Elastic VM for Rapid and Optimum Virtualized Resources' Allocation

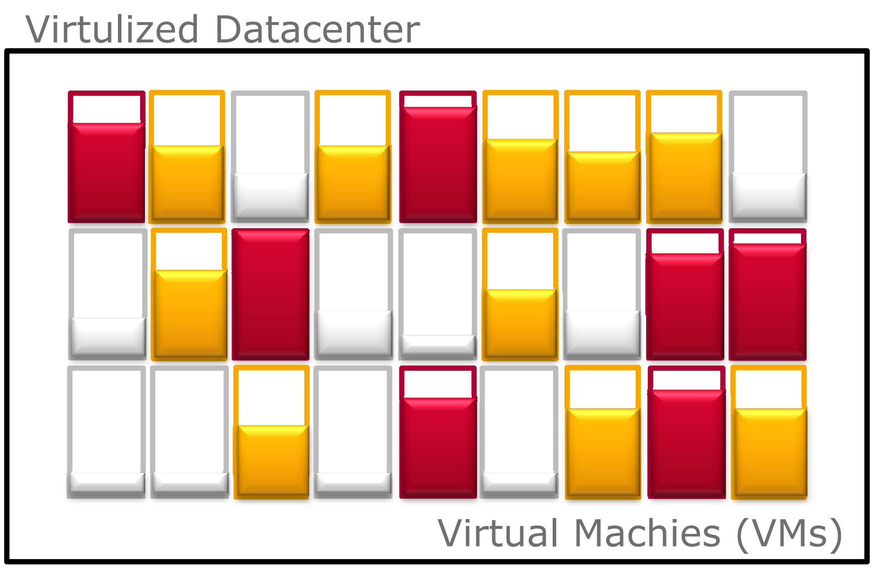
Wesam Dawoud
PhD. Student
Hasso Plattner Institute
Potsdam, Germany

5th International DMTF Academic Alliance Workshop on Systems and Virtualization Management: Standards and the Cloud
24 October, 2011 Paris, France

Motivation

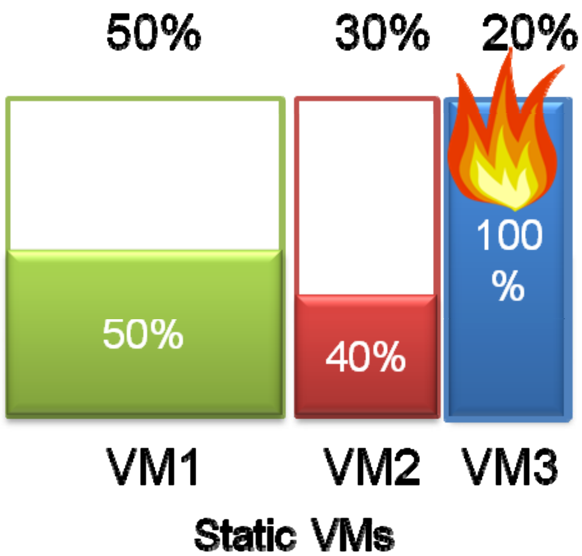
2

- 11.8 million servers in the USA in 2007. Most of those machines run at 15% capacity or less *
- Virtualized Datacenters rise server utilization rates to as high as 80 percent **
- Idle server consumes 50% of the power consumed by a highly utilized server



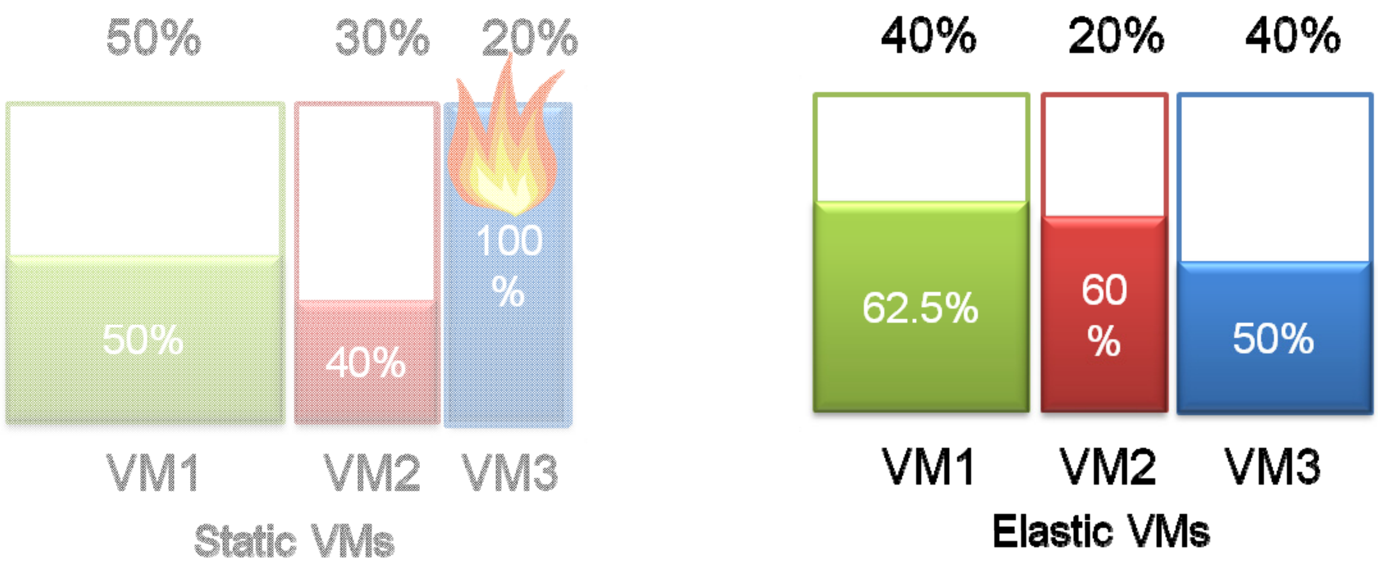
* http://blogs.computerworld.com/data_center_utilization_15_of_11_8_million_is_a_big_number
** http://www-03.ibm.com/systems/virtualization/news/view/vdc.html

Elastic VM



- Elastic VM is a VM benefits from virtualization technology features to enable on-the-fly resources scaling without interrupting the service or rebooting the system
- The hosting hypervisor is extended with interfaces that enable modifying VMs resources

Elastic VM

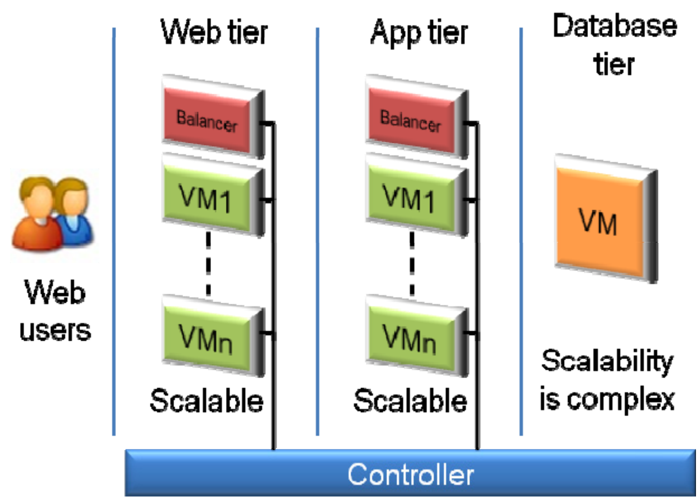


- Elastic VM is a VM runs a modified kernel supports on-the-fly resources scaling feature without interrupting the service or rebooting the system
- The hosting hypervisor is extended with interfaces that enable modifying VMs resources

Current Cloud Computing Elasticity

Hasso Plattner Institut

5 Multi-instances scaling *



1. Coarse-grained scaling
2. It is not the best scalability solution for all applications (e.g., Databases, Load balancers, and Applications with expensive licenses)
2. Scaling-down can interrupt sessions-based applications
3. Scale-out overhead causes SLO violation

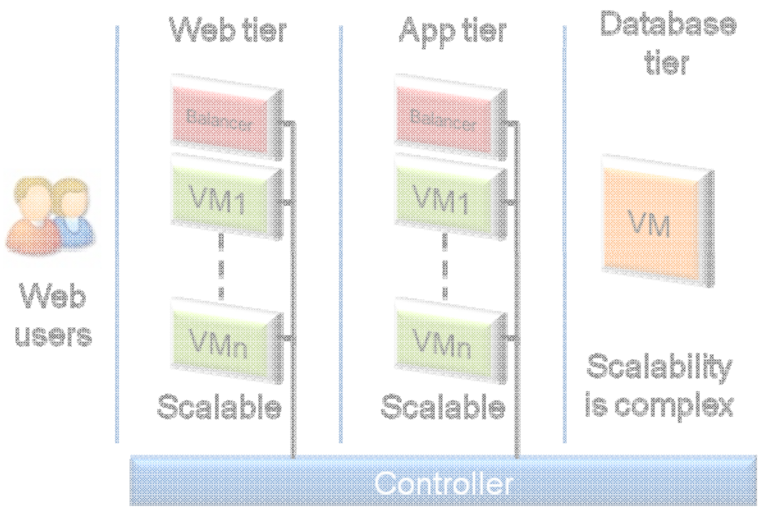
* http://media.amazonwebservices.com/AWS_Web_Hosting_Best_Practices.pdf

Objectives

- Optimize resources allocation (i.e., fine grained dynamic scale-up)
- Reduce the power consumption (i.e., run less physical hosts)
- Reduce current scale-out overhead
- Maintain Service Level Objectives (SLOs)

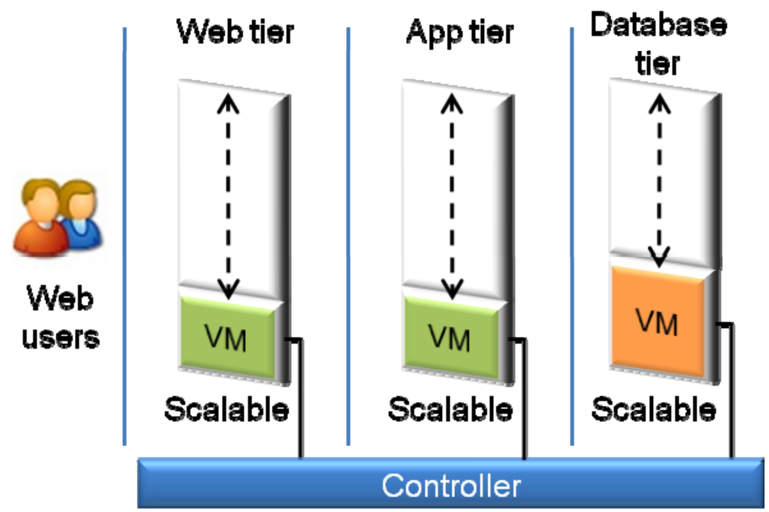
Elastic VM

Multi-instances scaling



1. Coarse-grained scaling
2. It is not the best scalability solution for all applications
3. Scaling-down can interrupt sessions-based web connections
4. Scale-out overhead causes SLO violation

Elastic VM scaling

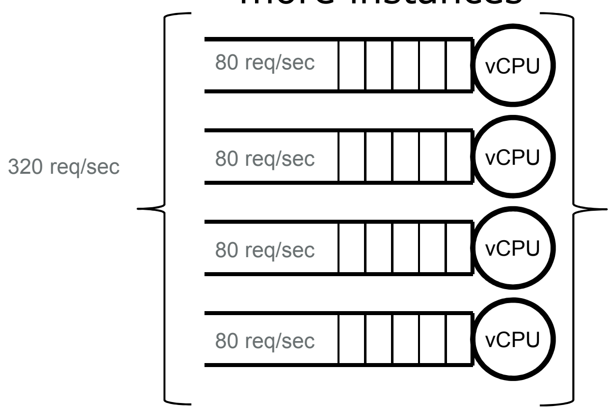


1. Fine-grained scaling
2. Applicable to any tier
3. Supports sessions-based web connections
4. Reduces the scaling-up overhead and mitigates SLO violation

Mutli-instances v.s. Elastic VM scaling

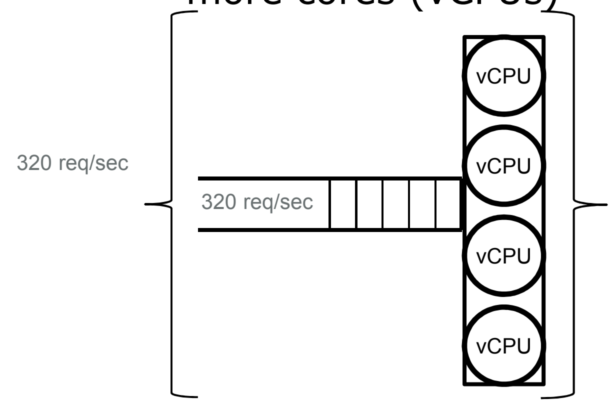
Performance:*

Mutli-instances scale by adding more instances



Each instance modeld as M/M/1

Elastic VM scales by adding more cores (vCPUs)



One instance with many cores modeld as M/M/c

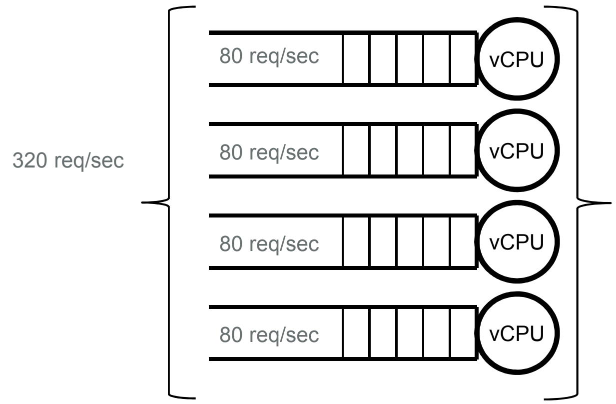
c	1	2	3	4	5	6	7	8
λ	80	160	240	320	400	480	560	640
Average response time (seconds)	0.05	0.027	0.02	0.017	0.015	0.014	0.013	0.013
Average wait time (seconds)	0.04	0.018	0.01	0.007	0.006	0.004	0.003	0.003

* Wesam Dawoud, Ibrahim Takouna and Christoph Meinel , “Elastic Virtual Machine for Fine-grained Cloud Resource Provisioning”, ObCom 2011, Vellore, TN, India, Springer Berlin Heidelberg (2011)

Mutli-instances v.s. Elastic VM scaling

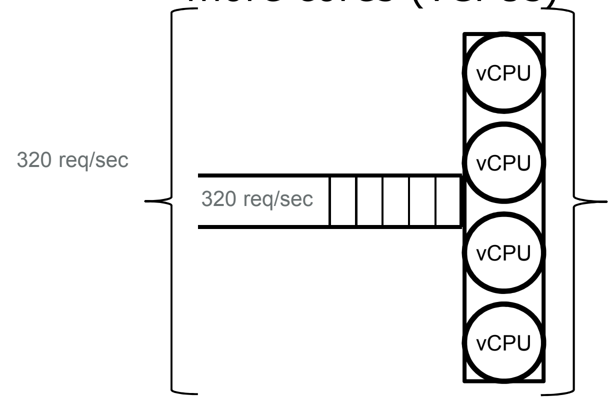
Performance:

Mutli-instances scale by adding more instances



Each instance modeld as M/M/1

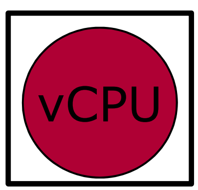
Elastic VM scales by adding more cores (vCPUs)



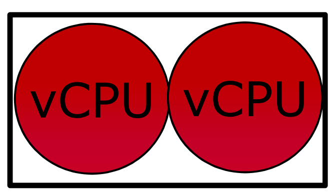
One instance with many cores modeld as M/M/c

c	1	2	3	4	5	6	7	8
λ	80	160	240	320	400	480	560	640
Average response time (seconds)	0.05	0.027	0.02	0.017	0.015	0.014	0.013	0.013
Average wait time (seconds)	0.04	0.018	0.01	0.007	0.006	0.004	0.003	0.003

Power:*



26 watts



17 watts

- Same workload
- Same throughput
- 50% less power consumption

* I. Takouna, W. Dawoud, and C. Meinel. "Accurate Mutlicore Processor Power Models for Power-Aware Resource Management", In Proceedings of the International Conference on Cloud and Green Computing (CGC 2011), December 2011

Evaluation

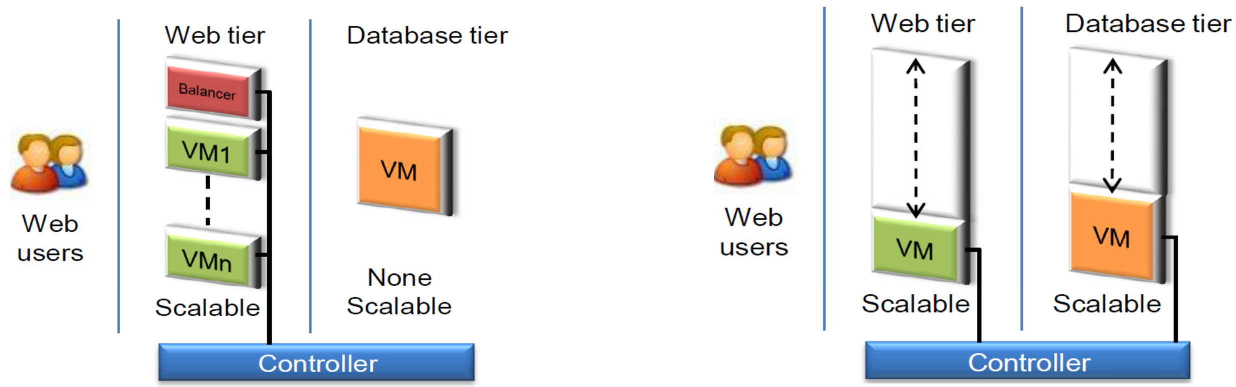


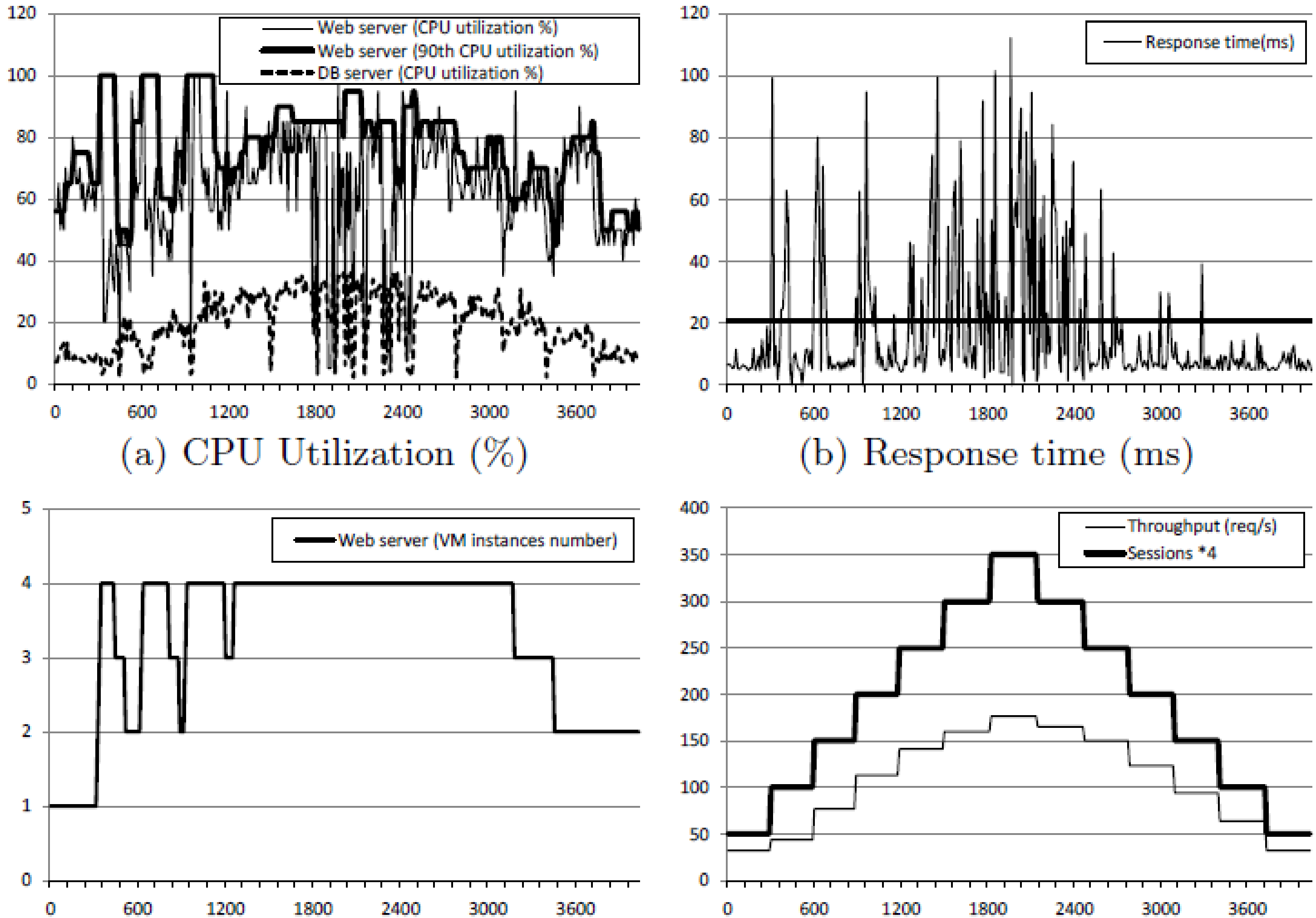
Table 1: Most significant parameters that control Amazon Scaling Model

Parameter description	Value
Minimum number of running instances	1
Maximum number of running instances	4
Monitored metric	CPU Utilization
Monitored metrics' measurement period	5 seconds
Lower threshold of measured metric	80
Upper threshold of measured metric	90
Breach duration	60 seconds
Lower breach increment	-1
Upper breach increment	1

The workload is generated by RuBBoS *

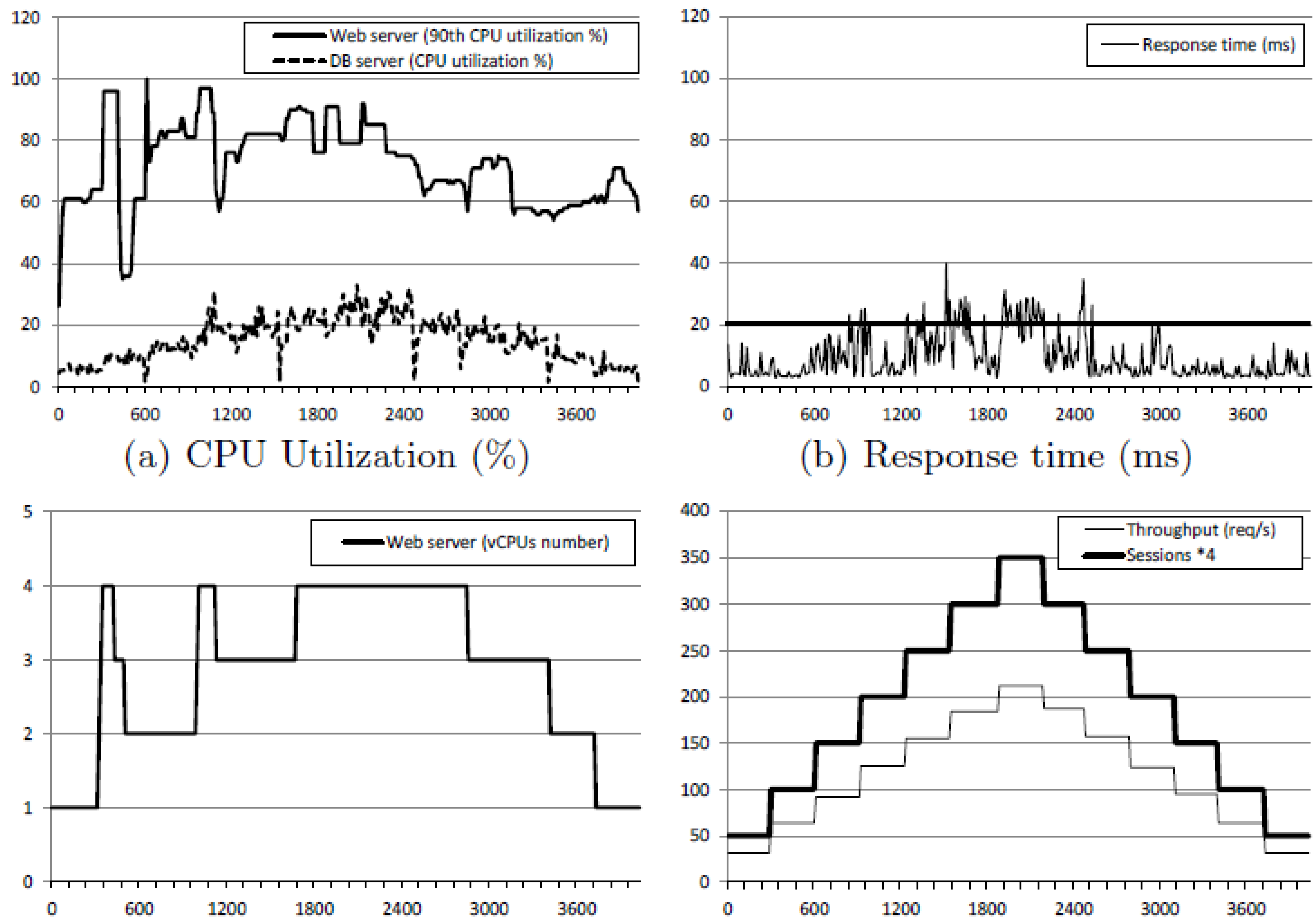
* Amza, C., Cecchet, E., Ch, A., Cox, A.L., Elnikety, S., Gil, R., Marguerite, J., Rajamani, K., Zwaenepoel, W.: Bottleneck Characterization of Dynamic Web Site Benchmarks (2002)

Evaluation: Multi-instances architecture implemented into web-tier



(a) CPU Utilization (%)
(b) Response time (ms)
(c) Number of VM instances
(d) Workload & Throughput

Evaluation: Elastic VM implemented into web-tier



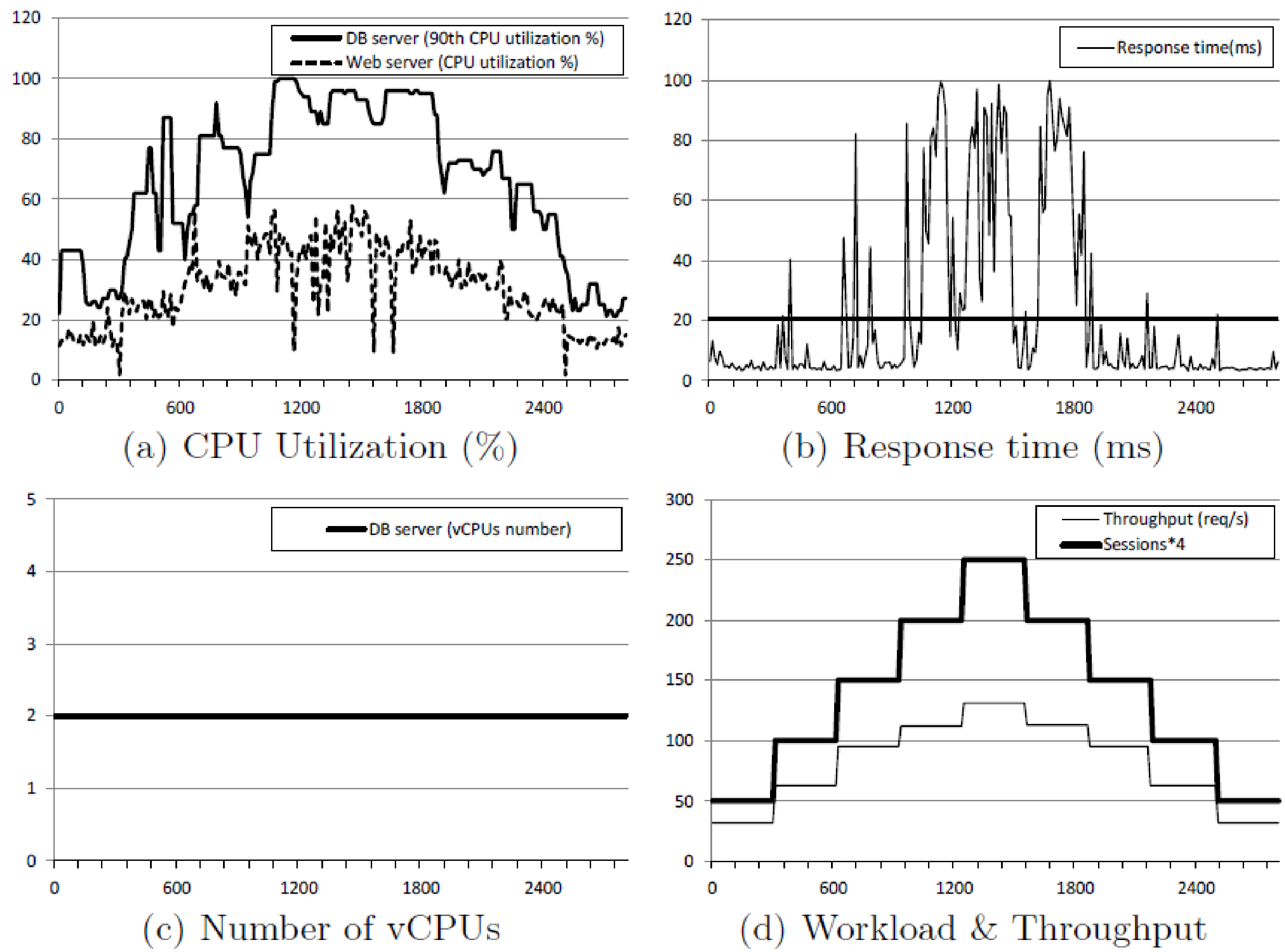
(a) CPU Utilization (%)
(b) Response time (ms)
(c) Number of vCPUs
(d) Workload & Throughput

Evaluation : Mutli-instances throughput degradation

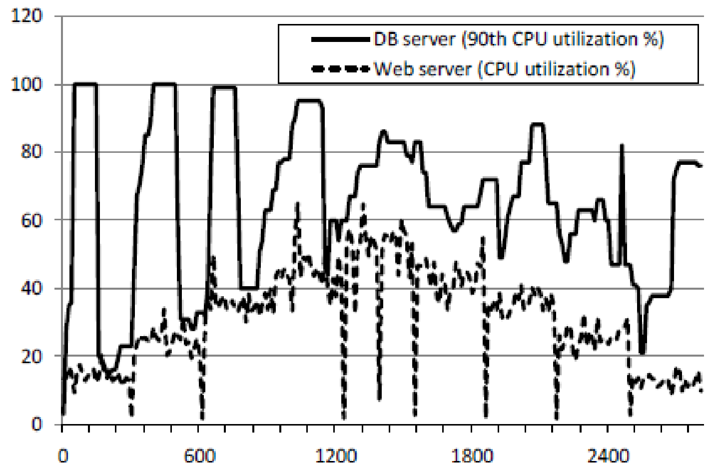
Table 2: Throughput degradation in Multi-instances architecture

Sessions Number	Elastic-VM throughput (req/sec)	Multi-instances throughput (req/sec)	Multi-instances throughput degradation (%)
200	32	32	0
400	64	44	31
600	92	77	16
800	125	113	10
1000	155	141	9
1200	184	160	13
1400	212	176	17
1200	187	165	12
1000	157	150	4
800	124	123	1
600	95	94	1
400	64	64	0
200	32	32	0

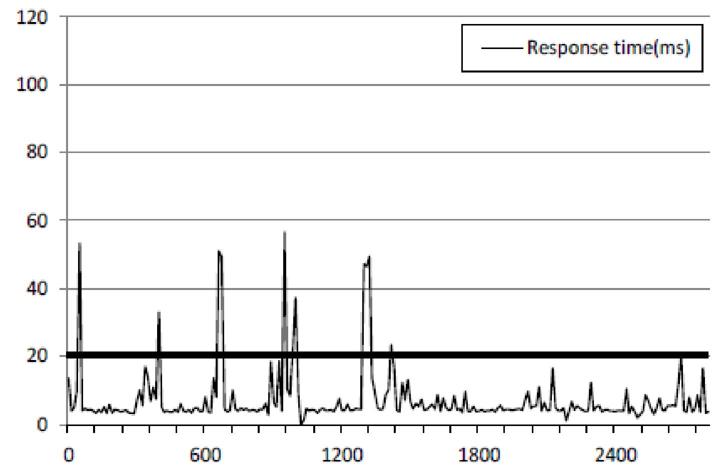
Evaluation: Multi-instances architecture implemented into database-tier



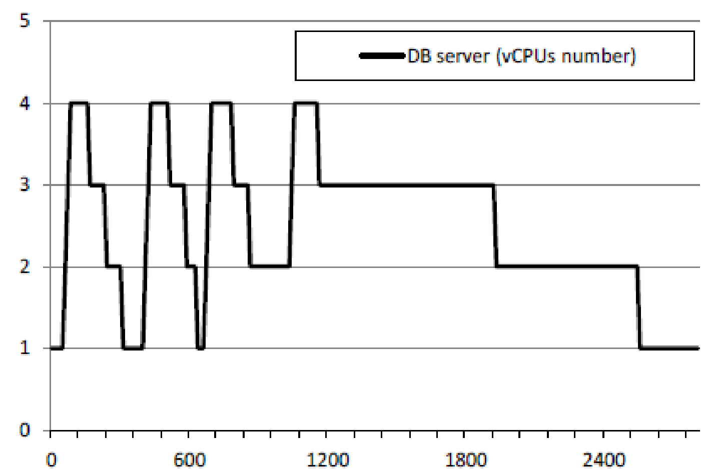
Evaluation: Elastic VM implemented into database-tier



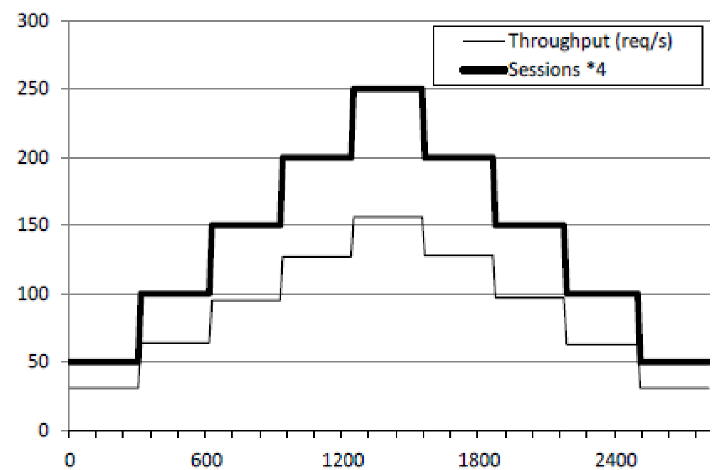
(a) CPU Utilization (%)



(b) Response time (ms)



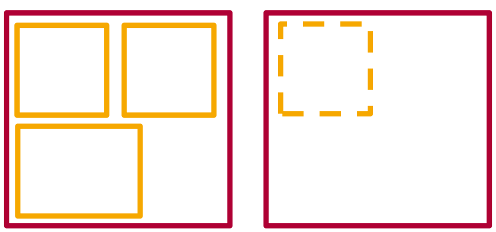
(c) Number of vCPUs



(d) Workload & Throughput

Challenges

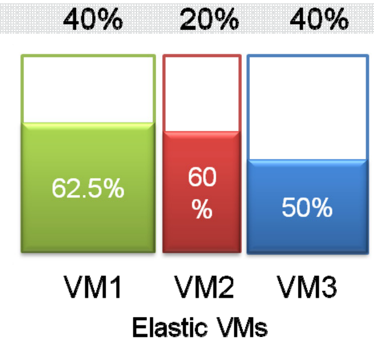
- Not all applications are aware of on-the-fly resources scaling
  - Solution:
    Most applications have performance metrics *
- Elastic VM scalability is limited to one physical machine
  - Solutions:
    1-VMs migration
    2-Hybrid architecture (i.e. Multi-instances integrated with Elastic VM architecture)
- Variant sizes of instances increase the bin-packing algorithm complexity



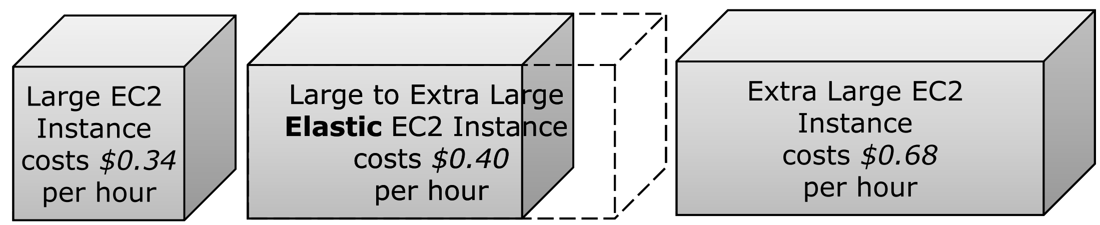
* W. Dawoud, I. Takouna, and C. Meinel, Elastic VM for Cloud Resources Provisioning Optimization, Eds. Springer Berlin Heidelberg, 2011

Who can use Elastic VM?

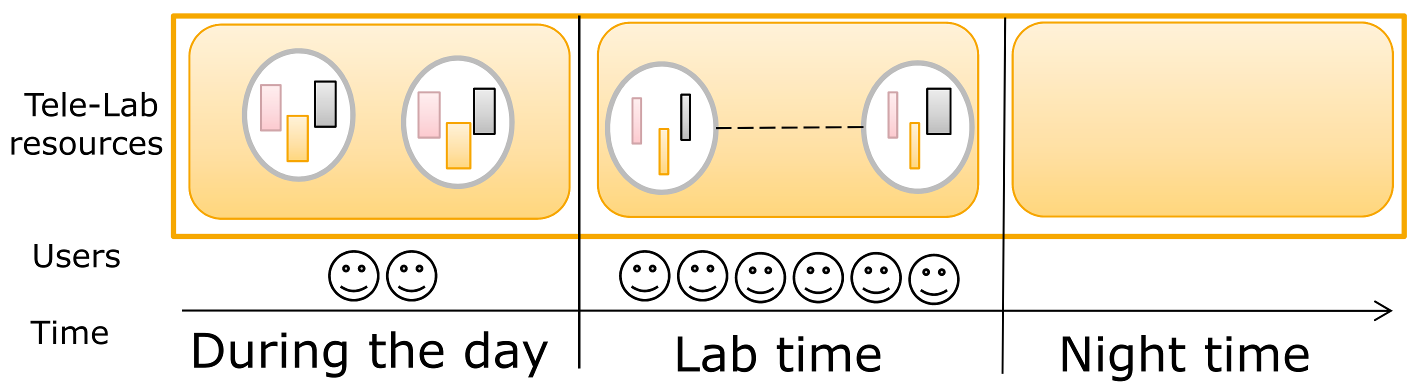
- Web hosting service providers



- IaaS Providers



- Tele-Lab (http://www.tele-lab.org/)



Conclusion & Future work

18

Conclusion:

- Experiment results confirm the theoretical analysis and show that proposed Elastic VM:
  - Mitigates SLOs violation
  - Maintains a higher throughput

- Elastic VM supports scaling applications, such as databases and expensive license software, with lower cost and complexity

Conclusion & Future work

Future work:

- Go over the current challenges:
  - One physical host
  - Bin-packing of variant size VMs
- Integrate Elastic VM with other managment techniques (e.g., VMs migration and workload redirection)
- Implemet Elastic VM into Tele-Lab (http://www.tele-lab.org/)

References

[1] Iqbal, W., Dailey, M.N., Carrera, D.: SLA-Driven Dynamic Resource Management for Multi-tier Web Applications in a Cloud. In: 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. pp. 832-837. CCGRID '10,IEEE, Washington (2010)

[2] Bhuvan Urgaonkar, G.P.: An analytical model for multi-tier internet services and its applications. In: In Proc. of the ACM SIGMETRICS2005. pp. 291-302 (2005)

[3] Dawoud, W., Takouna, I., Meinel, C.: Elastic VM for Cloud Resources Provisioning Optimization, Communications in Computer and Information Science, vol. 190. Springer Berlin Heidelberg (2011)

[4] Takouna, I., Dawoud W., and Meinel C. "Efficient Virtual Machine Scheduling-policy for Virtualized Heterogeneous Multicore Systems", In Proceedings of the 2011 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), July 2011.

[5] Hellerstein, J.L., Diao, Y., Parekh, S., Tilbury, D.M.: Feedback Control of Computing Systems. John Wiley & Sons (2004)

[6] Dubey, A., Mehrotra, R., Abdelwahed, S., Tantawi, A.: Performance modeling of distributed multi-tier enterprise systems. ACM SIGMETRICS Performance Evaluation Review 37(2), 9 (Oct 2009)

[7] Padala, P., Hou, K.Y., Shin, K.G., Zhu, X., Uysal, M., Wang, Z., Singhal, S., Merchant, A.: Automated control of multiple virtualized resources. European Conference on Computer Systems pp. 13-26 (2009)

[8] Heo, J., Zhu, X., Padala, P. Wang, Z.: Memory Overbooking and Dynamic Control of Xen Virtual Machines in Consolidated Environments. In: Proceedings of IFIP-IEEE Symposium on Integrated Management IM09 miniconference. pp. 630-637. IEEE (2009)

[9] Dawoud, W., Takouna, I., Meinel, C., "Elastic Virtual Machine for Fine-grained Cloud Resource Provisioning", ObCom 2011, Vellore, TN, India, Springer Berlin Heidelberg (2011)

Hasso Plattner Institut

Thanks!

Contact: wesam.dawoud@hpi.uni-potsdam.de