Not Too Big, Not Too Small. How Cisco IT Makes Virtual Machines Just Right, with Cisco Workload Optimization Manager (CWOM)

EXECUTIVE SUMMARY

CHALLENGE

- Increase data center efficiency
- Improve application performance
- Eliminate manual virtual machine placement

SOLUTION

- Automatically downsize virtual machines that are overprovisioned, with no downtime
- Automatically upsize virtual machines that need more resources

RESULTS

- Avoided $1 million in data center investments
- Reclaimed 52TB of memory, an average of more than 20GB per virtual machine.
- Recovered 4200 vCPUs
- Improved application performance by significantly reducing host contention

LESSONS LEARNED

- Start with a small virtual machine and let it grow as needed.
- Educate application owners
- Include all virtual machines in elastic compute program unless their owners opt out
- Consider peak utilization over several months before downsizing a virtual machine

NEXT STEPS

- Move virtual machines anywhere in the data center, using Cisco ACI and Tetration

Virtual machines are automatically upsized or rightsized. No human involvement. No downtime for upsizing.

Challenge

Two years. That's how long Cisco IT had before we'd run out of rack space in our Research Triangle Park (RTP), North Carolina data center. The Cisco Unified Computing System (UCS) hosts 60,000 virtual machines, and the count grows by 20 percent year over year.

Rather than constantly scrambling to add space, Cisco IT is working to free up 40 percent of the compute capacity in our global data centers. "Efficiency is extremely important in private clouds," says Todd Glenn, Cisco IT program manager. One of our tactics is to automatically downsize and resize virtual machines so that they're not reserving resources they don't need.

In fact, many virtual machines in RTP were bigger than necessary. One reason: server administrators had to manually resize virtual machines when their resource requirements grew. To save time, they sometimes provisioned the resources they predicted the virtual machine would need in two years. Unused resources wasted space.

Oversized virtual machines also worsened application performance. "If you're a party of four and request a table for eight, it takes longer to get seated and you're no better off with the larger table," says Glenn. "Virtualization works the same way. If the virtual machine is larger than needed, getting services from the Cisco UCS host takes longer." The system gets clogged, and other workloads also slow down.

What if virtual machines could configure themselves, grabbing more resources when needed and releasing them when no longer needed—all without human intervention? That's called elastic compute, and we wanted it.

Solution

Now every workload in our RTP data center gets exactly the resources it needs. No less and no more. We use Cisco Workload Optimization Manager (CWOM) to monitor infrastructure utilization in real-time. CWOM is a combination of UCS and Turbonomics. If utilization is less or more than the thresholds we've defined, the tool decides on the appropriate action: upsize or downsize.

Upsizing a virtual machine optimizes performance. If average utilization over the last 10 minutes exceeds our threshold, the tool adds vCPUs or memory—with no human intervention or downtime. Work isn't interrupted. Automatic upsizing means we can feel free to build smaller virtual machines. The standard is 2 vCPUs and 4GB memory. "Starting small and adding resources just in time improves performance and efficiency," says Chris Kosowan, Cisco IT engineer. "No virtual machine has too many or two few resources."

Downsizing a virtual machine improves efficiency. Downsizing requires a reboot, so we schedule downsizing for a maintenance window. To make sure that downsizing won't affect application performance, we base 90 percent of our decision on peak utilization over the last 120 days, and 10 percent on average utilization.

Process

We scheduled the transition to elastic compute for the same day as a planned shut down for maintenance.

Every virtual machine was enrolled by default for elastic compute—unless self-managed or used for disaster recovery. We built a self-service dashboard where application owners could opt out or decommission their virtual machines. The only valid reasons to opt out were if the workload could not take advantage of additional resources or the application had strict CPU license requirements. The FAQ we shared with application owners reassured them that their applications wouldn't suffer for lack of resources: "Automated upsizing enables 'hot add' of CPU and RAM resources without any downtime. Excess capacity doesn't have to be wasted waiting for future demand."

One week before change day, we validated the list of virtual machines and the automation playbook and then performed a dry run. On change day, in February 2017, we launched the resizing script in the evening. Within 4 hours, more than 2500 virtual machines were resized. Approximately 6500 virtual machines (51 percent of eligible virtual machines) were placed in the elastic upsizing pool. By 8:00 a.m. the following morning, all hosts scheduled for rightsizing were changed. When we turned the infrastructure back on, zero problems.

Results

$1 million in cost avoidance

Elastic compute freed up capacity on existing infrastructure for 2200 new virtual machines. Purchasing that amount of capacity would have cost $1 million. Instead, resizing 2600 virtual machines allowed us to:

- Reclaim 52TB of memory, an average of more than 20GB per virtual machine.
- Recover 4200 vCPUs

Performance Improved

Automatically downsizing virtual machines significantly reduced contention among virtual clusters. Less contention leads to better performance. "Host health has improved," says Kosowan. "And better health means fewer performance issues."

For illustration, imagine a 64GB virtual machine. You might think that all that memory would improve performance. But it actually degrades performance because acquiring the resources takes so long. What's more, finding a new host, if needed, takes longer. The problems ripple out to other workloads sharing the same resources. It's more efficient to resize the virtual machine to, say, 4GB, and let it grow to the right size based on actual usage.

Cisco IT Saved Time

Now server administrators spend less time figuring out compute and storage requirements. The administrator provisions the virtual machine in the right place for current needs. The virtual machine moves automatically when needed.

Next Steps

Now we're bringing elastic compute to all Cisco production data centers. We're also working on alternatives when virtual machine owners decide to opt out of right sizing. We expect those actions to free up another 50 – 100TB of memory and improve the health and efficiency of all IT infrastructure.

Longer term, we plan to use Cisco ACI and Tetration to move workloads freely anywhere in the data center. They'll no longer be confined to a specific cluster, pod, or Cisco UCS domain.

Lessons Learned

- Start with a small virtual machine and let it grow as needed. You won't have to reclaim resources later and you'll avoid downtime.
- Reassure application owners that performance will not decrease.
- Include all virtual machines in elastic compute program unless their owners opt out.
- Consider peak utilization (not just average utilization) over several months before downsizing a virtual machine.
- Give virtual machine owners an incentive to opt in by passing along the savings.

For More Information

To read additional Cisco IT case studies on a variety of business solutions, visit Cisco on Cisco: Inside Cisco IT www.cisco.com/go/ciscoit

Note

This publication describes how Cisco has benefited from the deployment of its own products. Many factors may have contributed to the results and benefits described. Cisco does not guarantee comparable results elsewhere.

CISCO PROVIDES THIS PUBLICATION AS IS WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Some jurisdictions do not allow disclaimer of express or implied warranties; therefore, this disclaimer may not apply to you.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**