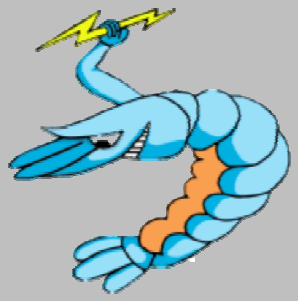
SAND2013-6890P



Hobbes
Node Virtualization Layer (NVL)

OS/R Program Kickoff Meeting
August 15, 2013
@ Argonne National Laboratory

Coordinator: Kevin Pedretti (SNL)
Participating: SNL, LANL, ORNL, NWU, PITT, UNM

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

NNSA

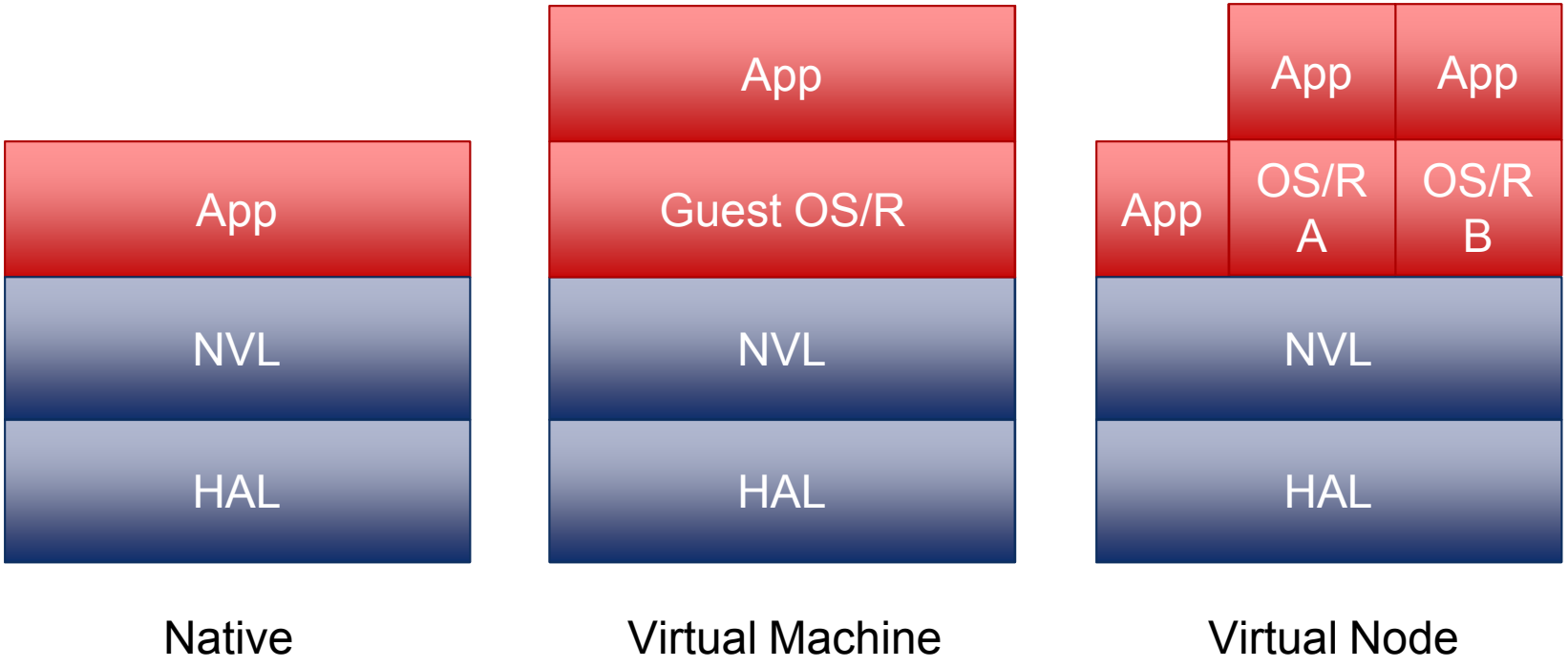
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Why Virtualization?

Sandia National Laboratories

- Flexibility, support multiple OS/R stacks simultaneously
  - There is likely to be no one-size-fits-all OS/R stack, lots of exploration
  - Co-location of VMs, efficient sharing of resources between enclaves
  - Native environment freed from legacy constraints
- Low overhead
  - Our past work has shown CPU and memory overheads negligible
  - Network I/O is still an issue, but tractable
- Industry momentum
  - Virtualization has been commoditized, is everywhere
  - Academic and student mindshare, where the jobs are
- Mostly orthogonal to “FusedOS” approach
  - FusedOS could run in NVL VM or natively, in the same machine
  - NVL could be co-designed with FusedOS

NVL Use Cases



- Virtual nodes allow physical node resources to be shared across co-located OS/R stacks
- New mechanisms and interfaces needed to manage sharing
- NVL interfaces with outside world to manage virtual nodes

Starting Point

Sandia National Laboratories

- Kitten Lightweight Kernel (LWK)
  - Sandia led development, 2008 – present
  - Influenced by Catamount, uses code from Linux (GPLv2'ed)
  - Supports x86_64, Linux ABI, multi-core, NUMA, PCI, Portals
  - Current work funded by DOE/ASCR XPRESS project, NNSA
- Palacios Virtual Machine Monitor (VMM)
  - Northwestern, New Mexico, Pittsburgh led development, 2006 – present
  - Influenced by Virtuoso, all new development (BSD licensed)
  - Supports x86_64, multi-core, NUMA, PCI pass-through
  - Current work funded by DOE/ASCR X-Stack 1 project, NSF, NNSA
- Both will be restructured to fit integrated HAL / NVL model
  - Have begun discussing internally
  - Interfaces between HAL and NVL to be improved, broadened

R&D areas

Sandia National Laboratories

- High Performance Virtualization
  (SNL, LANL, ORNL, NWU, PITT, UNM)
  - HAL / NVL architecture
  - Interfaces for virtual node composition
  - Network stack “virtualization”
  - Quality of Service
- High-Risk / High-Impact
  (NWU, PITT, UNM)
  - NVL-level autonomic adaption for specified power, energy, and/or performance goals (in Power Presentation)
  - Hybrid virtual machines for parallel language OS/R stacks, provide custom virtual cores specialized for language implementation
  - Para-native approach, run multiple native OS instances simultaneously without relying on hardware virtualization

HRHI: Hybrid Virtualization

Sandia National Laboratories

- Research question: How do we create a virtual environment that is better suited to parallel language run-times?
- Hybrid VM: commodity virtual cores for the general purpose OS and custom virtual cores for the language run-time
- Custom virtual core
  - Lightweight exit handling customized for run-time
  - Reduced or eliminated coherence
    - Initially emulated by VMM, hopefully eventually have hardware support
  - Specialized services in the VMM to support the run-time

HRHI: Hybrid Virtualization (cont.)

Sandia National Laboratories

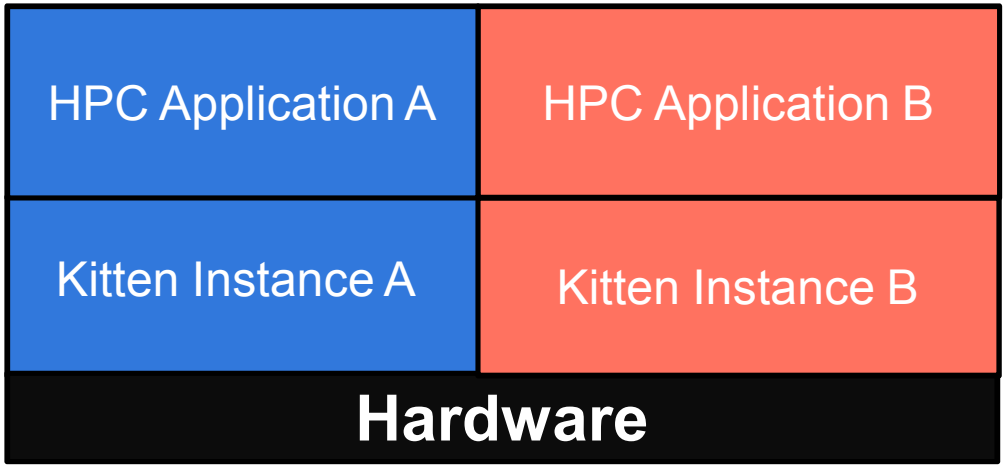
Preliminary Efforts:

- Racket Dynamic Language Run-time Efforts
  - Example Paper: J. Swaine, K. Tew, P. Dinda, R. Findler, M. Flatt, *Back to the Futures: Incremental Parallelization of Existing Sequential Runtime Systems*, OOPSLA 2010
  - Racket run-time port to Kitten (J. McClurg)
- VMM-based Hardware Transactional Memory Emulation
  - M. Swiech, K. Hale, P. Dinda, *VMM-based Emulation of Intel Hardware Transactional Memory*, NWU-EECS-13-03
- GEARS for services that extend into the guest
  - K. Hale, L. Xia, P. Dinda, *Shifting GEARS to Enable Guest-context Virtual Services*, ICAC 2012
  - K. Hale, P. Dinda, *Guarded Execution of Privileged Code in the Guest*, NWU-EECS-13-04

HRHI: Para-Native Architecture

Sandia National Laboratories

- Partition node resources between multiple native OS instances, each with direct hardware control
  - Strong isolation between OS instances
  - Not a stacked architecture, the multiple OSes are peers
  - Useful on systems with no hardware virtualization support
- Current prototype boots multiple instances of Kitten on a node



Sandia National Laboratories

Questions and Discussion

Participants

- Sandia (coordinating institution)
  - High perf. virtualization
- Los Alamos
  - High perf. virtualization, interfaces to node OS/R's
- Oak Ridge
  - Performance isolation, network stack sharing
- Northwestern U.
  - Autonomic adaption, hybrid virtual machines
- Pittsburgh U.
  - Para-native, multiple simultaneous native OS/R stacks
- U. New Mexico
  - High perf. virtualization, scheduling interfaces

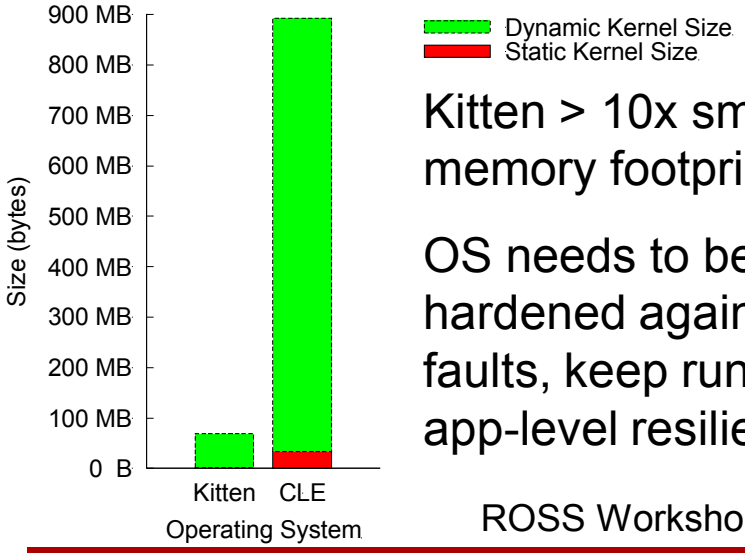
There are Many Levels of Virtualization

Sandia National Laboratories

- Full system virtualization
  - Supports unmodified guest OS/R stacks
- Para-virtualization
  - Supports slightly modified, cooperating guest OS/R stacks
- Java virtual machine
  - Idealized intermediate machine architecture, designed to be JIT'ed
- UNIX
  - Portable abstractions, mapped to underlying hardware by OS

Exploring Full Spectrum of Virtualization

Results



Kitten > 10x smaller memory footprint

OS needs to be hardened against faults, keep running for app-level resilience

ROSS Workshop Paper

Operating System	Round-trip Task Migration Time (task migrates from core A to B and back)
Linux 2.6.35.7	4435 ns
Kitten 1.3	2630 ns

Kitten integrated SST/gem5 to enable rapid prototyping and reproducibility

SimuTools'12 Paper

SST CPU and Memory Model Implemented by Palacios VM