LINUX PLUMBERS CONFERENCE

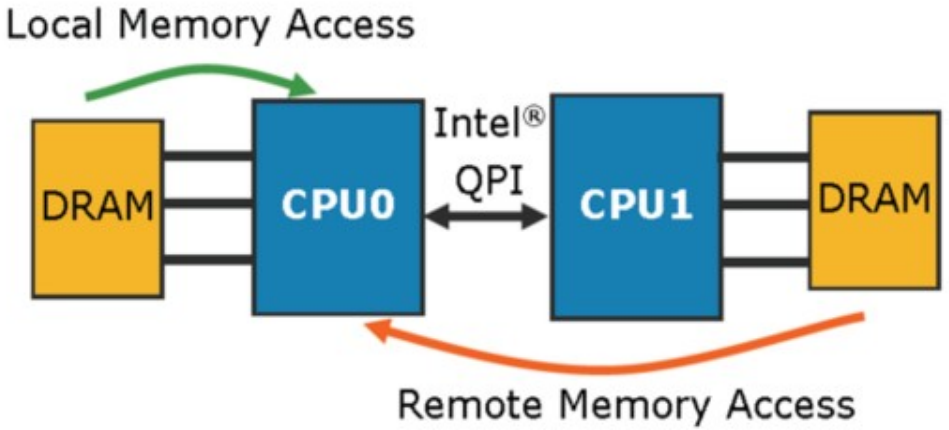
NUMA and Virtualization, the case of Xen

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

AUGUST 29-31, 2012 SAN DIEGO, CALIFORNIA

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

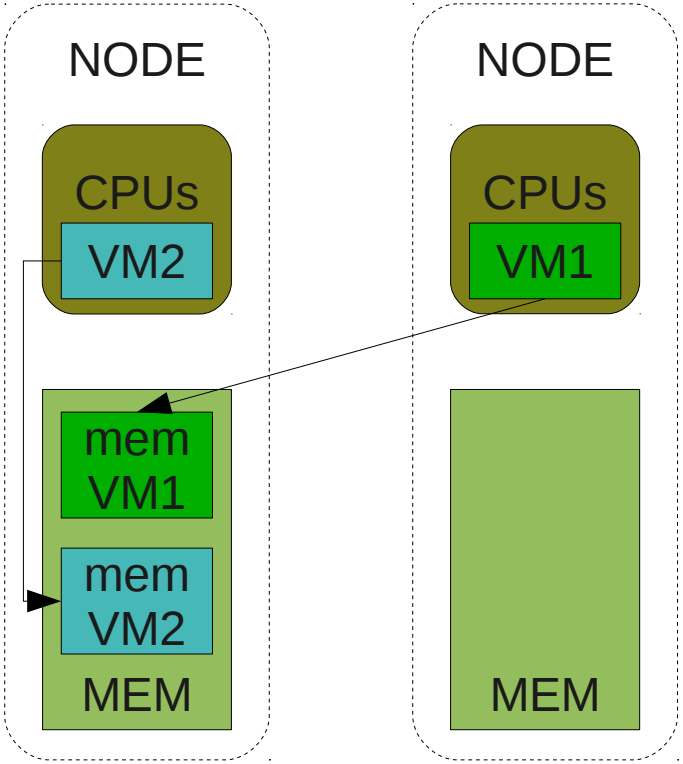


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

NUMA and Virtualization

What we wan to avoid:

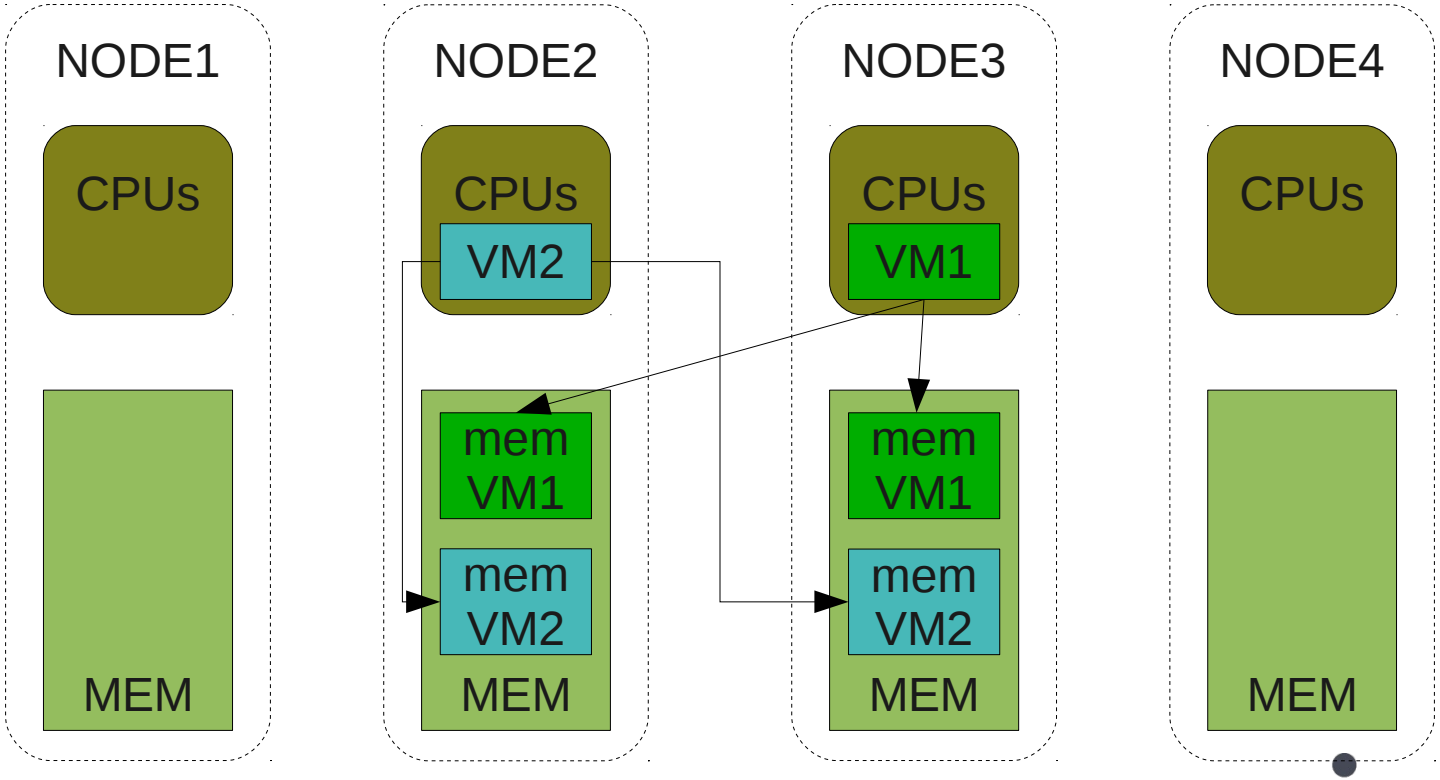


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

NUMA and Xen

What we used to have in Xen:

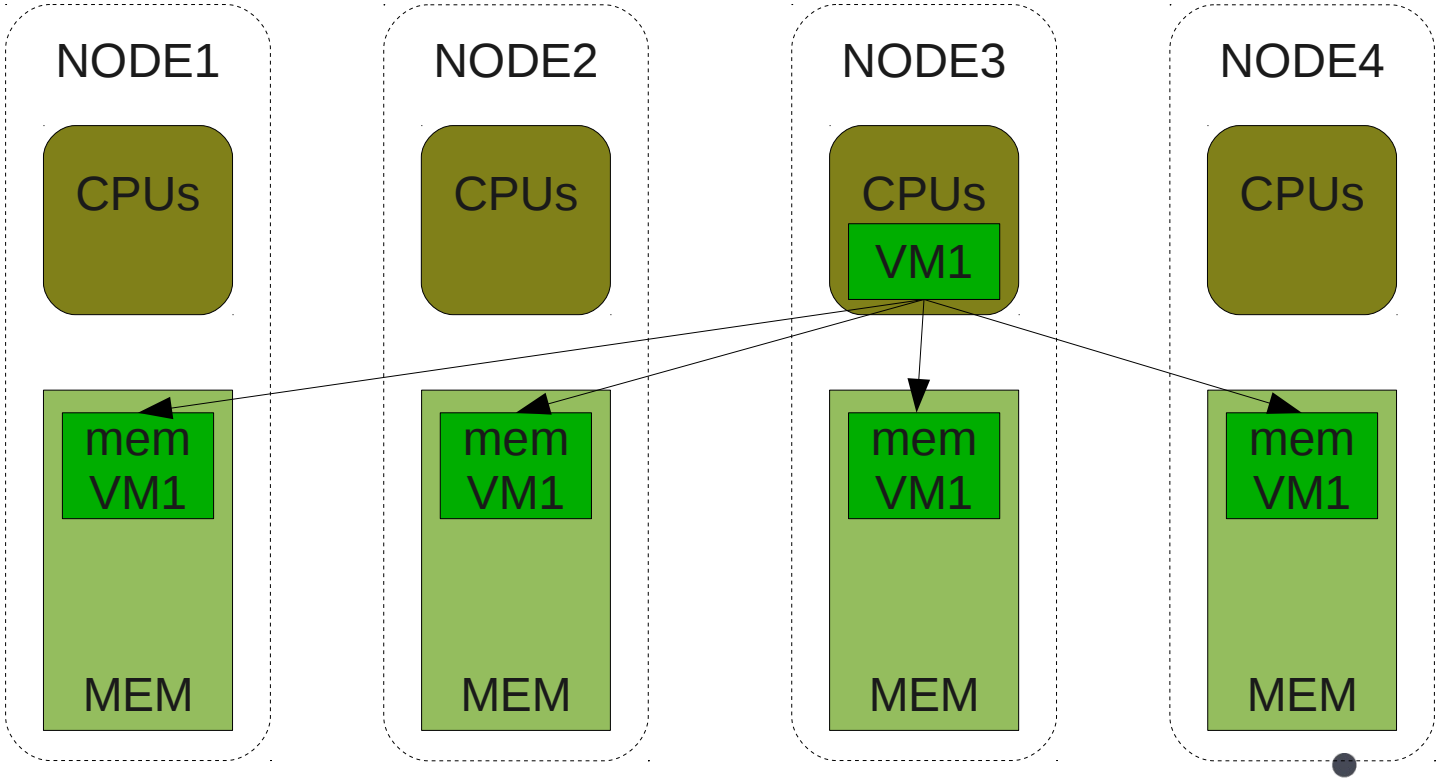


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

NUMA and Xen

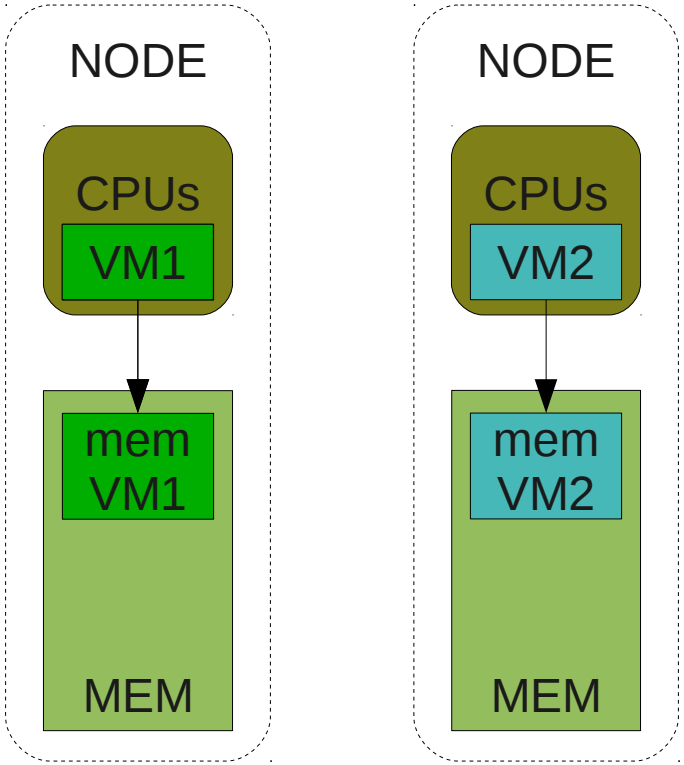
What we used to have in Xen (by default):



Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Automatic Placement



What we have now:

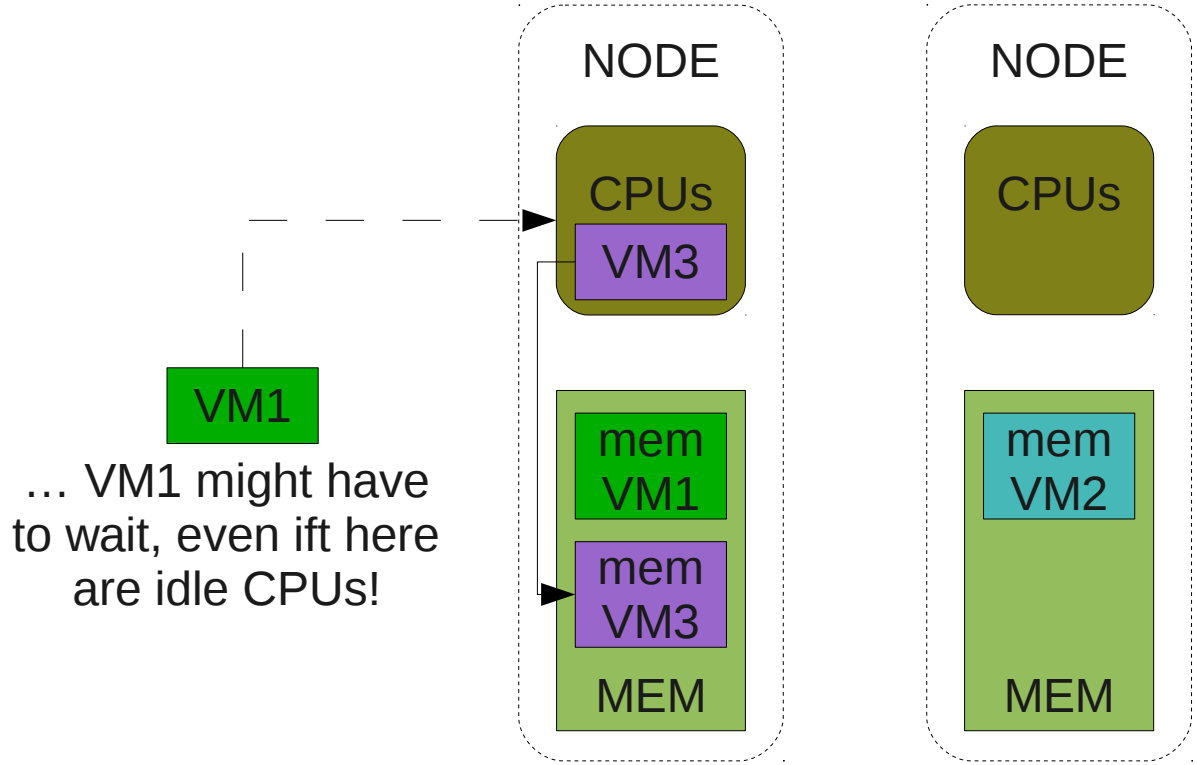
1. VM1 creation time: **pin** VM1 to the first node
2. VM2 creation time: **pin** VM2 to the second node, as first one already has another VM pinned to it

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

NUMA Aware Scheduling

However, if using **pinning** ...

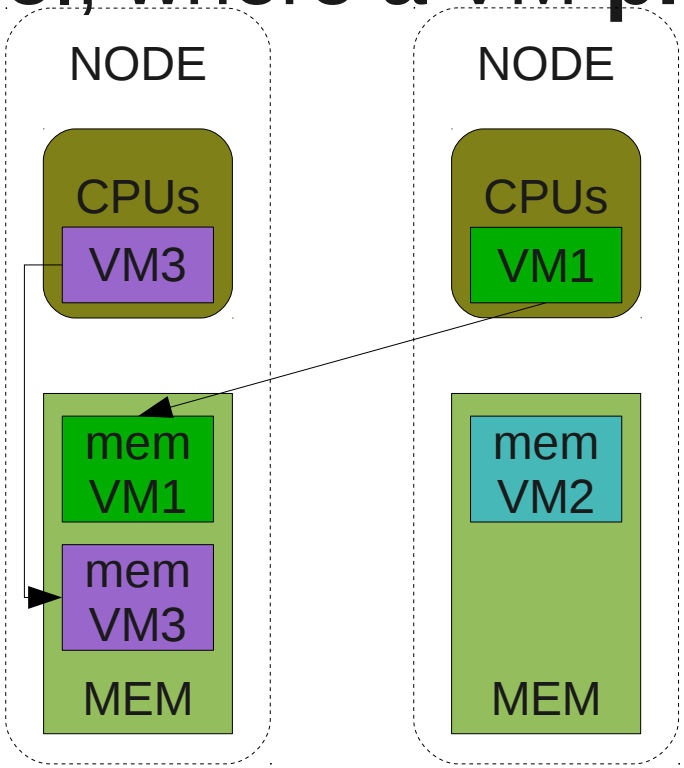


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

NUMA Aware Scheduling

What we will have in Xen (4.3 release): node affiniy, i.e., where a VM **prefers** to run



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

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

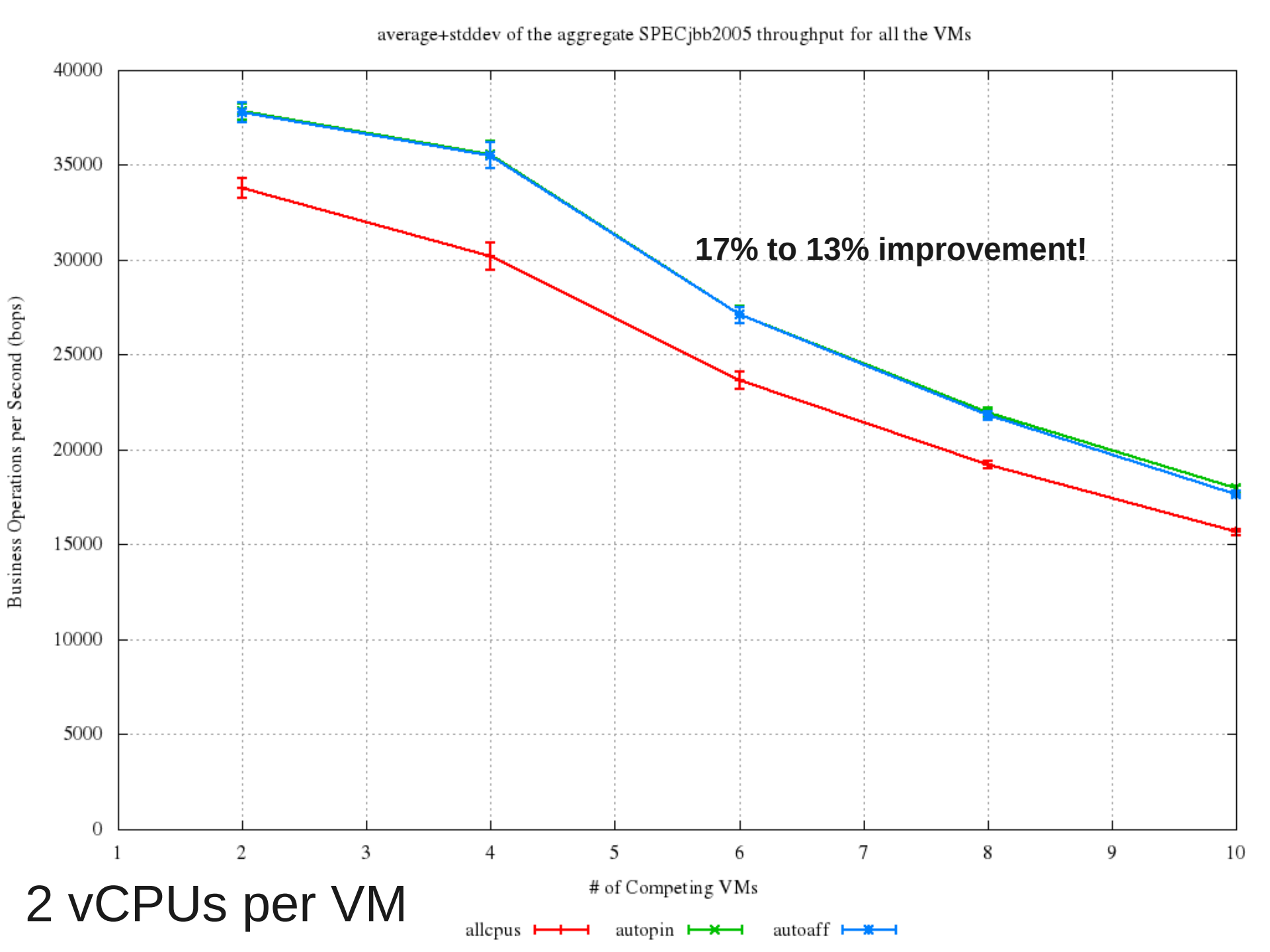
Performances Evaluation

- 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

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX



2 vCPUs per VM

Open Problems

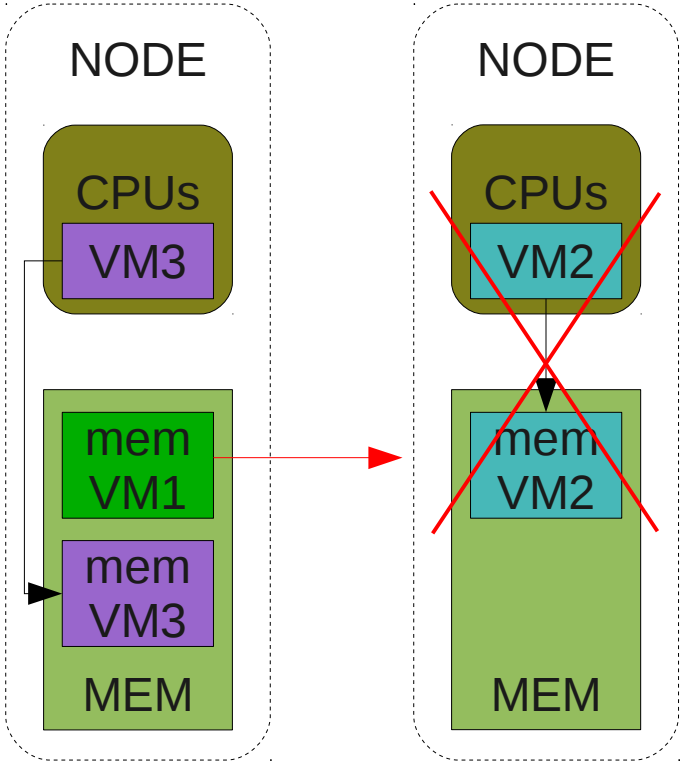
- Dynamic memory migration
- IO NUMA
- Guest (or Virtual) NUMA
- Ballooning and memory sharing
- Inter-VM dependencies
- Benchmarking and performances evaluation

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Dynamic Memory Migration

If VM2 goes away, we want move VM1's memory!

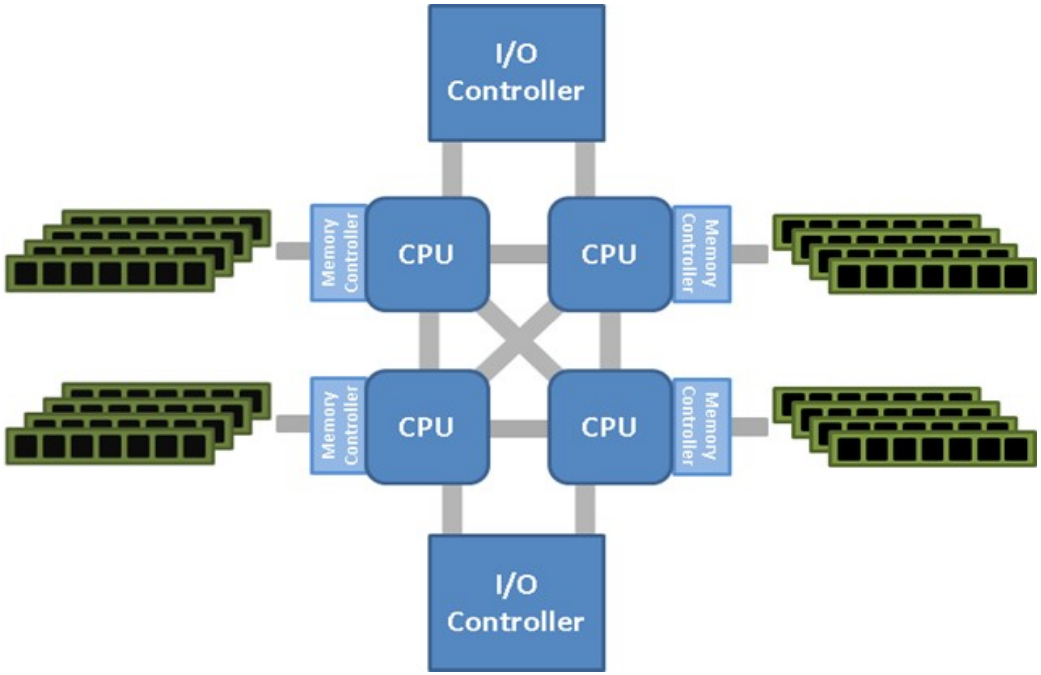


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

IO NUMA

Different devices can be attached to different nodes: needs to be considered during placement / scheduling

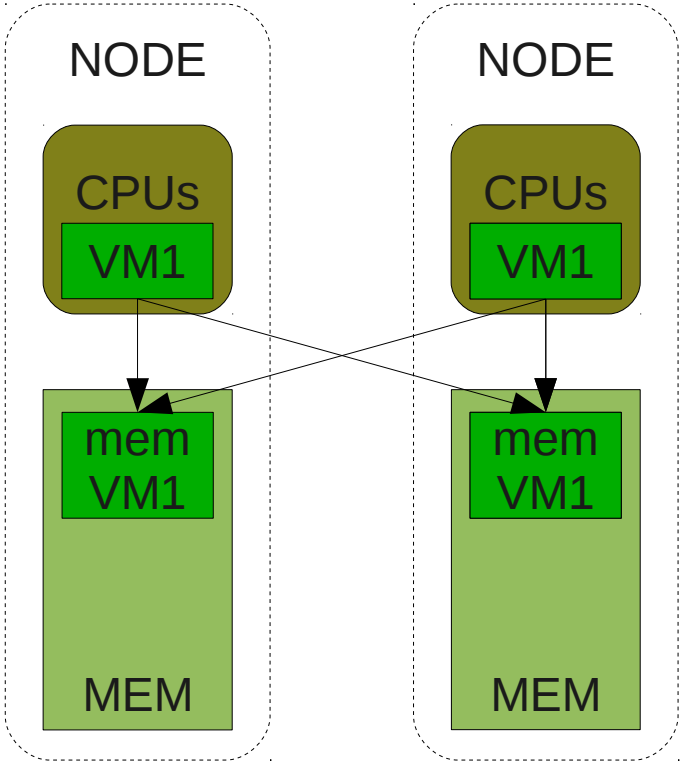


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

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

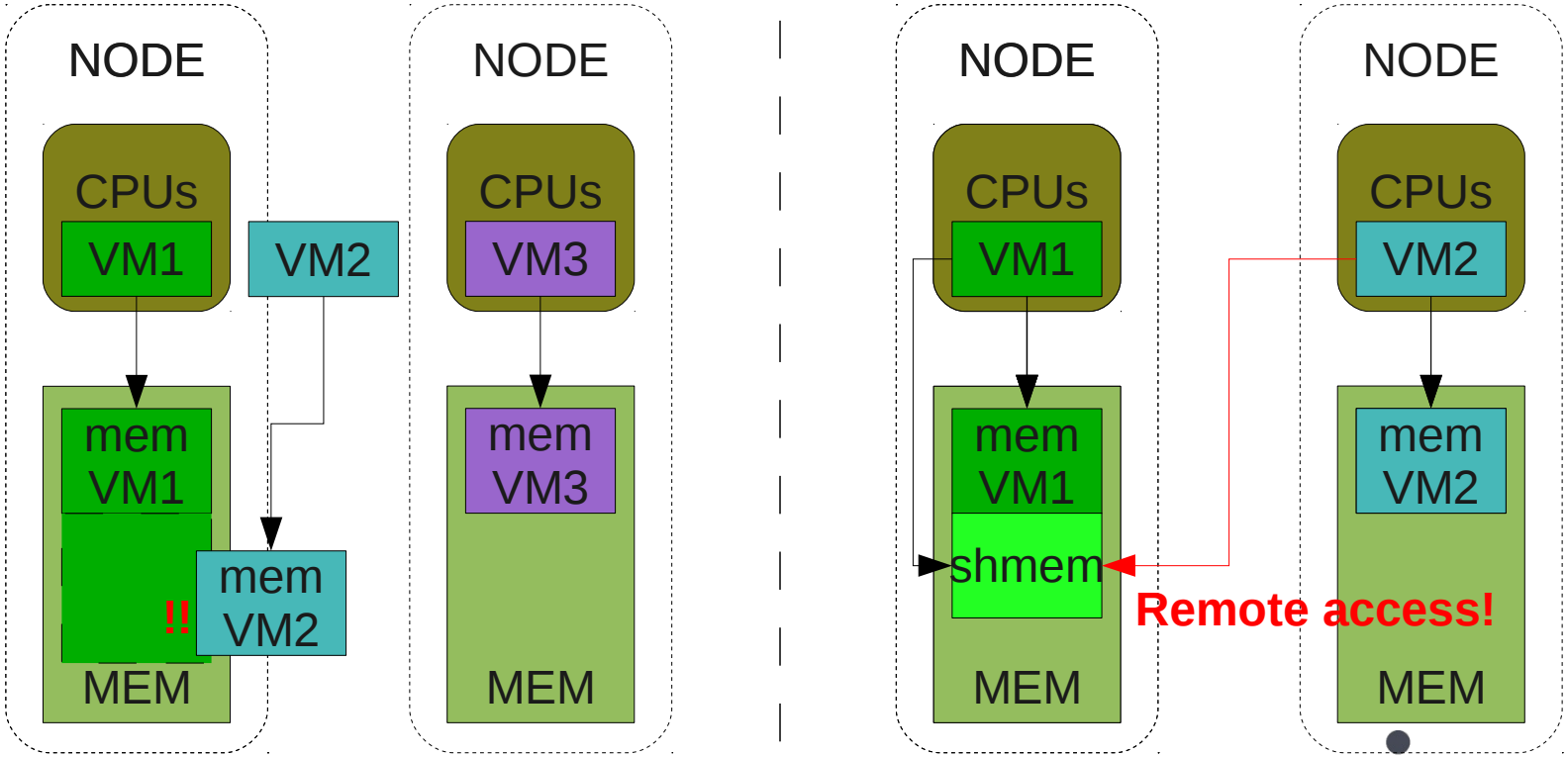
Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Ballooning and Sharing

Ballooning should be NUMA aware

Sharing, should we allow that cross-node?

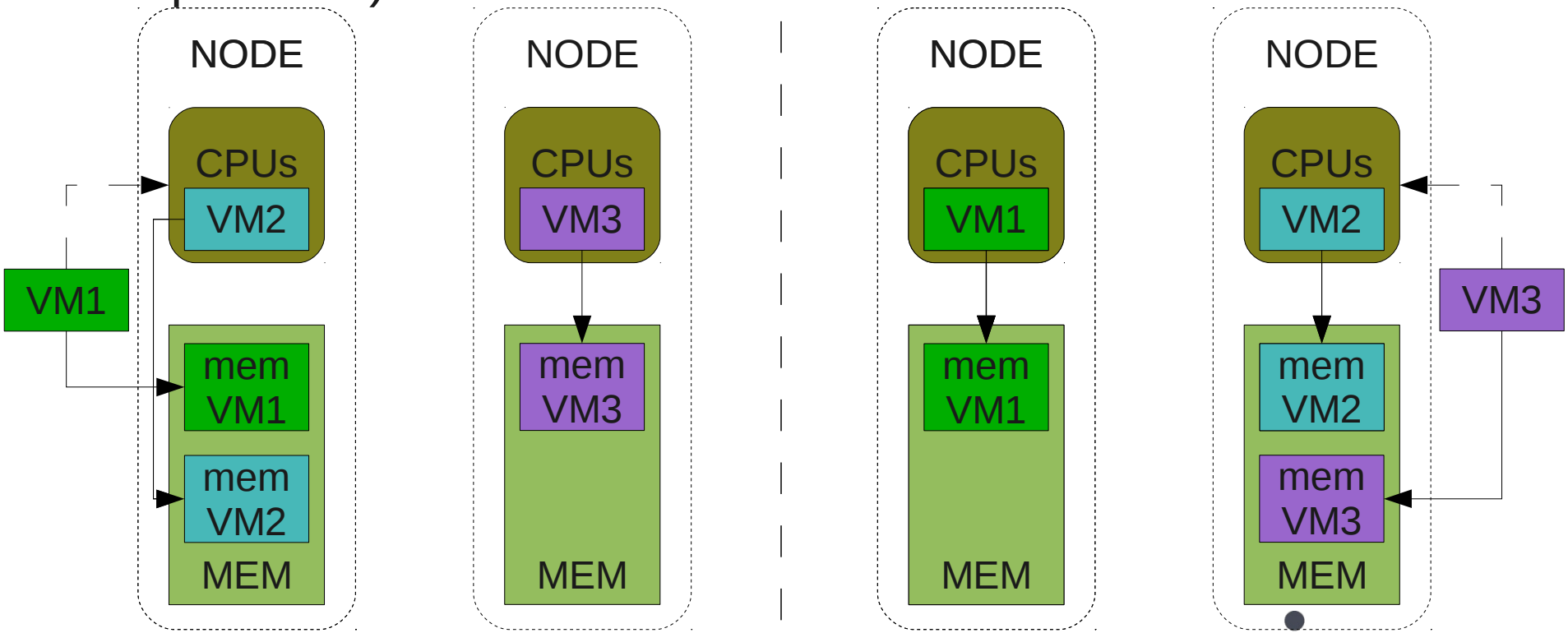


Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Inter-VM Dependences

Are we sure the situation on the right is always better?
Might it be **workload dependant** (VM cooperation VS. competiotion)



Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Benchmarking and Performances Evaluation

How to verify we are actually improving:

- What kind of workload(s)?
- What VMs configuration?

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX

Thanks!

Any Questions?

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

Dario Faggioli,
dario.faggioli@citrix.com

CITRIX