OPW - Xen vNUMA

Elena — OPW Xen Project Intern

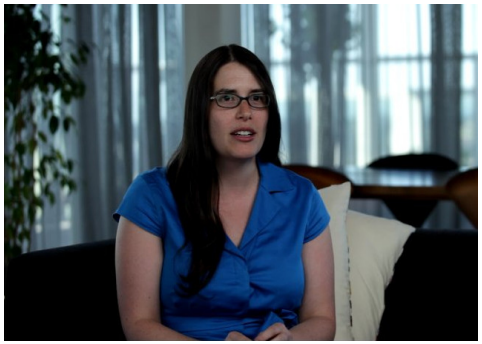
- was working on OPW all summer 2013
- had fun and really enjoyed coding
- best mentors in town

+

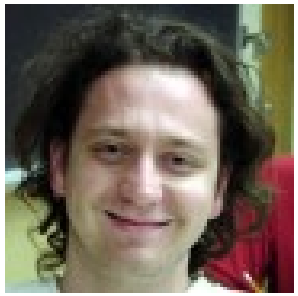
+

Mentors

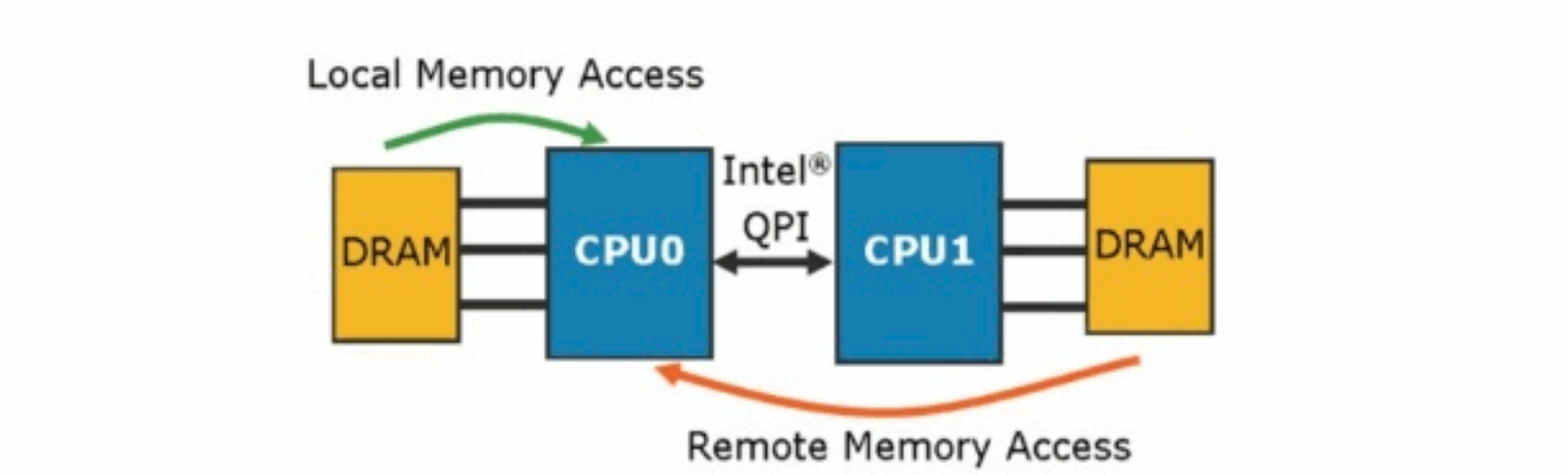
OPW
Mentors



Xen and kernel
mentors



NUMA and Linux



NUMA topology parsing

NUMA aware scheduling

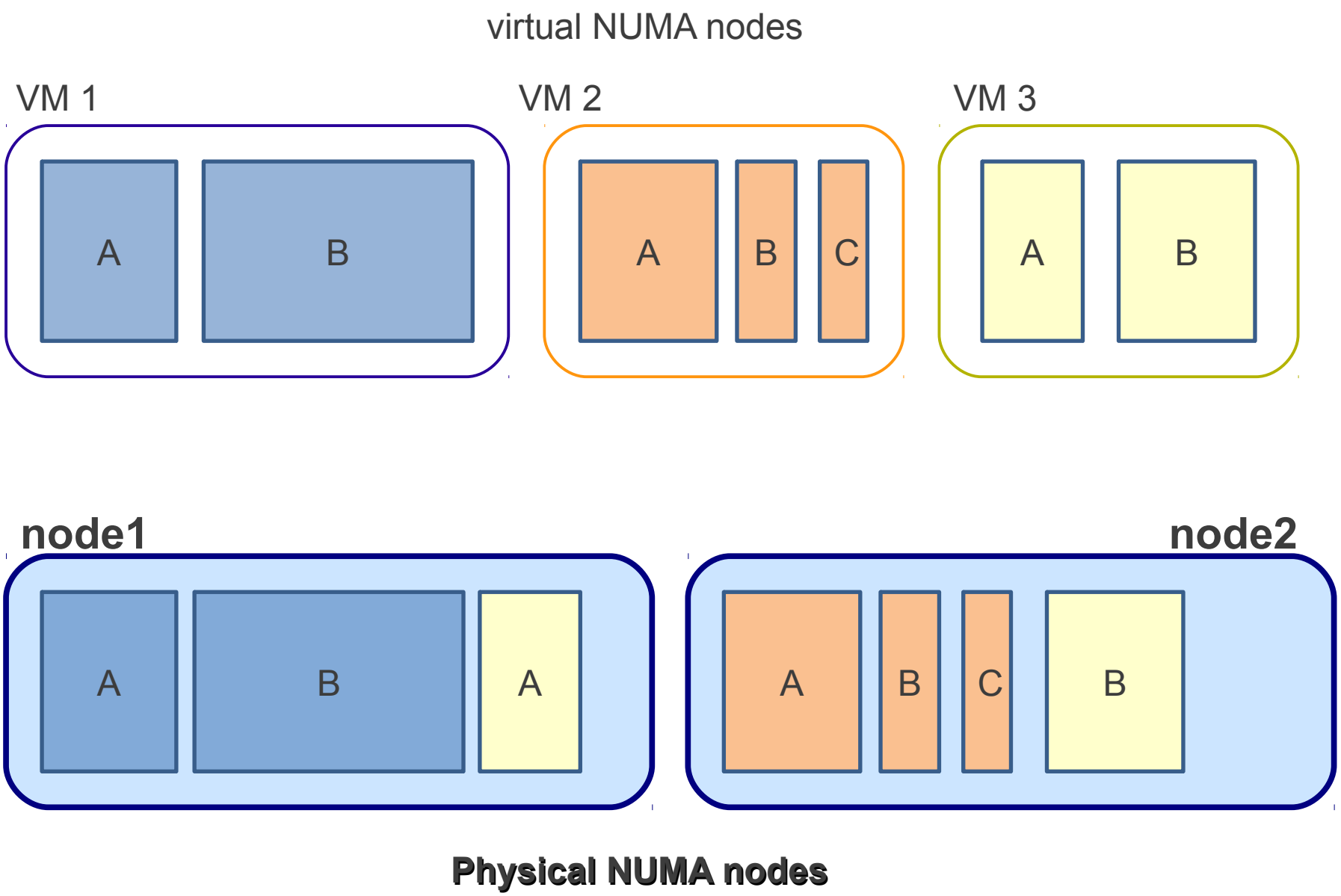
Automatic NUMA balancing

libnuma

Xen vNUMA

- 'enlighten' guest with vNUMA topology
- performance in mind
- let user decide on virtual topology
- bind to NUMA

pNUMA + vNUMA


virtual NUMA nodes
VM 1
A
B
VM 2
A
B
C
VM 3
A
B
node1
A
B
A
node2
A
B
C
B
Physical NUMA nodes

The Kernel Xen internship

- OPW and Linux Foundation
- Meet my mentors
- Facing dev community
- Grasp the idea
- Learn source code

What was done

- Xen and Linux patches
- Understand code
- Comments from community
- Feature freeze for Xen 4.4

What was done exactly

Configure vNUMA topology

```
name = "vm1"
memory = 4096
vcpus = 4
vnodes = 2
vnumamem = [2048, 2048]
vdistance = [10, 20]
vnuma_vcpumap =[1, 0, 1, 0]
vnuma_vnodemap = [0, 1, 1, 0, 1, 1]
```

```
(XEN) Memory location of each domain:
(XEN) Domain 0 (total: 2569511):
(XEN)     Node 0: 1416166
(XEN)     Node 1: 1153345
(XEN) Domain 5 (total: 4194304):
(XEN)     Node 0: 2097152
(XEN)     Node 1: 2097152
(XEN)     Domain has 4 vnodes
(XEN)         vnode 0 - pnode 0  (4096) MB,
(XEN)         vnode 1 - pnode 0  (4096) MB,
(XEN)         vnode 2 - pnode 1  (4096) MB,
(XEN)         vnode 3 - pnode 1  (4096) MB,
(XEN)     Domain vcpu to vnode: 0 1 2 3
(XEN) Domain 6 (total: 1048576):
(XEN)     Node 0: 524288
(XEN)     Node 1: 524288
(XEN)     Domain has 2 vnodes
(XEN)         vnode 0 - pnode 1  (2048) MB,
(XEN)         vnode 1 - pnode 0  (2048) MB,
(XEN)     Domain vcpu to vnode: 1 0 1 0
```

Boot vNUMA aware guest

```
root@heatpipe:~# numactl --hardware
available: 4 nodes (0-3)
node 0 cpus: 1
node 0 size: 511 MB
node 0 free: 488 MB
node 1 cpus: 0
node 1 size: 512 MB
node 1 free: 473 MB
node 2 cpus: 3 4
node 2 size: 512 MB
node 2 free: 496 MB
node 3 cpus: 2 5
node 3 size: 512 MB
node 3 free: 486 MB
node distances:
node   0   1   2   3
  0:  10  20  20  20
  1:  20  10  20  20
  2:  20  20  10  20
  3:  20  20  20  10
root@heatpipe:~# numastat -zc

Per-node numastat info (in MBs):
                 Node 0 Node 1 Node 2 Node 3 Total
                 ------ ------ ------ ------ -----
Numa_Hit             77    149     76     84   387
Interleave_Hit        8      8      8      8    32
Local_Node           67    149     66     73   354
Other_Node           10      0     11     11    32
```

What has to be done

- A lot! :)
- Dom0/HVM NUMA awareness
- Automatic NUMA balancing support
- Performance evaluations
- Heuristics/Statistics

What I have learned

- Reading source code
- Patches
  - Big patches are bad
  - Review process
  - Git is great
- Xen and para-virtual guests
- Linux, booting, memory management
- Time management

Thank you

- Linux Foundation and GNOME
- XenProject, Citrix, Oracle

- Xen project mentors and Linux Kernel Mentors

Dario Steffano George **Konrad** *Sarah* Greg