MegaTux:

SAND2012-6701C

Using Lightweight Virtualizations for Multi-Million-Node Emulations

Discovery 2020

John Floren
Sandia National Labs, CA

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.

Motivations and background

- 2007: 100-node clusters on a laptop
- Can we make a cluster of this?
- If we have hundreds of nodes, each running hundreds of VMs, that grows pretty quickly
- Use it to model large-scale networks of computers
  - Peer-to-Peer systems
  - Bittorrent

Why large-scale simulation?

- Trying to do things at the Internet scale is difficult
  - You may see unexpected behavior
  - Current HPC still struggles beyond ~10,000 nodes
  - Faults are assured
- Testing at scale?
  - Very expensive, very few have the resources
- We're comfortable with micro analysis, but it won't show emergent behavior at scale
- Macro analysis is not as well developed

MegaTux 1.0

- Goal: Run lots of Linux VMs (over 1 million)
- Ran 1 million VMs on Thunderbird
  - 4,400 node supercomputer
  - Used small lguest VMs
  - Simple, flat network
- Write-up in the New York Times (worth reading)

Lessons learned

- Hardware limitations: switches crashed when they saw too many client MAC addresses
- DNSMasq doesn't scale to 1 million clients
- If a VM fails to boot 1 out of 100,000 times, you still end up with 100 missing VMs

Expanding the work

- We wanted to run Windows and Android
- Wanted more complex and flexible networks
- Needed better control over the cluster
  - Faster distribution of VM images is important too, especially for Windows

GProc

- Gproc is a large-scale cluster management tool written in Go
- Like LANL's BProc, it allows remote process creation and management, plus file distribution
- GProc arranges cluster nodes into a tree
  - This gives O(log n) execution time
- Commands and files are transmitted down the tree, reducing the load on the master and improving speed
- Output of commands gets passed back up the tree to the master

KANE

- A new cluster build specifically for MegaTux
  - Flexible
  - Fast recovery: nodes cycle in about 2 minutes
  - Expandable
  - Divisible
  - Cheap ($900 per node)
  - Core i7 930 processors, 12 GB RAM, Gb Ethernet
- Uses our own in-house scheduler

KANE (half of it)

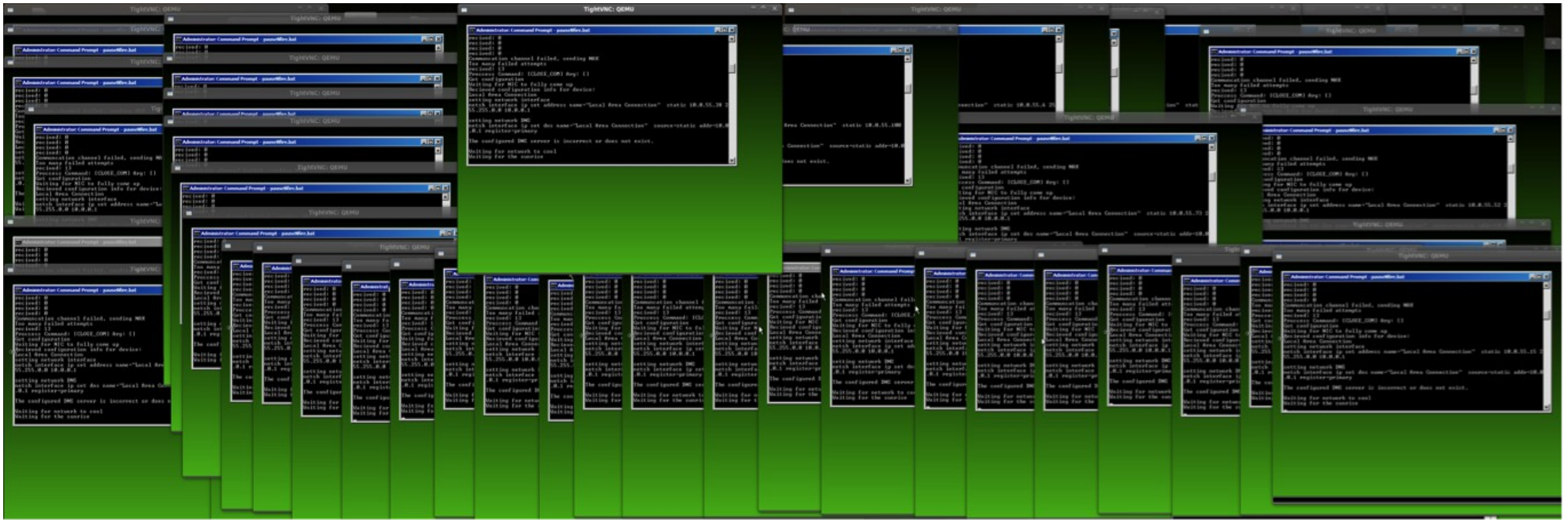
New VM launcher

- A new tool was written to launch VMs
- Allows different OSes at the same time
- More complex networks
  - VDE switching
  - Virtual routers (Quagga, Linux, etc)
- Uses KSM (Linux memory de-duplication) to fit 45 GB into 12 GB
- Control OOM killer to keep the system stable

Windows as a guest

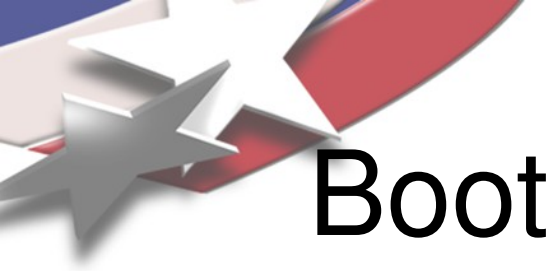
- Windows is a lot harder to boot than Linux
- Eventually booted Windows XP and 7
- 200 VMs per host for Win 7, 230 per host for Win XP
- Over 65,000 total Windows instances

Windows 7



Slimming down Windows

- Stock Windows 7
  - 10 GB disk
  - 1 GB RAM
  - Fancy graphics effects
  - 45 second boot
- Stripped down Windows 7 Embedded
  - 600 MB disk (shared)
  - 128 MB RAM
  - Boots to desktop in a few seconds



Booting 100,000x Windows 7

- Boot 520 Linux hosts (approx 2 minutes)
- Push disk images to hosts, total ~0.5 TB (pushed via GProc, 2-4 minutes)
- Run launcher to start VM boot process
  - Start a batch of guests in paused state
  - Run KSM
  - Repeat those steps until physical memory is full
  - Start and configure virtual network
  - Unpause VMs
  - Virtual machines run the experiment

MegaDroid

- Boot lots of Android VMs
- Initial result: booted 300,000 VMs on KANE
- Android's implementation works well with KSM
- Intention: create a testbed for Android devices at a city scale
- Needs more information, such as GPS location, nearby wifi, cell towers in range, etc.
  - Some of these are implemented
- Sample use case: run popular location-based software on simulated devices as they “walk” around the city

Other current work

- We're iterating on the VM/network configuration and launching tool
  - Our aim is to simplify the process of setting up, running, and tearing down an experiment
  - Will be able to load a network topology from an Opnet model (other options also available)
- Working on applications for the Mega platforms

Team members

Elisha Choe 08965
Ken Chiang 08966
Casey Deccio 08966
David Evensky 08966
John Floren 08961
David Fritz 08966
Michael Karres 08966
Levi Lloyd 08966
Ron Minnich (ex-08961)
Don Rudish (ex-08961)
Andrew Sweeney (ex-08965)
Keith Vanderveen 08961
Jamie Van Randwyk (ex-08966)

Questions?