Empirical Virtual Machine Models for Performance Guarantees

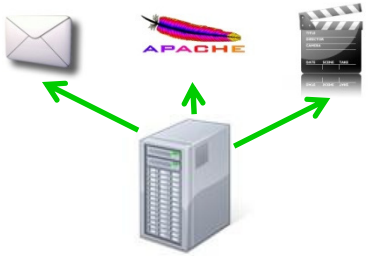
Andrew Turner, Akkarit Sangpetch, Hyong S. Kim
Carnegie Mellon University

LISA 2010
11th November 2010

Electrical & Computer ENGINEERING

Introduction

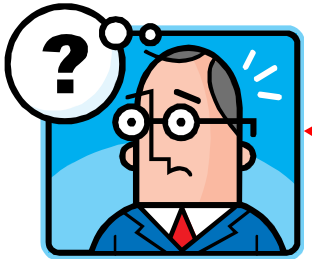
Good application performance



Hosts run multiple VMs

Performance bottlenecks

Resource allocation levels



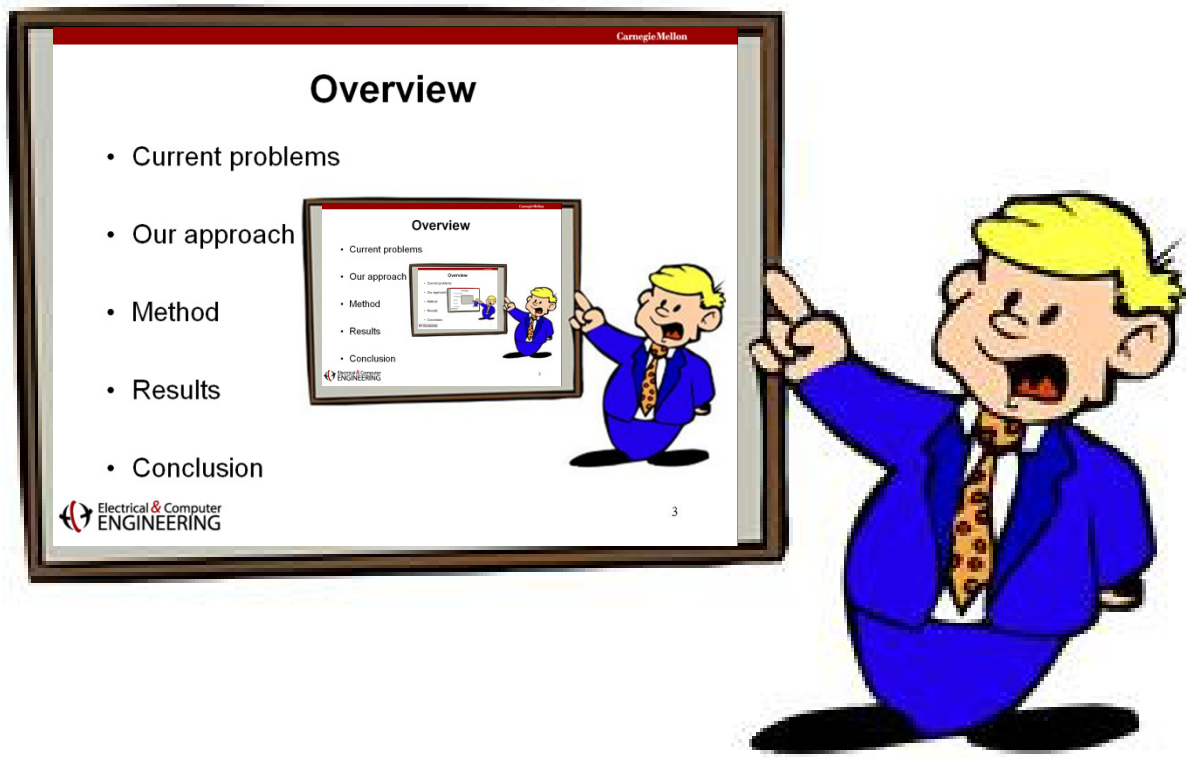
You Now

New You →  = 

Automatic resource allocations levels

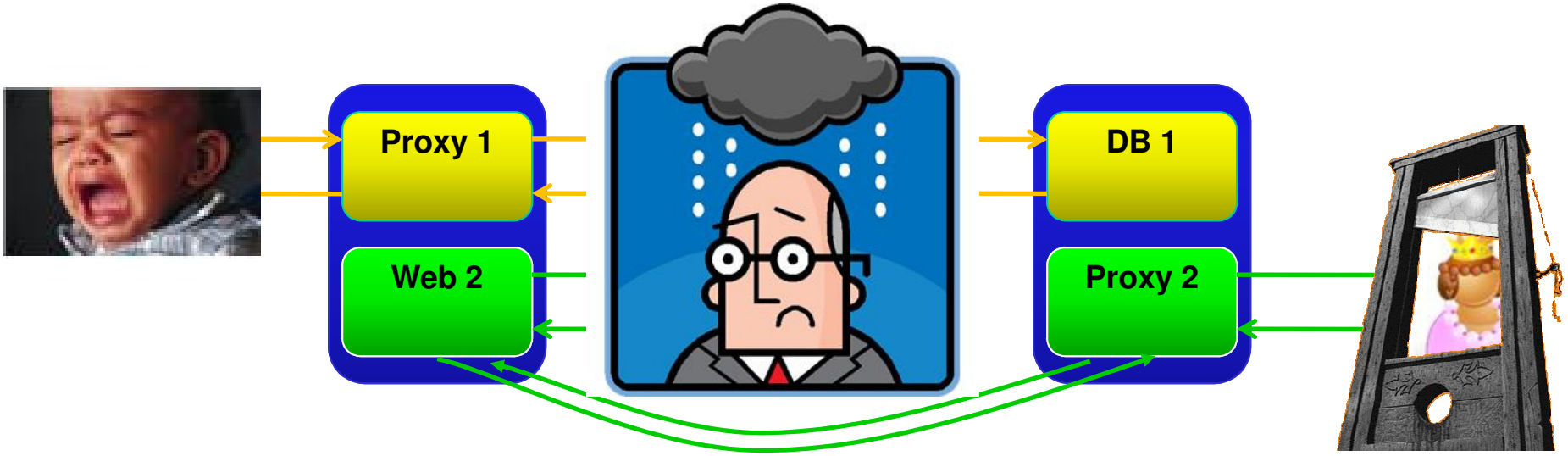
Overview

- Current problems
- Our approach
- Method
- Results
- Conclusion



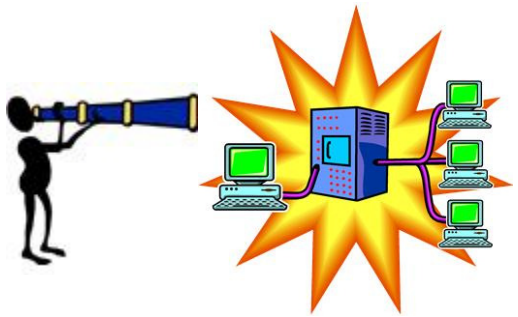
Electrical & Computer ENGINEERING

Current Problems



- Multiple application tiers on different hosts
- Resources needs
- F(resource allocation) = performance?
- Needs change throughout the day
- Over-provisioning wastes energy and resources
- Unhappy users

Our approach

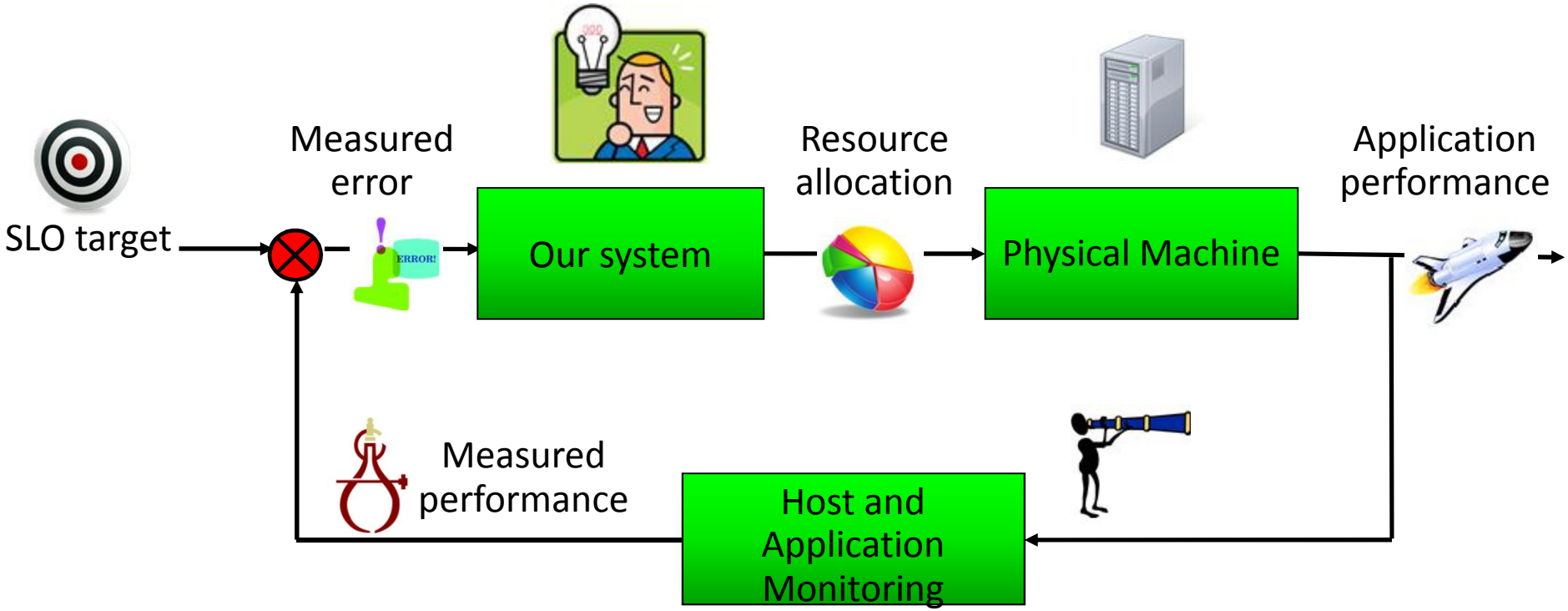


- Observe performance

- Create online model

- Calculate required resources

Our approach



Control loop constantly checks performance and recalibrates resource allocation levels

Benefits of our system

- Automatically identifies performance bottlenecks
- Automatically sets resource allocation levels
- Provides more performance per resource allocated
- Reduces energy and hardware usage
- Allows SLOs to be met

Assumptions

- We can monitor application performance
- We can control resource access or scheduling
- Application performance is convex

Data used


T – SLO target


R – Real performance


W – Workload level


M – Performance Model


E – Probability T achieved


C – Contention level


A – Resource allocation

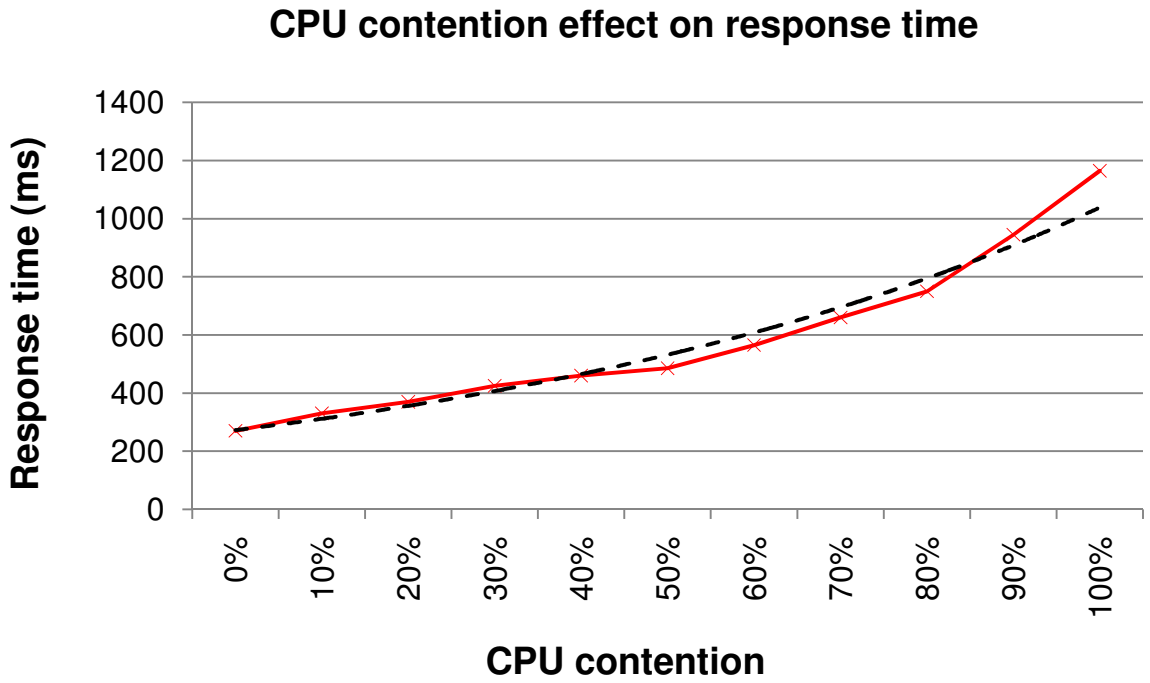
$$P(R \geq T) \geq E$$

$$P(R \geq T) = P(M(W, C, A) \geq T)$$

$$P(R \geq T) = \iint_{0..100\%} P(W = w)P(C = c)P(M(w, c, A) \geq T)dwdc$$

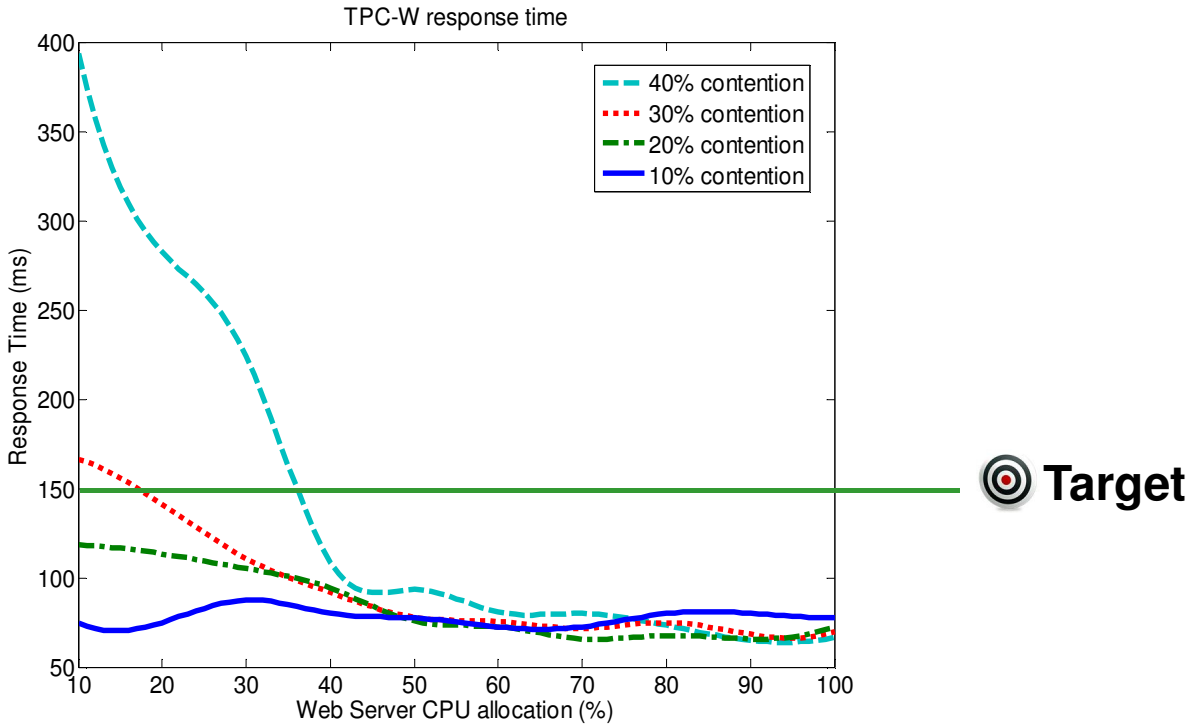
Creating the model

- Created online
- Use previously observed data
- Curve fit to fill unobserved areas



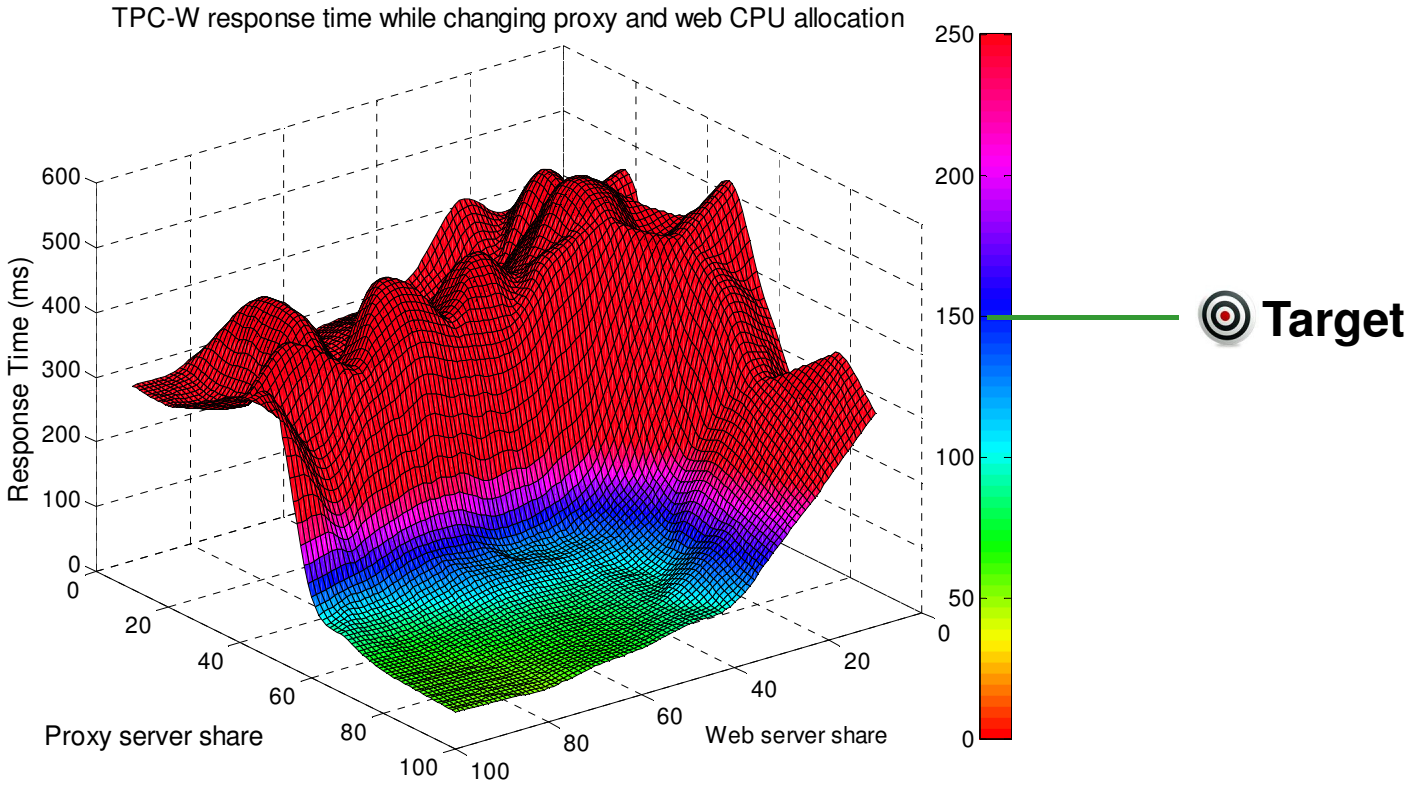
Deciding resource assignment

- Hyperplane at target performance
- Choose allocation that crosses plane



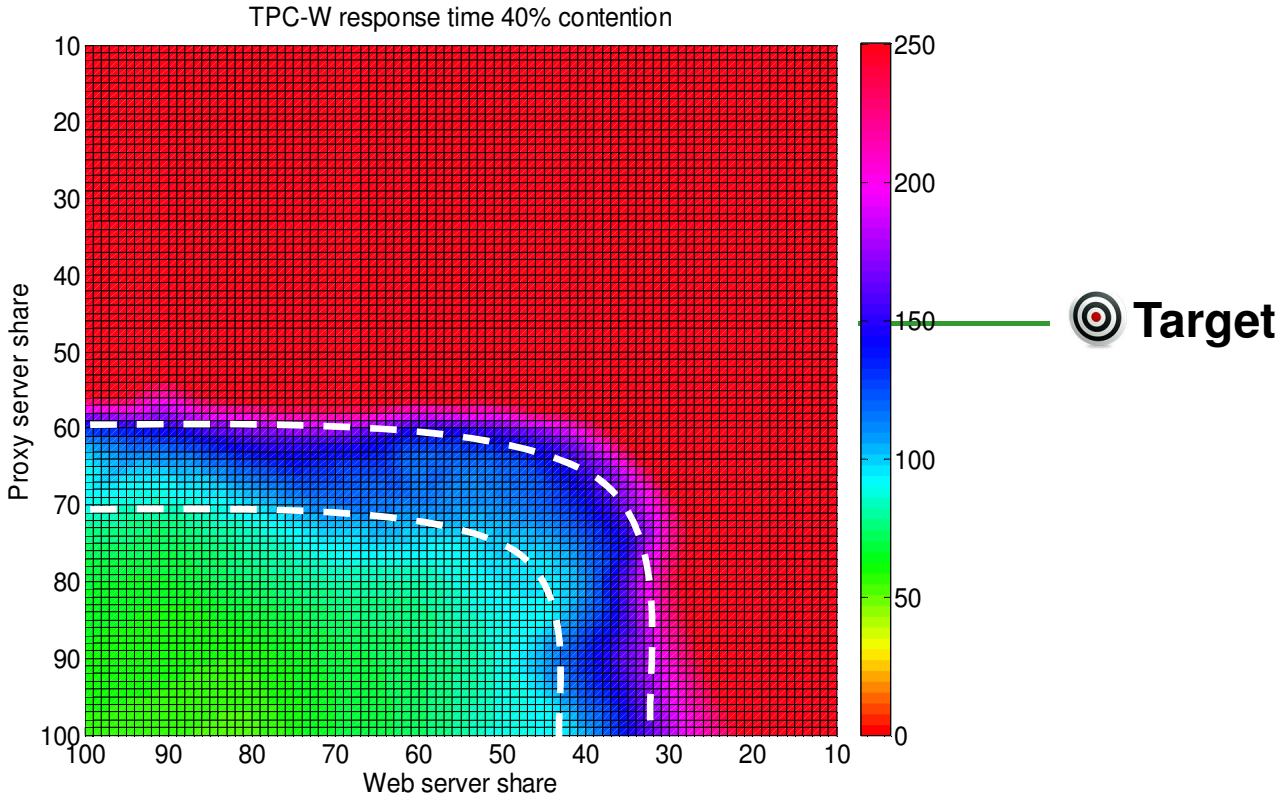
Deciding resource assignment

- Resource allocation is vector of allocations
- E.g. (proxy = 65%, web = 55%) or (proxy = 80%, web = 35%)



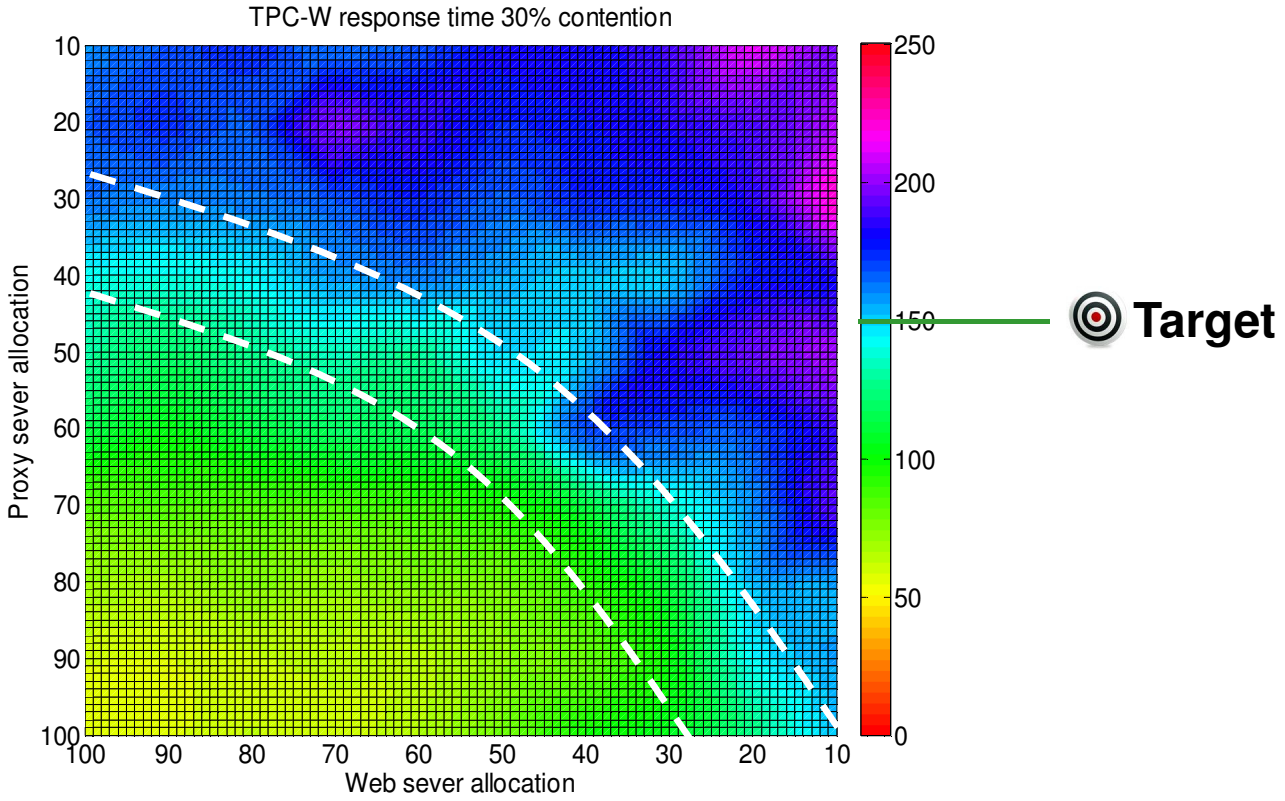
Deciding resource assignment

- Resource allocation is vector of allocations
- E.g. (proxy = 65%, web = 55%) or (proxy = 80%, web = 35%)



Deciding resource assignment

- Resource allocation is vector of allocations



Deciding resource assignment

A – the potential resource allocations

X – chosen resource allocations

Q – Priority level of application

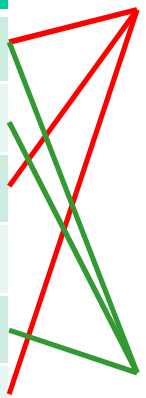
	Hosts			Priority	Start X	End X
App 1 solutions	30	0	60	0.8	?	0
	50	0	50	0.8	?	0
	60	0	30	0.8	?	1
App 2 solutions	70	20	50	1	?	0
	20	20	90	1	?	0
	40	80	50	1	?	1

- Minimize: $X^T A\ 1^T Q$
- Subject to: $X^T A <= 100\%,\ 1^T X >= 1,\ X >= 0$

Reducing model dimensions

- Which resources are important to model?
- Use regression to find impact of each resource

Time	CPU Contention	Disk Contention	Performance
1	10%	10%	130ms
2	40%	12%	180ms
3	14%	90%	135ms
4	12%	50%	132ms
5	30%	75%	160ms
6	10%	40%	130ms

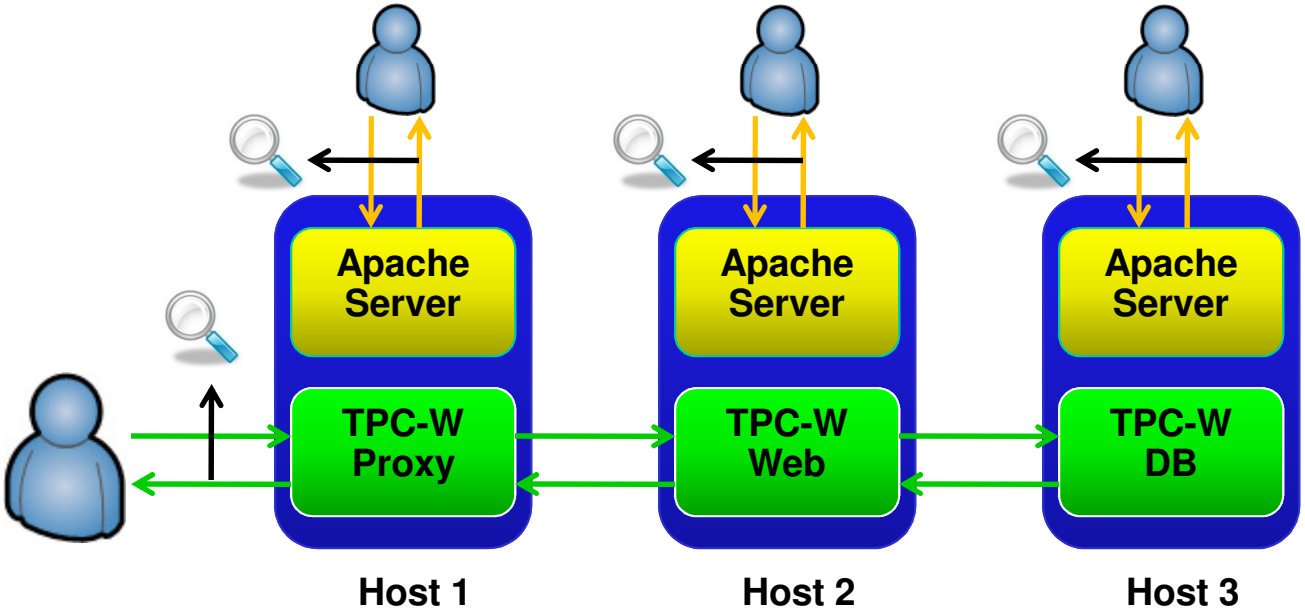


Disk has no affect

CPU does have affect

Experimental Evaluation

- We test TPC-W and a dynamic web page
- Measure response time



Experimental Evaluation

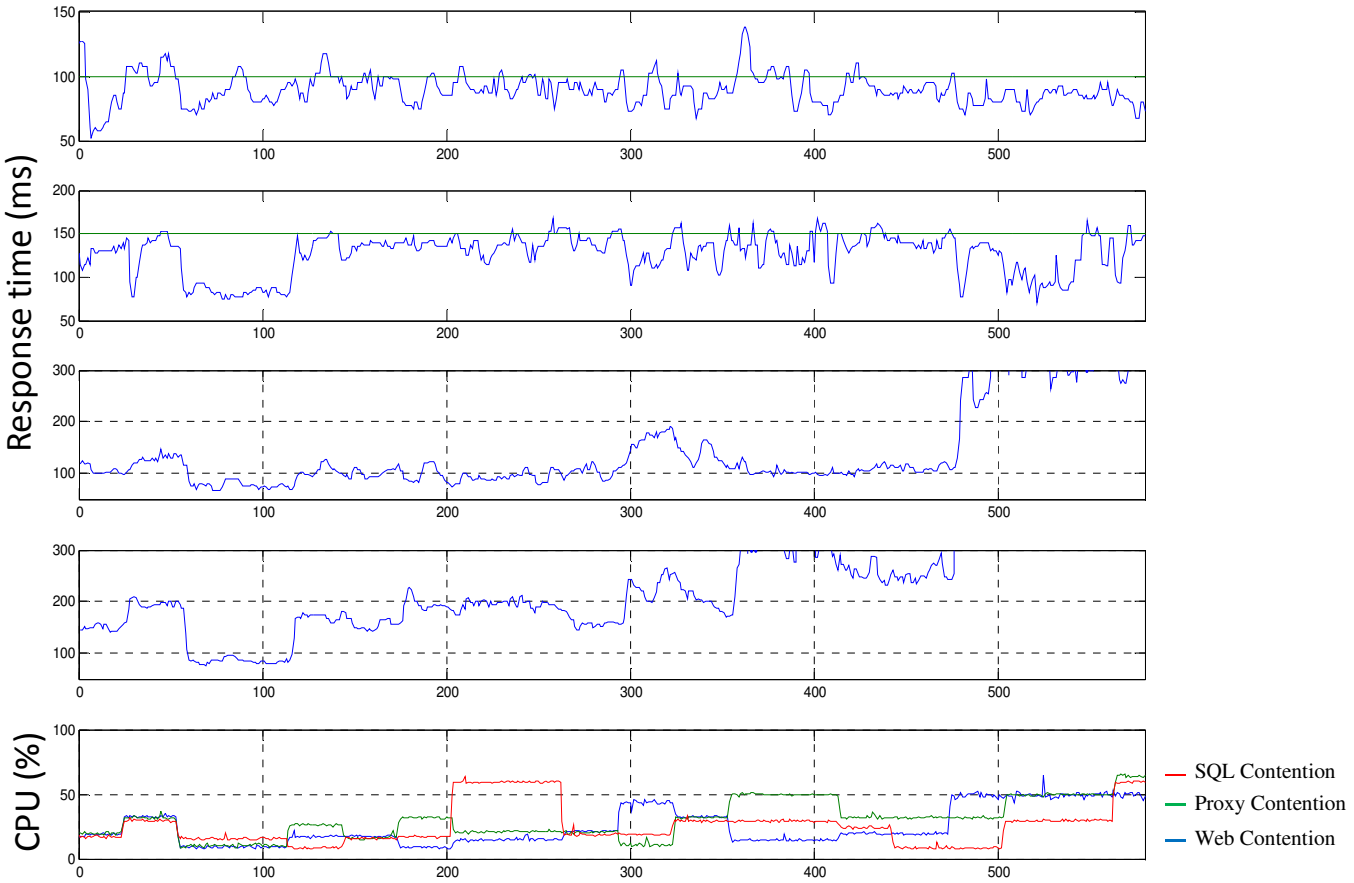
SLO 100ms

SLO 150ms

50% resource allocation

10% resource allocation

CPU contention levels



- System keeps response time close to target

Experimental Evaluation

- Dynamic resource assignment helps meet SLOs
- Use less resources that static allocation

Test	RT average	Resource allocation average	Apache VM average
SLO = 100ms	89ms	48%	125ms
SLO = 150ms	127ms	35%	107ms
50% resource allocation	150ms	50%	120ms
10% resource allocation	355ms	10%	83ms

Experimental Evaluation

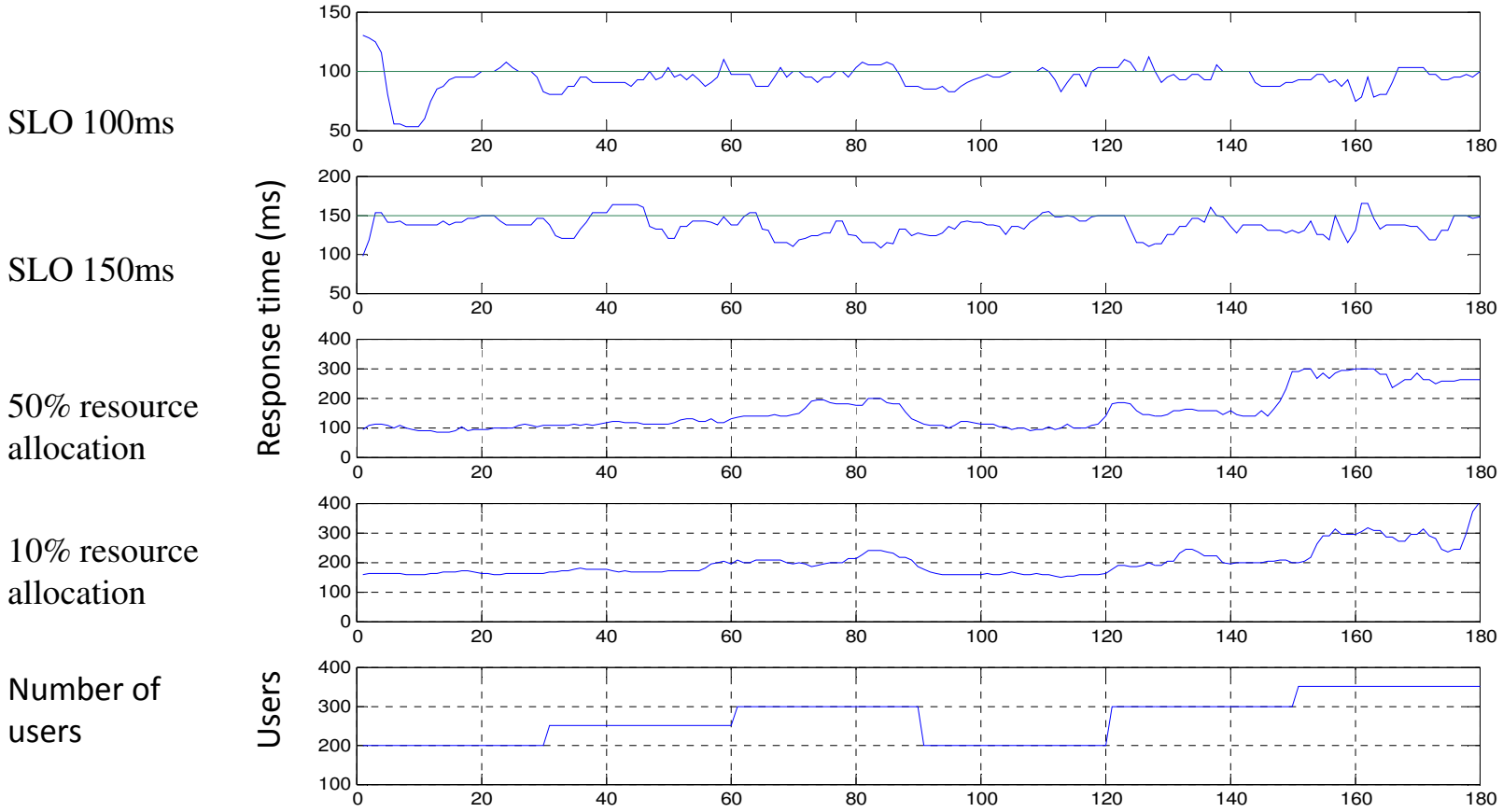
SLO 100ms

SLO 150ms

CPU contention levels

- No increase in resource allocation as DB is not bottleneck

Experimental Evaluation



- Meets target time despite changes in workload

Conclusion

- We automatically calculate required resources
- Works on generic multi-tiered applications
- Helps to meet SLOs
- Better performance per resource assigned
- Simplifies resource management