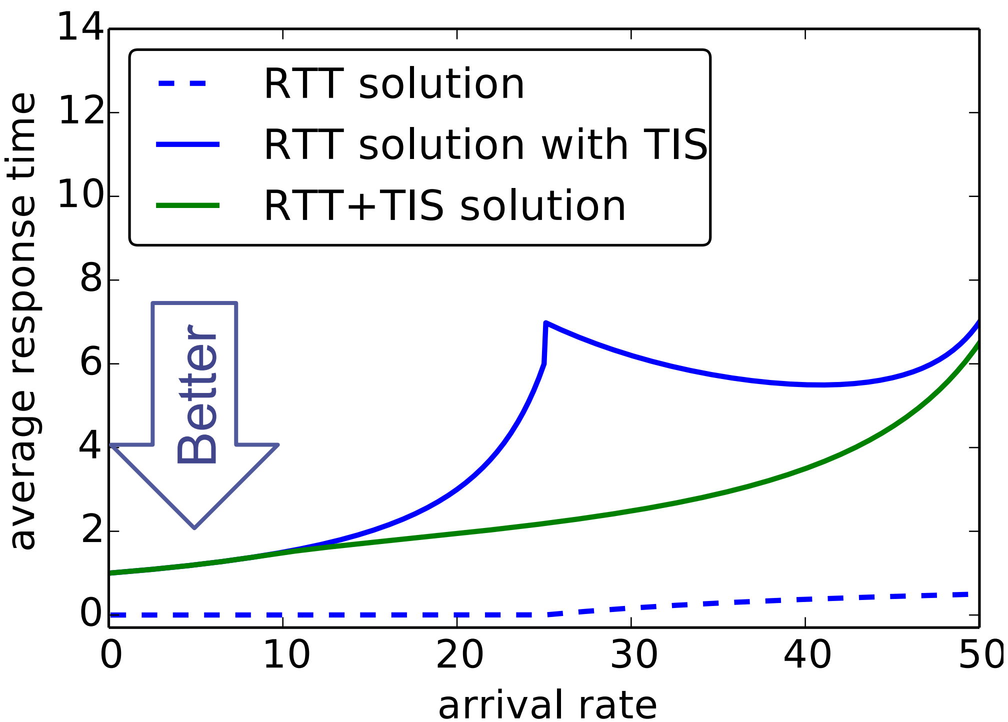
- Demand assignment
  - Facility Location Solution with RTT only

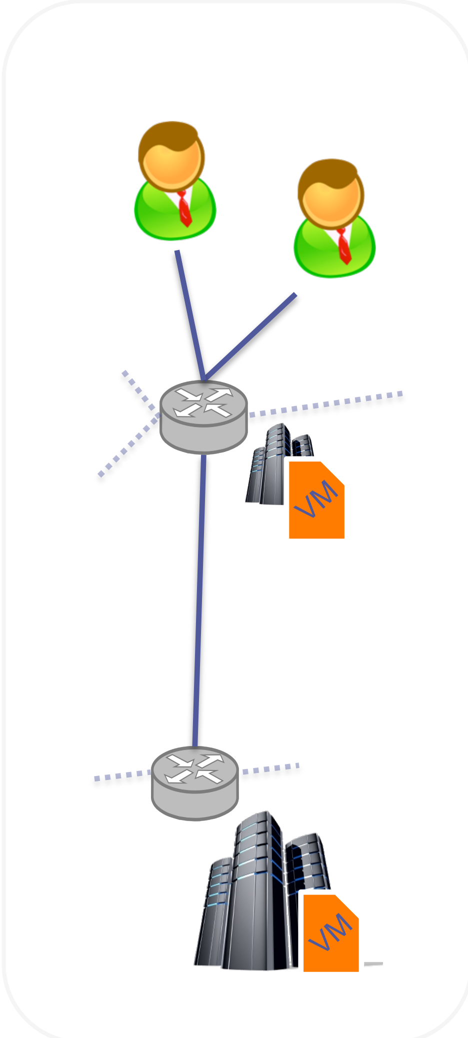




Example: RTT + TIS

- Demand assignment
  - Facility Location Solution with RTT only
  - With RTT + TIS





Goal

Given

- Network
- Data centres

Objective

- Minimize response time

Means

- Allocation of *n* VMs at data centres

Goal

Given

- Network
- Data centres

Objective

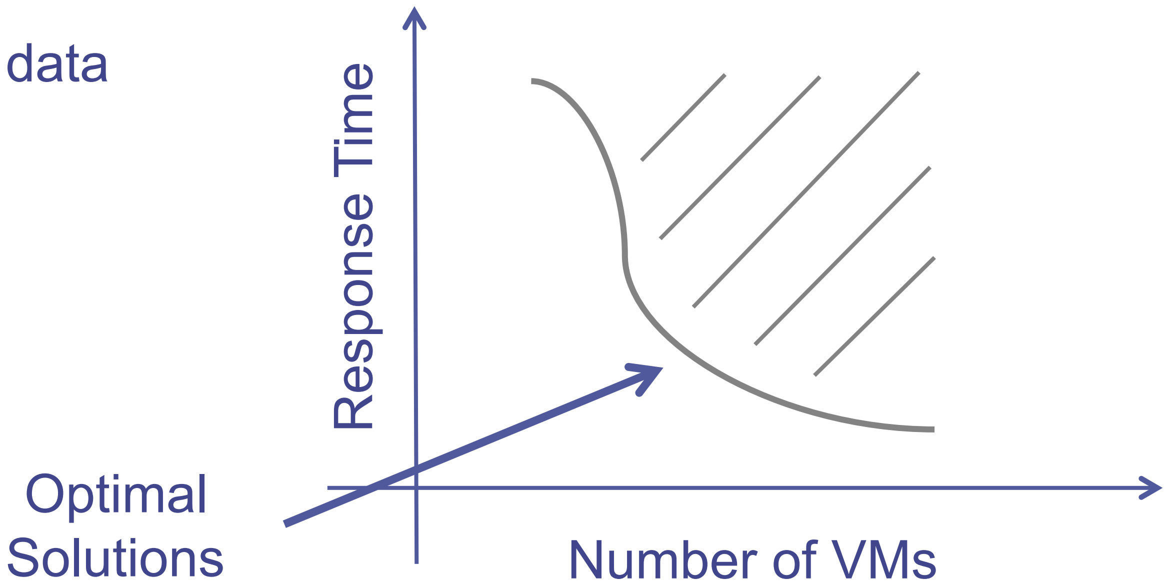
- Minimize response time

Means

- Allocation of *n* VMs at data centres

Characterise:

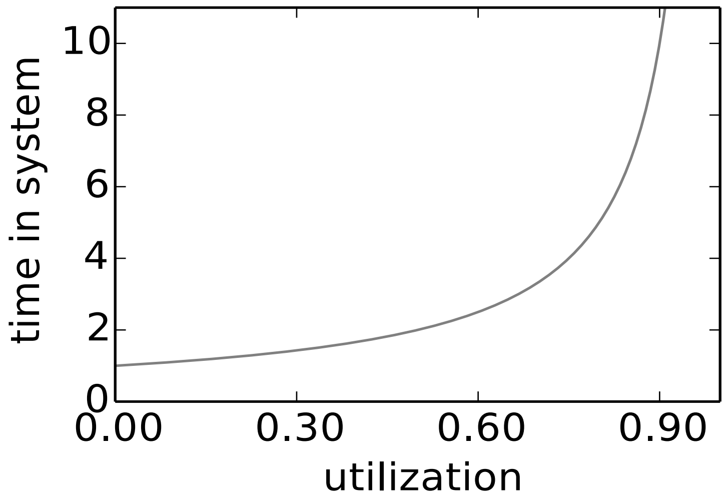
- How does response time depend on number *n* of VMs?



Two Approaches

Accurate Solution

- Mixed Integer **Convex** Problem
- Convex TIS function for each data centre

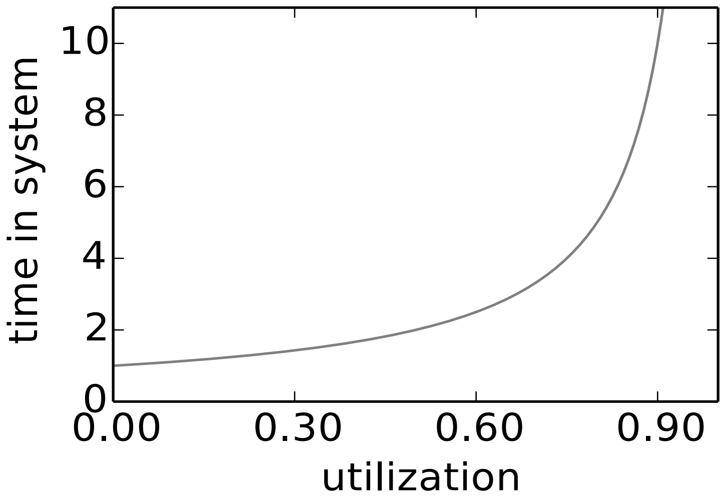


- Tough to solve – slow?

Two Approaches

Accurate Solution

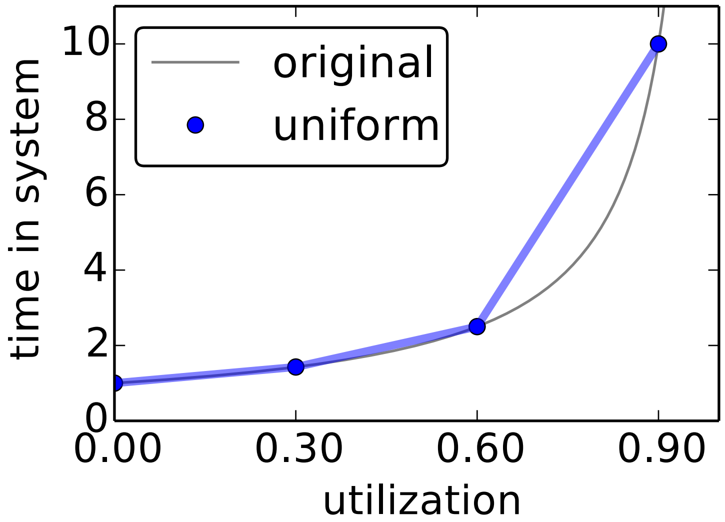
- Mixed Integer **Convex** Problem
- Convex TIS function for each data centre



- Tough to solve – slow?

Approximate Solution

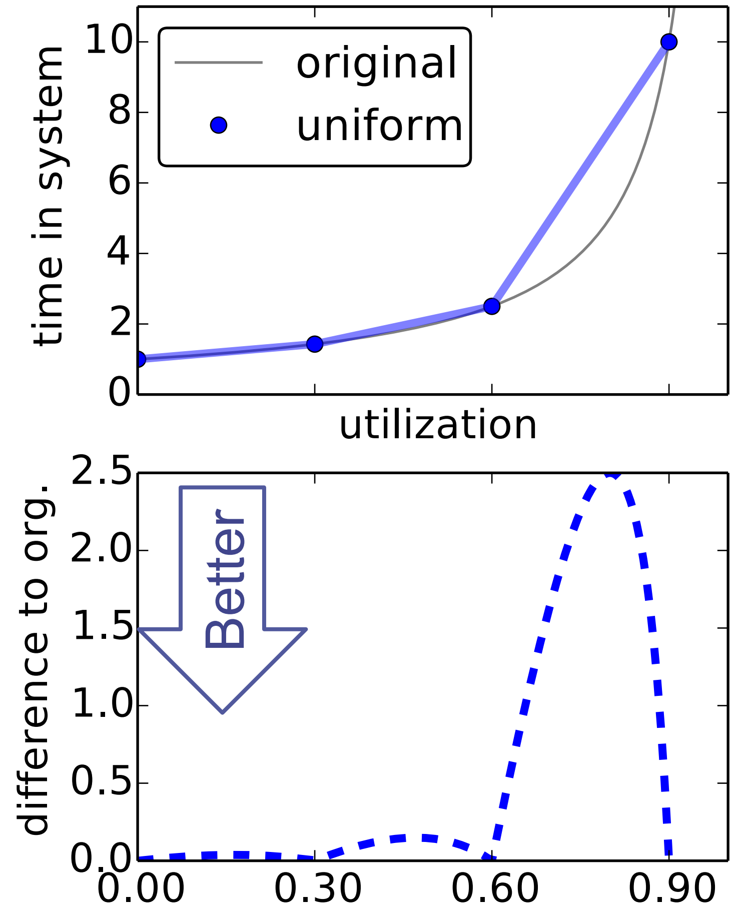
- Reformulation: Mixed Integer **Linear** Problem
- Linearization of TIS function



- Accuracy? Speed?

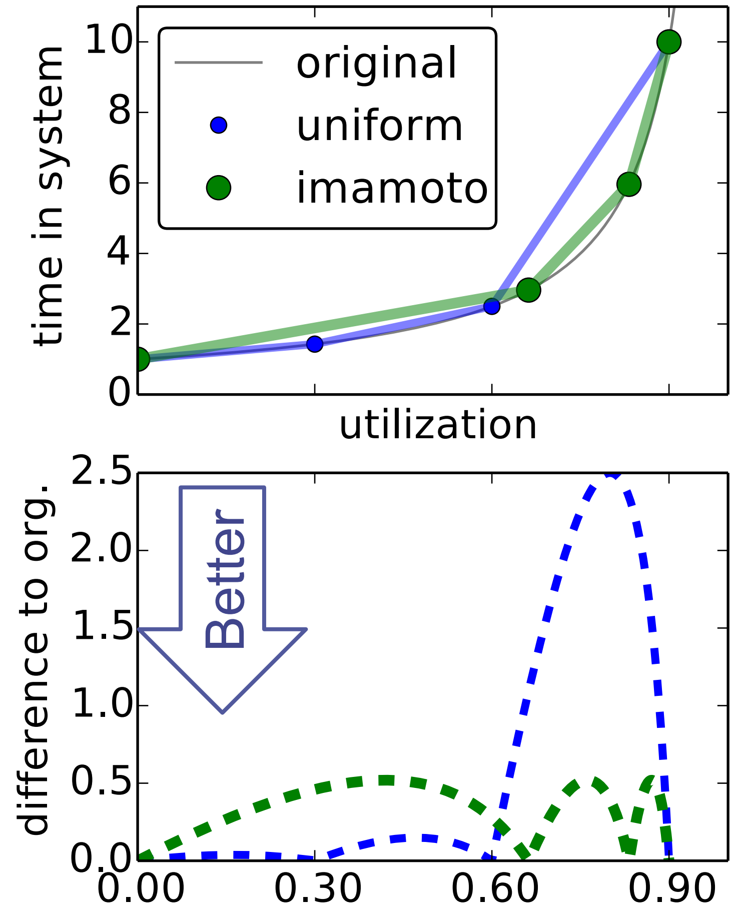
Improve accuracy of linearization

- Objective:
  - Minimize the maximum difference
- Control knobs
  - Number of basepoints
  - End point at asymptote
  - Basepoint positions



Improve accuracy of linearization

- Objective:
  - Minimize the maximum difference
- Control knobs
  - Number of basepoints
  - End point at asymptote
  - Basepoint positions
- Evaluation in Paper



Evaluation of both approaches

Convex Problem

- Reference Solution
- Tough to solve – slow?

Linear Problem

- Approximate Solution
- Accuracy? Speed?

Linearization

Evaluation of both approaches

Convex Problem

- Reference Solution
- Tough to solve – slow?

Linear Problem

- Approximate Solution
- Accuracy? Speed?

Linearization

Configurations

- 6 topologies, 12 – 54 nodes
- à 50 random demand realizations
- 10 data centre fix

Evaluation of both approaches

Convex Problem
- Reference Solution
- Tough to solve – slow?

Linear Problem
- Approximate Solution
- Accuracy? Speed?

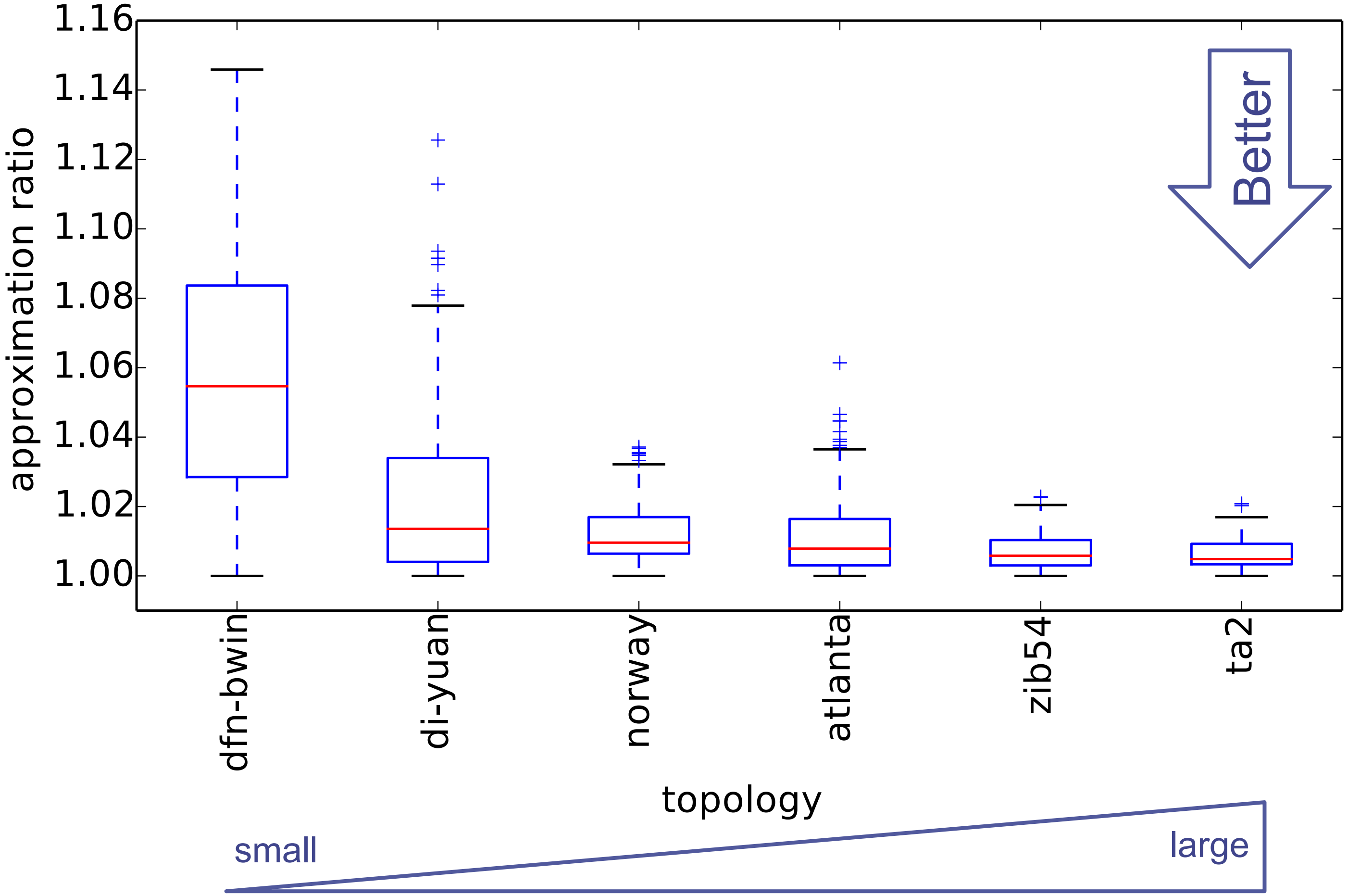
Linearization

VM limit: 5 – 10

Configurations
- 6 topologies, 12 – 54 nodes
- à 50 random demand realizations
- 10 data centre fix

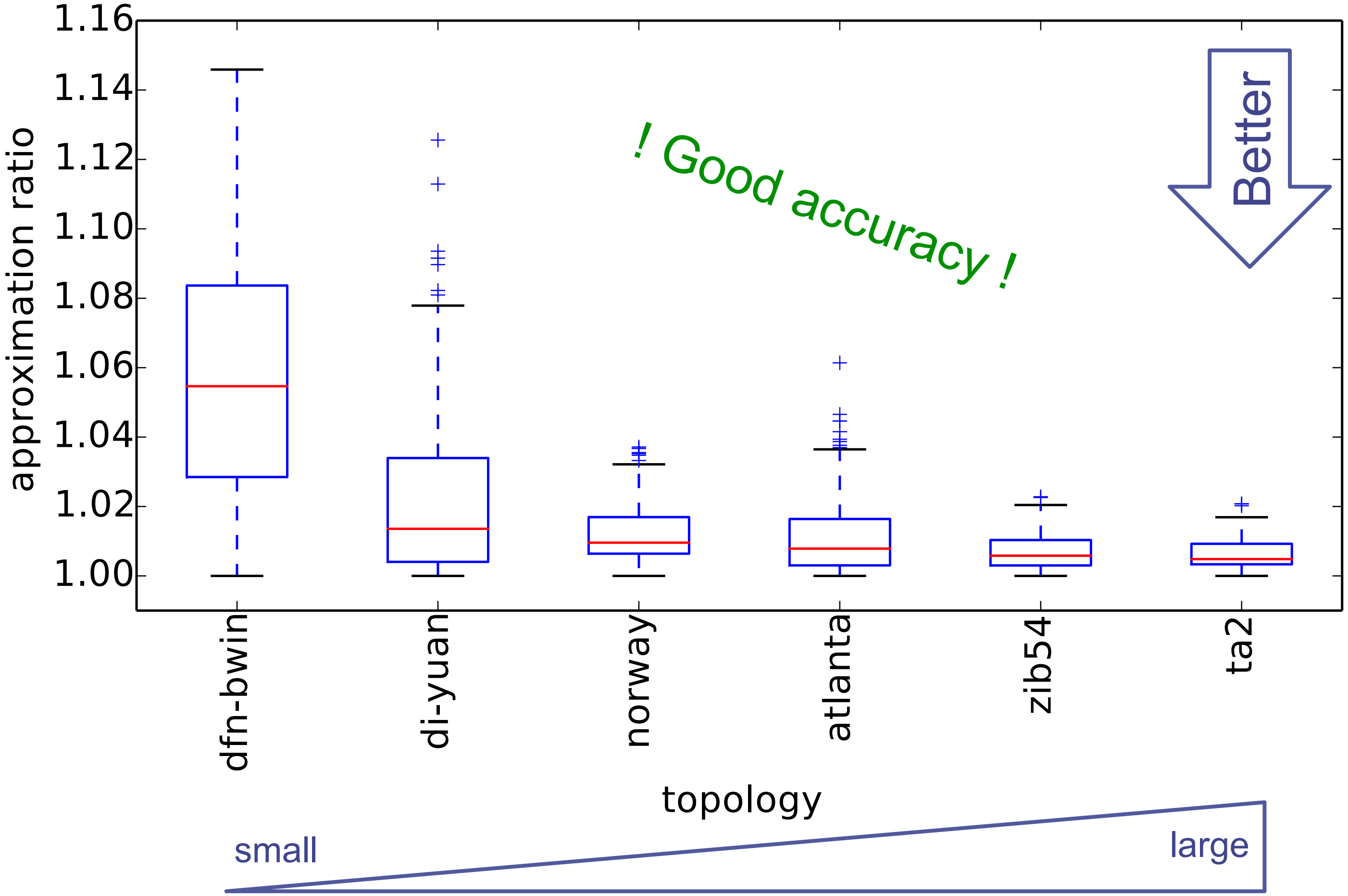
Results – Approximation Ratio

$$\text{approx. ratio} = \frac{\text{Resp.time}_{\text{Linear}}}{\text{Resp.time}_{\text{Convex}}}$$



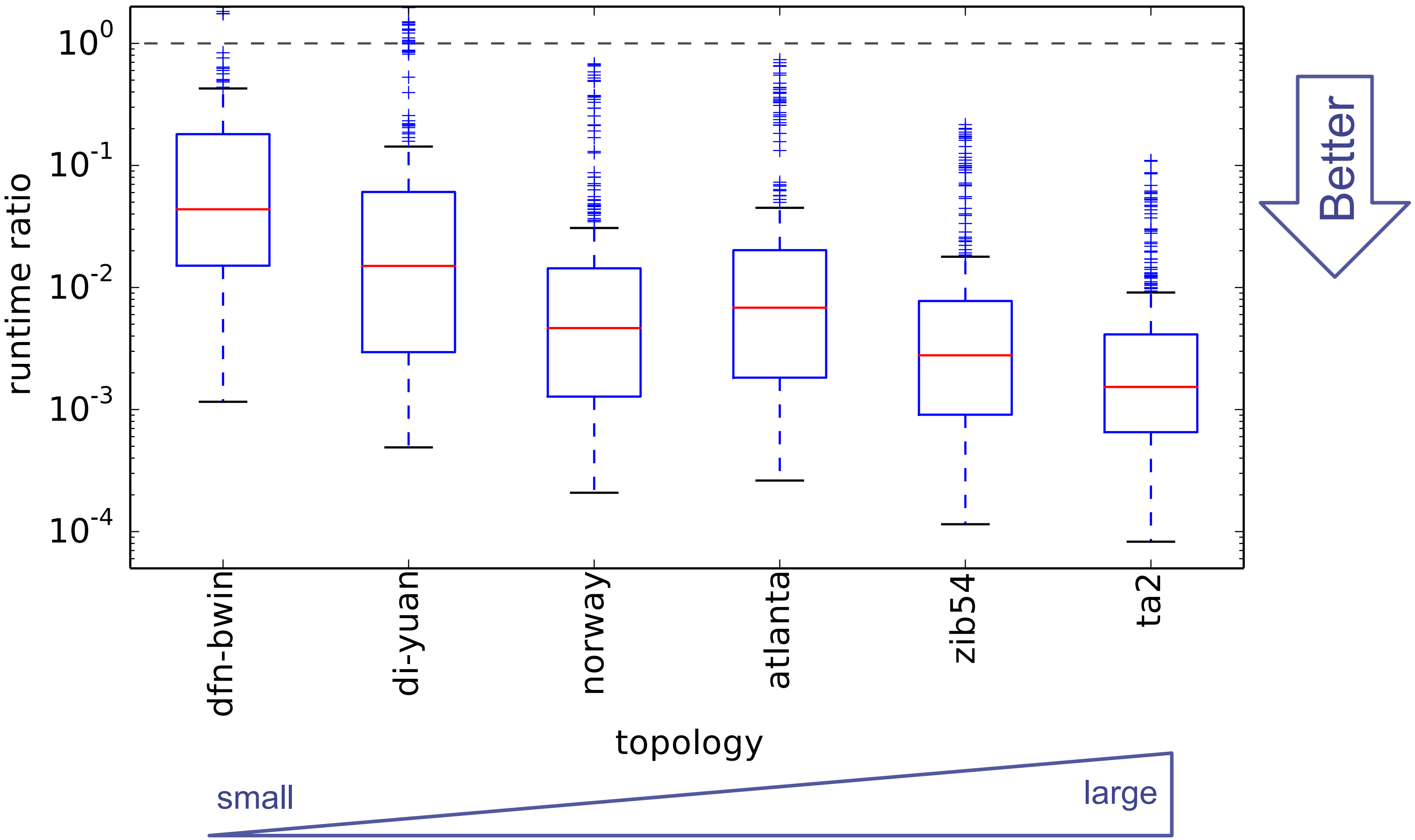
Results – Approximation Ratio

$$\text{approx. ratio} = \frac{\text{Resp.time}_{\text{Linear}}}{\text{Resp.time}_{\text{Convex}}}$$



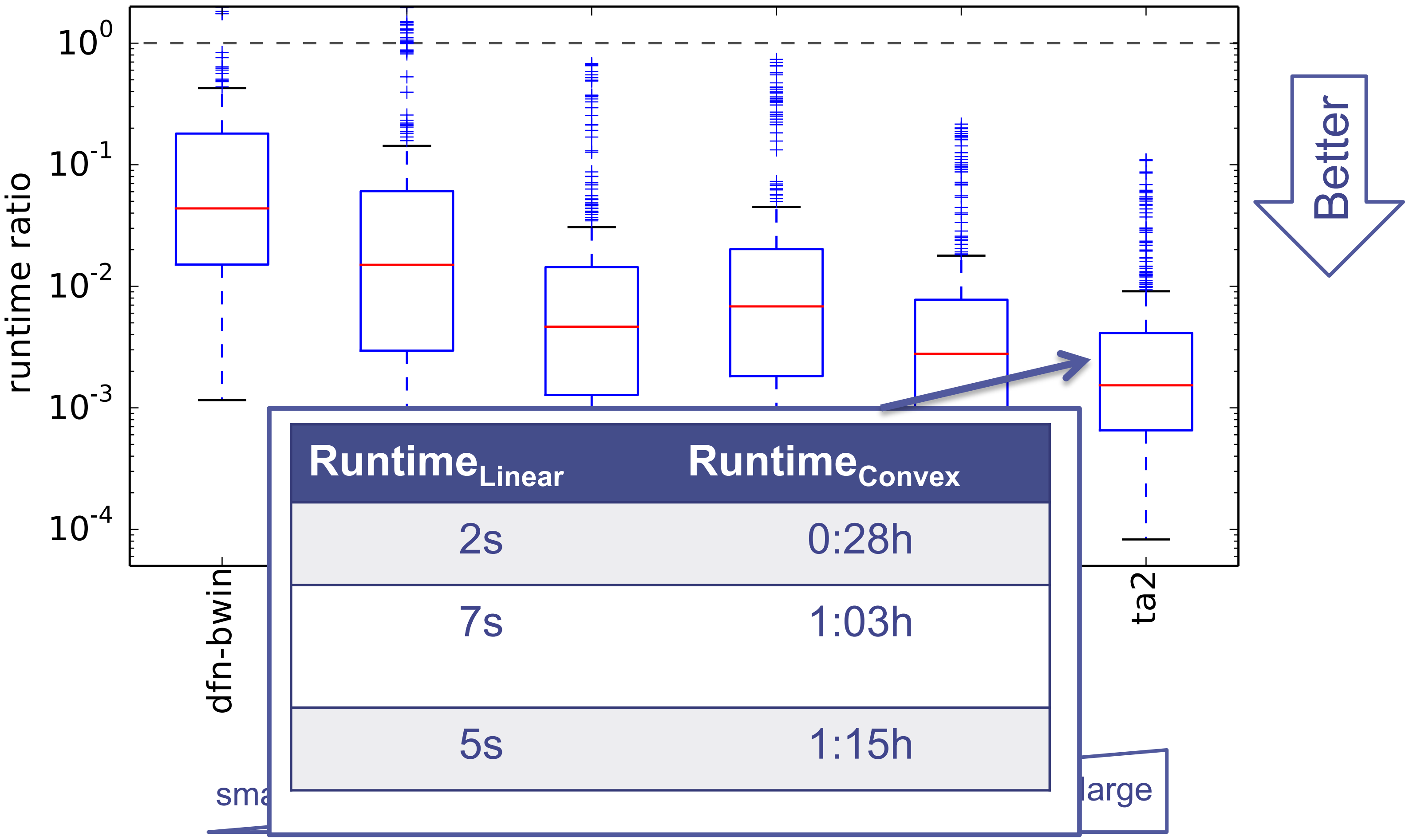
Results – Runtime Ratio

$$\text{runtime ratio} = \frac{\text{Runtime}_{\text{Linear}}}{\text{Runtime}_{\text{Convex}}}$$



Results – Runtime Ratio

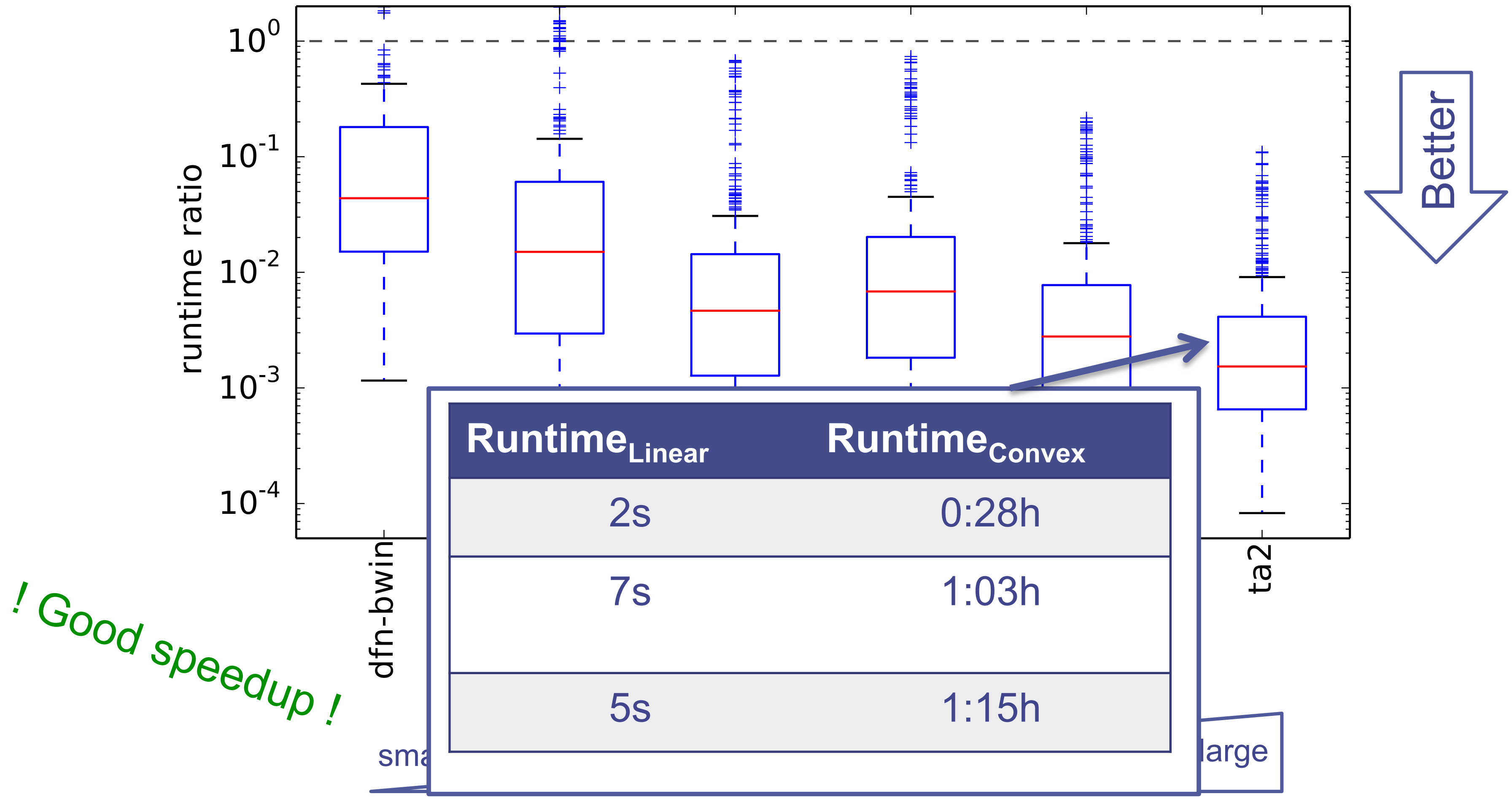
$$\text{runtime ratio} = \frac{\text{Runtime}_{\text{Linear}}}{\text{Runtime}_{\text{Convex}}}$$



$\text{Runtime}_{\text{Linear}}$	$\text{Runtime}_{\text{Convex}}$
2s	0:28h
7s	1:03h
5s	1:15h

Results – Runtime Ratio

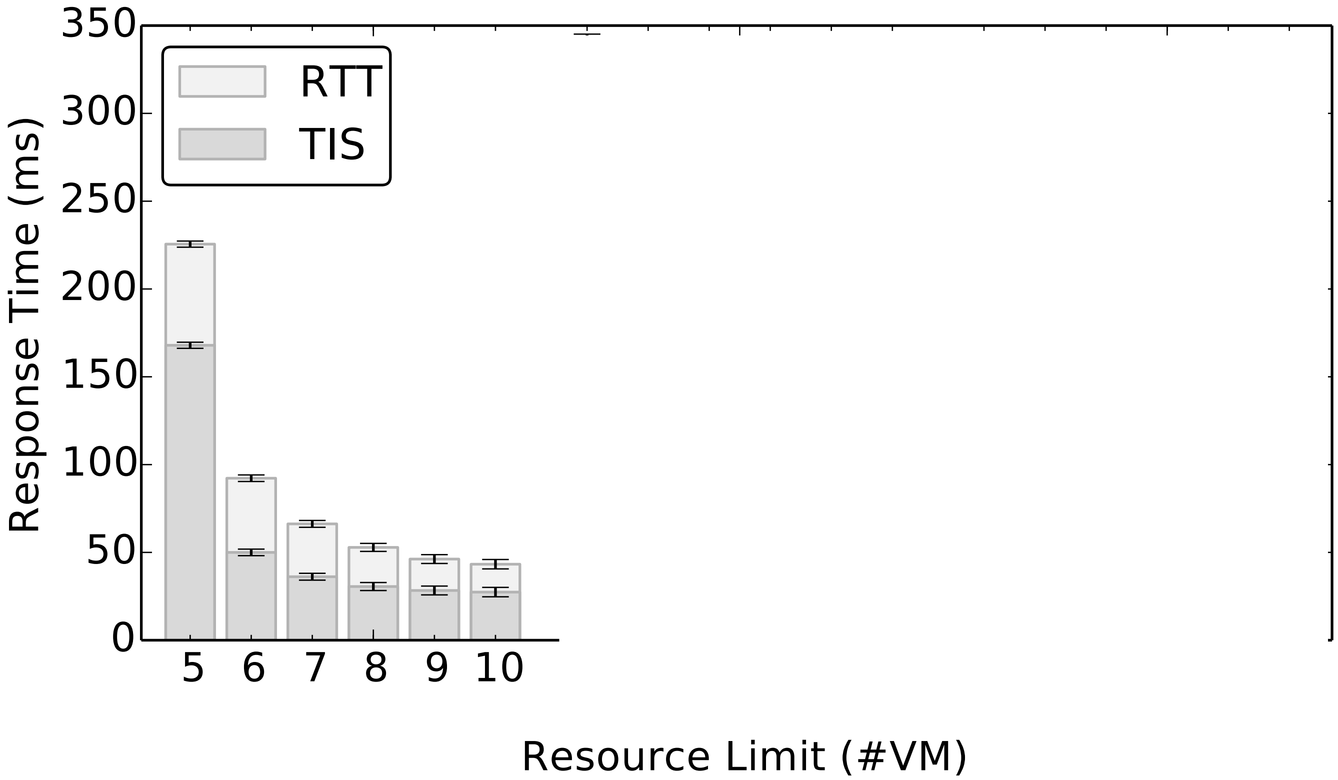
$$\text{runtime ratio} = \frac{\text{Runtime}_{\text{Linear}}}{\text{Runtime}_{\text{Convex}}}$$



Runtime$_{\text{Linear}}$	Runtime$_{\text{Convex}}$
2s	0:28h
7s	1:03h
5s	1:15h

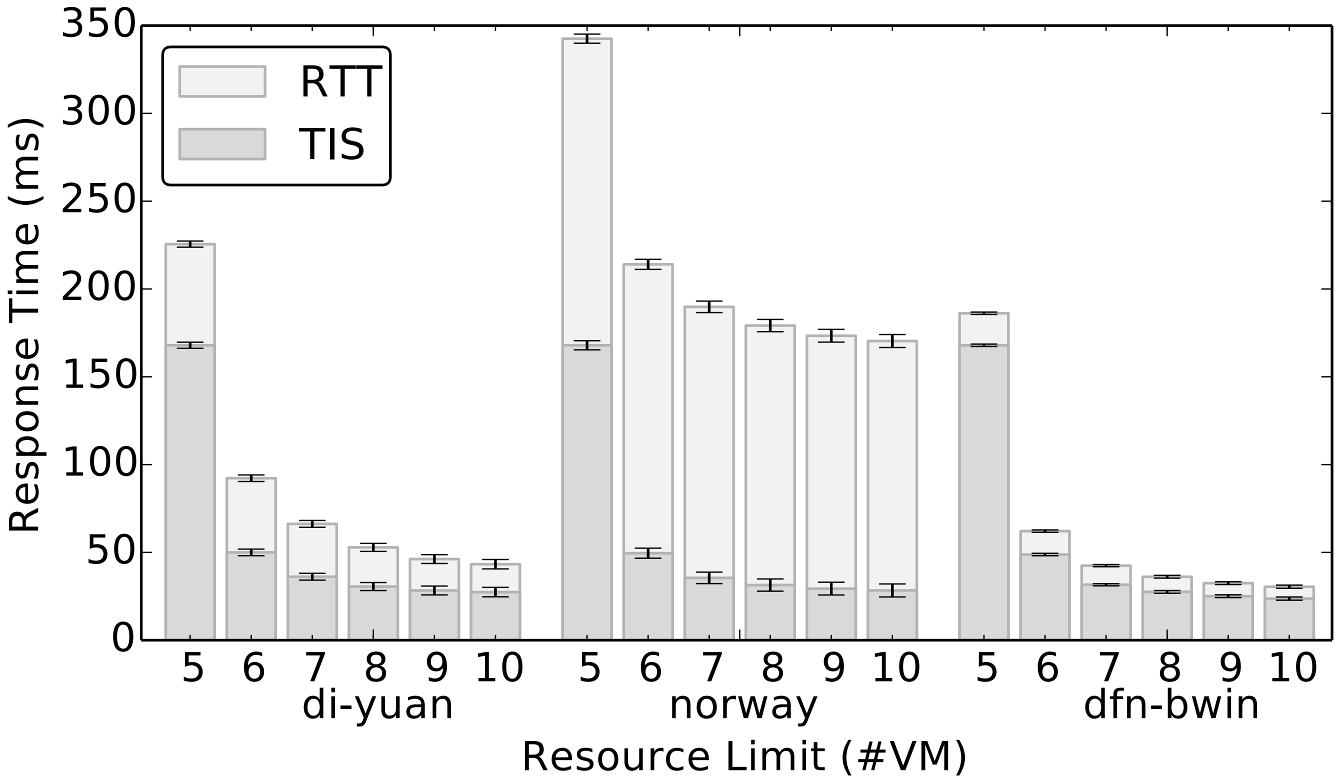
Results – Optimal Solutions

- More Resources:
  - Shorter time in queuing system
  - VMs at closer data centres



Results – Optimal Solutions

- More Resources:
  - Shorter time in queuing system
  - VMs at closer data centres



In the paper…

- Convex/Linear Problem Formulation
  - Facility Location Problem & queuing model
  - P-median facility location + convex cost function
  - P-median facility location + piecewise linear cost function
  - Piecewise Linear Function: Minimize maximal difference
  - Convexity Proof
- Evaluation
  - Pareto optimal solutions
  - Compare linear/convex problem
    - Approx. Ratio
    - Runtime

In conclusion…

… adjust your latency-sensitive service:

- Faster!
  - Adapt to demand fluctuations swiftly

Linear
Approximation

- Accurate!
  - With queuing delay – no surprises at runtime

Processing
Queue Model