NUMA and Virtualization, the case of Xen

Dario Faggioli, Senior Software Engineer, Citrix
dario.faggioli@citrix.com
What is NUMA

- **Non-Uniform Memory Access**: it will take longer to access some regions of memory than others.

- Groups of processors (NUMA node) have their own local memory.

- Any processor can access any memory, but accessing remote memory will be slower.
NUMA and Virtualization

What we wan to avoid:

NODE
- CPUs
- VM2
- mem
- VM1
- mem
- VM2
- MEM

NODE
- CPUs
- VM1
- mem
- VM2
- MEM
NUMA and Xen

What we used to have in Xen:

- NODE1: CPUs, MEM
- NODE2: CPUs, VM2, mem VM1, mem VM2, MEM
- NODE3: CPUs, VM1, mem VM1, mem VM2, MEM
- NODE4: CPUs, MEM
NUMA and Xen

What we used to have in Xen (by default):

NODE1
- CPUs
- mem
- VM1
- MEM

NODE2
- CPUs
- mem
- VM1
- MEM

NODE3
- CPUs
- mem
- VM1
- MEM

NODE4
- CPUs
- mem
- VM1
- MEM

August 29-31, 2012, San Diego, CA, USA
Dario Faggioli, dario.faggioli@citrix.com
Automatic Placement

What we have now:

1. VM1 creation time: **pin VM1 to the first node**

2. VM2 creation time: **pin VM2 to the second node, as first one already has another VM pinned to it**
However, if using **pinning** ...

... VM1 might have to wait, even if there are idle CPUs!
What we will have in Xen (4.3 release): node affinity, i.e., where a VM \textit{prefers} to run

VM1 can run immediately: remote accesses are better than not running at all!
Performances Evaluation

- Host: Intel Xeon(R) E5620, 16 cores, 12 GB RAM, 2 NUMA nodes
- VMs: 2, 4, 6, 8 and 10 of them, 2 vCPUs, 960MB RAM
- SPECjbb2005 executed concurrently in all VMs
- 3 configurations: all-cpus, auto-pinning, auto-affinity
- Exp. repeated 3 times per each configuration
average+stddev of the aggregate SPECjbb2005 throughput for all the VMs

17% to 13% improvement!

2 vCPUs per VM
Open Problems

- Dynamic memory migration
- IO NUMA
- Guest (or Virtual) NUMA
- Ballooning and memory sharing
- Inter-VM dependencies
- Benchmarking and performances evaluation
Dynamic Memory Migration

If VM2 goes away, we want to move VM1's memory!
Different devices can be attached to different nodes: needs to be considered during placement / scheduling
Guest NUMA

If a VM is bigger than 1 node, should it know?

**Pros:** VM performances

**Cons:** what if that needs to change?
- suspend/resume
- live migration

August 29-31, 2012, San Diego, CA, USA

Dario Faggioli, dario.faggioli@citrix.com
Ballooning and Sharing

Ballooning should be NUMA aware

Sharing, should we allow that cross-node?

Remote access!
Inter-VM Dependences

Are we sure the situation on the right is always better? Might it be workload dependant (VM cooperation VS. competition)
Benchmarking and Performances Evaluation

How to verify we are actually improving:

- What kind of workload(s)?

- What VMs configuration?
Thanks!

Any Questions?