Auto-Scaling Cloud Applications Using Machine Learning

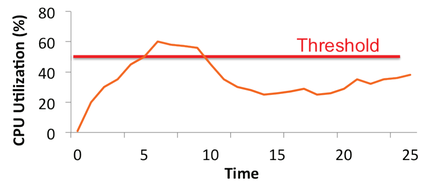
XLR8

Goal: Meet Application SLO

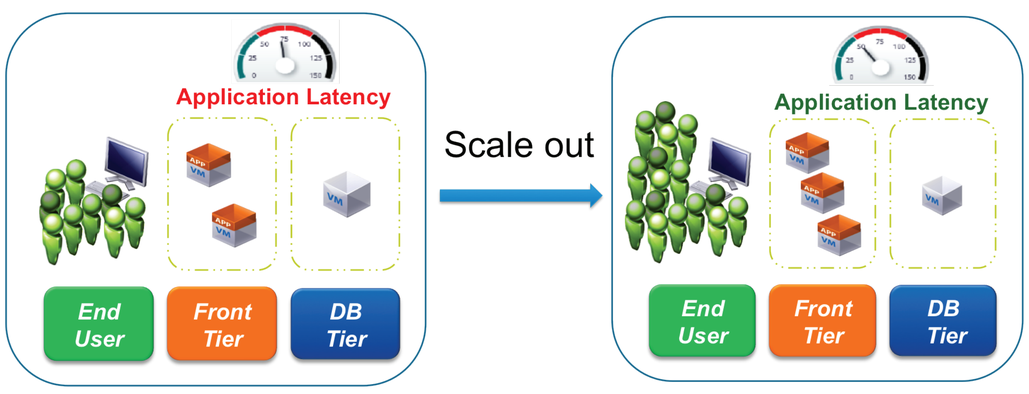
• **Allow** a cloud application to meet its service level objective (SLO) by automatically scaling its resource allocations
• **Reduce** SLO violations by meeting increased demands
• **Save** IT costs by removing excess VMs from an elastic tier or by right-sizing over-provisioned VMs

Traditional Approach

• User-defined threshold on a specific metric
• Scale out/up when threshold is violated
• Problems
- How to determine threshold values?
- How to handle multiple tiers?
- How to handle multiple resources?

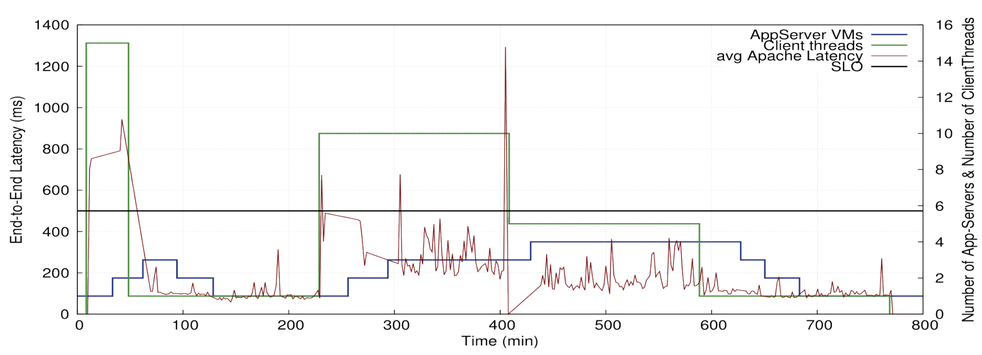


Horizontal Scaling

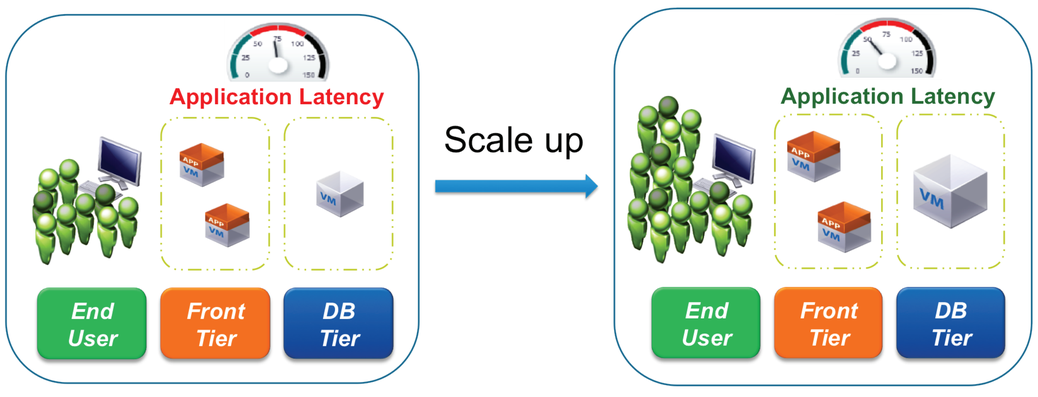


Approach:
• Uses machine learning to auto-learn application behavior
• Uses heuristics to seed the learning process
• Handles multiple resources and tiers

Evaluation:
• 3-tier Dukes Bank application with dynamic workload
• Learns quickly and adapts to demand changes automatically

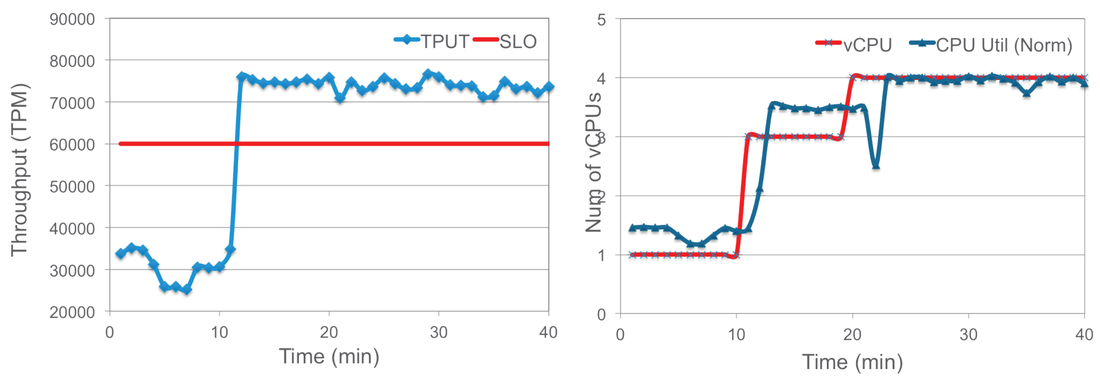


Vertical Scaling (Right-Sizing)

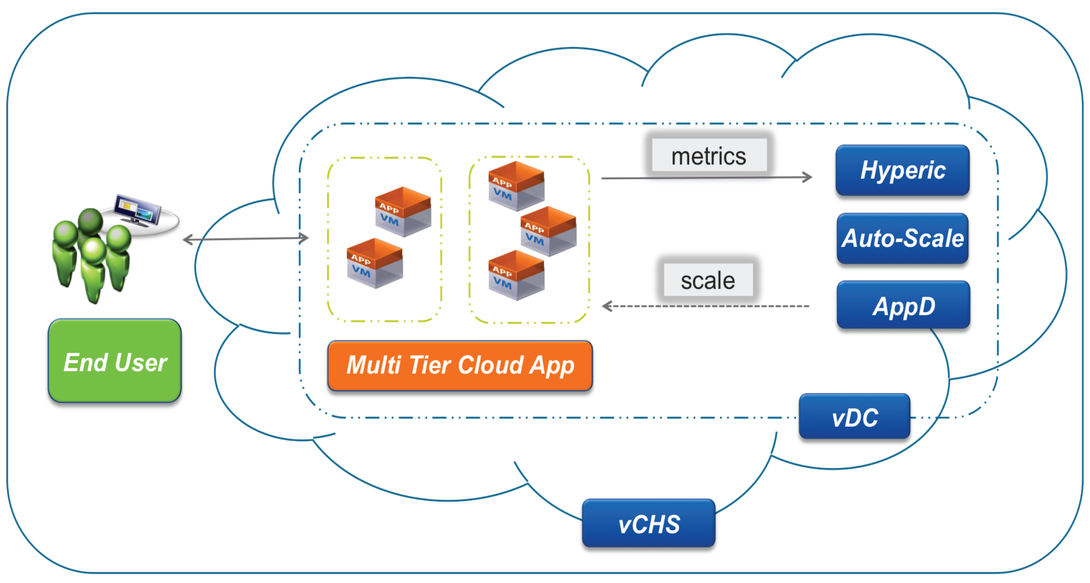


Approach:
• Uses learning to correlate performance with resource usage
• Automated online adaptation of resource configuration
• Reduces need for offline capacity planning and load testing

Evaluation:
• TPCC benchmark on MS SQL Server with varying user demand
• Automatically adjusts no. vCPUs and memory size via HotAdd



vCHS Auto Scale



* Check out our demo at the VMware OCTO booth!

Status

• **Technical papers published**
- Scaling of cloud applications using machine learning, VMTJ 2014.
- Application-Driven dynamic vertical scaling of virtual machines in resource pools, NOMS 2014.
- Runtime vertical scaling of virtualized applications via online model estimation, SASO 2014.

• **Ongoing Alpha program**
Email autoscale@vmware.com if you're interested in participating

vmware®