KVM on s390: what's next?
Agenda

- Current status
- Exploring the limits of our kvm port with the flower shop scenario
- Next steps
Current status

- Kernel components upstream in 2.6.26
- Intermediate userspace “kuli”
- Kuli is not a supported customer scenario
- Features:
  - Very low intercept rate and performance overhead
  - VirtIO block, console and network
  - No channel subsystem
  - Up to 64 virtual cpus per guest
  - Nested page tables, guest and host demand paging
  - CPU timer, and vtimers clock cycle granularity
  - Clock cycle granularity time accounting (usr,sys,idle,wait,steal,guest)
  - Can run on z/VM and LPAR, on all 64bit machines
VirtI/O on s390

- **Cannot use virtio_pci**
- **Transport similar to lguest**
  - Synchronous disk I/O
  - Network connection only via TAP
  - Only ~80 devices per guest
  - No hotplug
  - Very stable, but needs functional improvement
- **Issue with virtio_console**
  - Based on hvc_console which uses request_irq/free_irq
  - Split notification method for hvc_console, work in progress
The flower shop scenario

- **200 Linux images hosted inside a single KVM host**
- **Guests:**
  - 2 CPUs each, tested up to 64 CPUs each
  - 640 Mbytes memory each
  - IBM WebSphere application server, with PlantsByWebsphere Demo
- **Host:**
  - Logical partition (LPAR) on System z9 enterprise class
  - 12 shared CPUs @1.7 Ghz (out of 54 total)
  - 44 Gbytes of memory (out of 256 total)
  - 200 Gbytes swap
## Hardware Management Console

**Select** | **Name** | **Status** | **Activation Profile** | **Last Used Profile** | **OS Name** | **OS Type** | **OS Level**
---|---|---|---|---|---|---|---
T63LP29 | Operating | T63LP39 | | | BCET6329 | zVM | 5.3.0 - 0701
T63LP30 | Operating | T63LP30 | | | BCET6330 | zVM | 5.3.0 - 0702
T63LP31 | Exceptions | T63LP31 | | | | | |
T63LP32 | Not Operating | T63LP32 | T63LP32 | | | | |
T63LP33 | Operating | T63LP33 | | | | | |
T63LP34 | Operating | T63LP34 | | | CECFINTRA | | |
T63LP35 | Operating | T63LP35 | | | | | 00000000002061a
T63LP36 | Operating | T63LP36 | | | | | |
T63LP37 | Operating | T63LP37 | | | | | |
T63LP38 | Operating | T63LP38 | | | | | |
T63LP39 | Operating | T63LP39 | | | | | |

**Tasks: T63**

- CPC Details
- Toggle Lock
- Daily
- Recovery
- Service
- Change Management
- Remote Customization
- Operational Customization
- Configuration

**Status: Exceptions and Messages**

- Exception
- Information
- Warning
- Critical

**Transferring data from lnxhmc5...**
## KVM on s390

### Current Status

<table>
<thead>
<tr>
<th>Name</th>
<th>conn</th>
<th>cpu</th>
<th>disk</th>
<th>files</th>
<th>info</th>
<th>memory</th>
<th>msgs</th>
<th>ports</th>
<th>proc</th>
<th>ssh</th>
<th>trends</th>
</tr>
</thead>
<tbody>
<tr>
<td>kvm1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kvm30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Done
PLANTS BY WEBSHARE

Gardens of Summer
They all start with the right flowers...
and we've got them all

Tips
Preserve extra grass seed by keeping it dry. Tape boxes and bags closed, or seal them into plastic bags. Be sure to remove extra air from the bags. Store all seed in a cool, dry area such as a garage or basement.

Specials
- Bonsai Tree
  - $30.00 each
- Red Delicious Strawberries
  - $3.50 (50 seeds)
- Tulips
  - $17.00 (10 bulbs)
The image shows a terminal window displaying the output of the `top` command, which is a system monitoring utility in Unix-like operating systems. The terminal output includes system statistics such as CPU usage, memory usage, process ID (PID), user, priority level, and memory usage details. The `top` command provides real-time information about system processes and resource usage. The displayed output includes:

- **System Information**:
  - `top - 18:14:52 up 5 days 3:20 3 users load average: 44.94, 15.19, 6.53` indicating the system has been running for 5 days and 3 hours with 3 users active.
  - `Tasks: 446 total 1 running 445 sleeping 0 stopped 0 zombie` showing the number of system processes.

- **CPU Usage**:
  - CPU usage breakdown for each core (Cpu0, Cpu1, etc.), including percentages for user, system, idle, and other states.

- **Memory Usage**:
  - Total memory and swap space usage.

- **Process Details**:
  - Table with columns for PID, USER, PR, NI, VIRT, RES, SHR, S, %CPU, %MEM, TIME+ and COMMAND, detailing individual process statistics and resource usage.

This terminal output is typically used to monitor system performance, troubleshoot issues, and manage system resources.
Befehlsfenster - Konsole <3>

Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe

Every 1,0s: kmvert once | sort -n -r

7047963: exit_instruction
6553500: instruction_sigp_sense
739409: exit_wait_state
384833: exit_null
207935: deliver_virtio_interrupt
192563: instruction_sigp_emergency
192302: deliver_emergency_signal
186115: userspace_handled
108733: diagnose_44
100408: exit_external_request
6450: instruction_stsi
300: deliver_program_interruption
200: instruction_stfl
200: instruction_stap
150: instruction_stidp
100: instruction_spx
100: instruction_chsc
50: instruction_stscl
50: instruction_sigp_set_prefix
50: instruction_sigp_set_arch
50: instruction_sigp_restart
50: exit_validity
50: deliver_service_signal
Exploring the limits of our kvm port

- **Very brave behavior with little overcommitment [33xCPU/ 3xmem]:**
  - While compute intensive: >98% guest time, <2% user+system
  - I/O implementation causes significant overhead: <10% user+system
  - Fluid and responsive

- **Runs into issues with**
  - A lot of virtual cpus per guest
  - Extended memory overcommitment in the host
  - Without compat_sched_yield
The stop_machine_run issue

- **Scenario:**
  - Guests have 64 vcpus, host has only 12 vcpus to back that

- **Stop_machine_run does cpu_relax() loops on vcpus to wait for other vcpus**

- Circumvention by diagnose 0x44: yield() will schedule a different vcpu

- A storm of context switches with yield(), even with compat_sched_yield

- Rusty currently rewrites stop_machine_run to become more virtualization friendly
The memory overcommitment issue

- Scenario:
  - Guests start up, and utilize their memory, which exceeds the host memory size in total (200* 640MB = 128 GB versus 44 GB)
- one third of the memory is in inactive list, all dirty + anonymous
- vmscan starts writeback of dirty pages
- When the request queues of the swap disks runs full, pdflush cannot write back anymore (get_request_wait)
Flower shop scenario conclusion

- KVM on s390 runs stable
- No scalability issues in the KVM module
- The process scheduler in Linux is well suited for scheduling guest workload
- Core memory management has issues when handling a lot of anonymous memory
  - Track dirty pages and start writeback early?
  - Skip second chance pass on the inactive list if pdflush runs into the I/O limit?
  - Rick van Riel's optimizations?
Next steps

- Merge into the common KVM userspace
- Pseudo page fault interrupt
- Diagnose 0x10 “release pages” for ballooning
- Retrieve dirty pages log for migration
- Gdb stub
- Z90crypt virtualization over virtio
- Device passthrough for channel I/O
Questions?