An architecture for space sharing HPC and commodity workloads in the cloud

Jack Lange
Assistant Professor
University of Pittsburgh
HPC in the cloud

• Palacios VMM
  – HPC oriented VMM
  – Specifically targeting supercomputing environments

• Current state of supercomputing OSes
  – Linux appears to be the future
    • Not necessarily a good thing

• Palacios and Linux
  – Lightweight environments on heavyweight OSes
Palacios VMM

- OS-independent embeddable virtual machine monitor
- Open source and freely available
  - 1000s of unique downloads
- Users:
  - Kitten: Lightweight supercomputing OS from Sandia National Labs
  - MINIX 3
  - Linux

- Ideally suited to lightweight supercomputing OSes
- Successfully used on supercomputers, clusters (Infiniband and Ethernet), and servers

Palacios
An OS Independent Embeddable VMM
http://www.v3vee.org/palacios
A brief Sandia OS history

• **SUNMOS** (Sandia/UNM Operating System)
  – In response to Paragon issues
  – Library OS
  – Single task, no demand paging

• **PUMA/Cougar/Catamount**
  – Multitasking
  – Portals: High performance message passing API

• **Kitten**
  – Lightweight kernel based on Linux ABI
  – Built from the ground up
    • Ported selected features from Linux

• **CNL** (Compute Node Linux)
  – Cray
  – Modified version of Linux
Palacios as an HPC VMM

• Minimalist interface
  – Follows the “lightweight OS” philosophy

• Compile and runtime configurability
  – Create a VMM tailored to specific environments

• Low noise

• Contiguous memory allocations

• Passthrough resources and resource partitioning
HPC Virtualization

• Virtualization is very useful for HPC, but...
  Only if it doesn’t hurt performance

• Virtualized RedStorm with Palacios and Kitten
  – Evaluated with HPC system benchmarks

  Cray XT4
  204,200 cores
  284 TFlops
  2.5 MegaWatts
Scalability at Large Scale (Weak Scaling)

Catamount Guest OS

Within 3%

Scalable

CTH: multi-material, large deformation, strong shockwave simulation
Supercomputing and Linux

• Lightweight kernels are great for application performance
  – But their days seem to be numbered
  – Probably not operational OS of choice for new supercomputers
• Linux seems to be taking over
  – Compute Node Linux (CNL), ZeptoOS, etc

• For some of us, this is not necessarily a “good thing”
  – Linux introduces significantly more overhead
  – Subset of applications are hurt (e.g. SAGE ~17% slowdown on CNL)
    – Ironically Linux is the more closed environment

• Challenge
  – Can we provide a lightweight environment on a heavyweight OS?
    – If we can do this on a supercomputer why not in the cloud?
HPC in the cloud

• Are supercomputers converging with the cloud?
  • No (at least not yet)
    – Noise issues
    – Consolidation
    – Non reserved resources
    – Uncontrolled topology

• Very bad for tightly coupled parallel apps
  – Can virtualization provide a solution?
Why not other VMMs?

• Virtualization is considered an “Enterprise” tool
  – Not for HPC

• Example: KVM design issues
  – Userspace handlers
  – Loose CPU affinities
  – Complicated memory management
  – HW resets on context switch
Palacios and Linux

• Palacios provides a lightweight environment
  – Internally manages node resources
  – Does not bother with “enterprise features”

• Claim: Palacios can provide a scalable HPC environment on commodity platforms
  – Turn a commodity cloud node into a supercomputer node
Palacios on Linux

• Bypass Linux management layers

• Palacios selectively takes over resource management
  – Memory, devices, CPUs
  – Repurpose existing mechanisms
    • Allocate large chunks of resources and manage them internally

• Kernel module
  – Does not require kernel modifications
  – Implements lightweight interface
  – Compatible with Fedora 15 kernel

Available in Palacios 1.3 (Nov. 2011)
Palacios and Linux

Hardware

Hadoop Processes  KVM VM  HPC VM

Linux/KVM  Palacios VMM

Resource Manager

Direct User Feedback

Application Performance Measurements
Palacios and Linux memory

- Linux disables selected memory blocks
  - Internally (via memory hotplug)
  - Large contiguous blocks (128MB / block)
  - Memory is still accessible!

- Palacios assumes responsibility of disabled blocks
  - Uses internal memory allocator

- Palacios and Linux coexist without interference
  - User chooses physical memory regions to assign
Palacios memory hierarchy

- HPC VM Application
- Guest Page Tables
- LWK allocator
- Palacios VM Memory
- Palacios Memory Manager
- Palacios Assigned Memory
- Linux Assigned Memory
- Linux SL*B allocator
- OS Memory manager
- Physical RAM
- Page Tables
- Linux Application

- Physical RAM
- OS Memory manager
- Page Tables
- Linux Application
- Linux Assigned Memory
- Linux SL*B allocator
Results

• Evaluations are currently underway
  – Have collected preliminary results

• HPC Benchmark (hpccg), run time (secs)
  – 1 VM, 4 cores

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Mean</th>
<th>Stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Kitten</td>
<td>15.15</td>
<td>0.02</td>
</tr>
<tr>
<td>Kitten guest (Palacios)</td>
<td>15.28</td>
<td>0.06</td>
</tr>
<tr>
<td>Kitten guest (KVM)</td>
<td>18.28</td>
<td>2.96</td>
</tr>
<tr>
<td>Host Linux</td>
<td>15.93</td>
<td>2.68</td>
</tr>
</tbody>
</table>

High standard deviation implies poor scalability
Open issues with I/O

• IOMMU’s and SRIOV are now prevalent
  – Direct access to network hardware
  – Palacios can provide high performance I/O

• Two issues
  – VM placement (possibly random topology)
  – Traffic contention

• These should be solvable
  – Intelligent placement and scheduling
  – Specialized network architectures
    • Optical networks, reservable networks, others
Infiniband on Commodity Linux

(Linux guest on IB cluster)

![Graph showing bandwidth measurement for Native and Virtualized systems vs message size.](image)

2 node Infiniband Ping Pong bandwidth measurement
Conclusion

• **Clouds are not supercomputers**
  – Despite their claims to the contrary
  – Virtualization can bring them closer to convergence

• Virtualization can bypass host OS overheads

• Palacios and Linux
  – Lightweight environments on heavyweight OS
Thank you

- **Jack Lange**
  - jacklange@cs.pitt.edu
  - http://www.cs.pitt.edu/~jacklange

- **V3Vee Project**
  - http://www.v3vee.org