

NUMA support in Xen

September 2006

Ryan Harper, IBM
Ryan Grimm, IBM

Linux



NUMA Requirements

- NUMA infrastructure in hypervisor
 - Data structures required to represent NUMA characteristics
 - Boot-time detection/discovery
- NUMA-aware memory allocators
 - Modify page allocator to optimize for locality
- NUMA-aware tools
 - allocate/pin VCPU to CPU with NUMA awareness
 - expose NUMA topology in userspace
- NUMA-aware Xen-Linux
 - expose NUMA characteristics of a domain to Guests



NUMA Infrastructure

■ Detection

- SRAT table parsing via Linux ACPI NUMA driver
- Use modified Linux X86_64 NUMA support to build CPU and Memory affinity structures for both 32 and 64 bit Xen

■ Memory Affinity

- Map physical memory address to nodes

■ CPU Affinity

- Map which cpus are in which node



NUMA-aware Memory Allocation

- Make heap allocator NUMA-aware
 - Introduce per-node bucket in Xen heap per Winter Summit
 - `heap[ZONE][NODE][ORDER]`
 - Determine locality during heap initialization
 - Reserve guard pages on node boundaries that aren't `MAX_ORDER` aligned to prevent cross-node merging in buddy allocator
 - Determine locality with added CPU parameter to heap API
- Modify domain memory reservations to use NUMA API
 - Use first VCPU in domain to determine locality
 - Assumption that most domains will fit within a node
 - Ensure VCPU/CPU mapping in place prior to memory allocation



NUMA-aware Tools

- Extend physinfo hypercall
 - Add NUMA machine characteristics (nr_nodes, memchunks)
- Modify tools to display NUMA information
 - xm info displays nr_nodes, array of memchunks, node_to_cpus
- Add program to probe heap and NUMA information
 - display free pages in heap (amounts per zone,node and order)
 - display a domain's page distribution (which nodes used)



Current Status

- Latest iteration against changeset 11134:ec03b24a2d83
 - Unisys mentioned possible dom0 slowdown, waiting on debug info from dom0
 - Will resubmit patches shortly to update to new split dom0 ops

