IBM Linux technology center

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

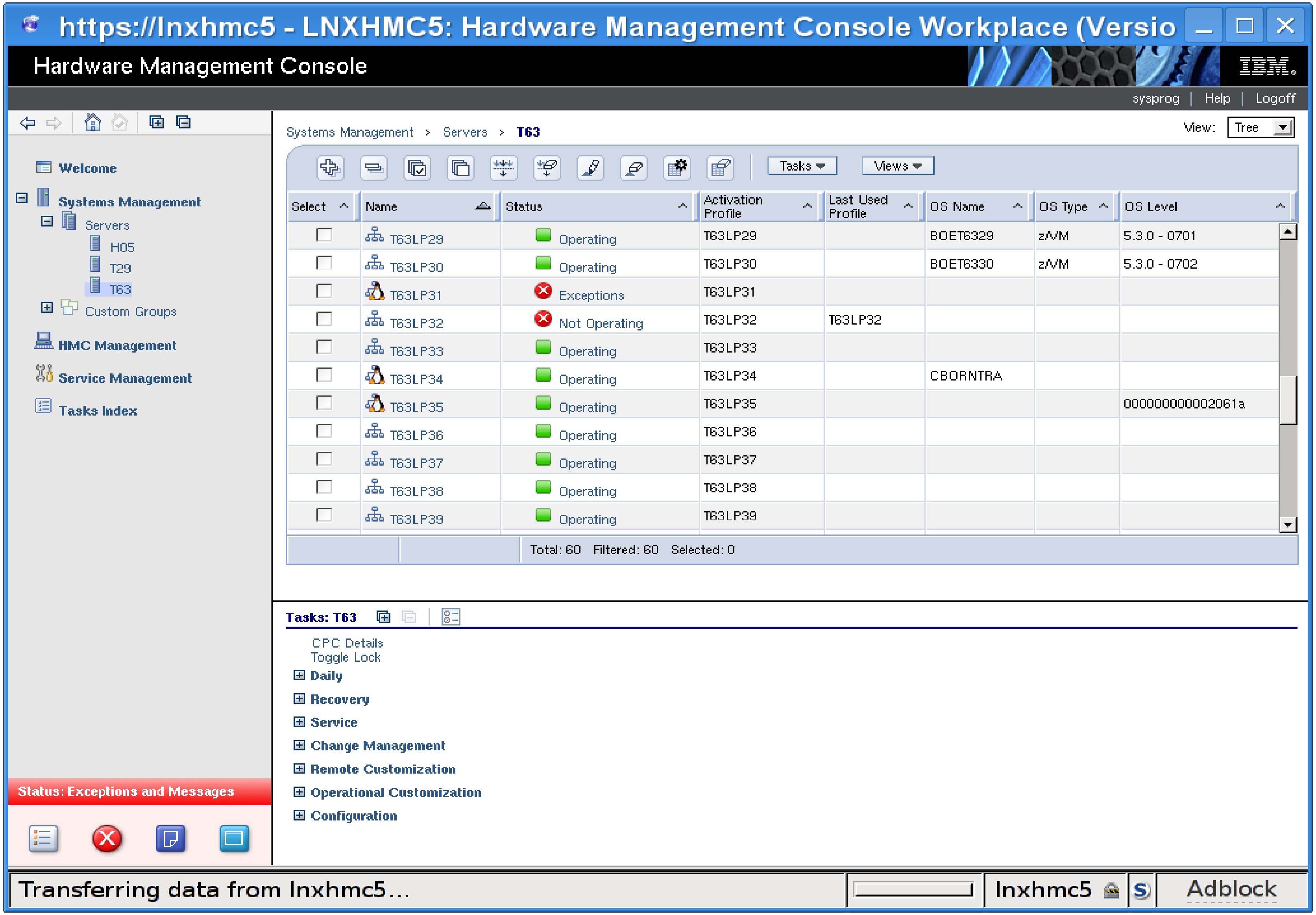
- **Kernel components upstream in 2.6.26**
- **Intermediate userspace "kuli"**
- **Kuli is not a supported customer scenario**
- **Features:**
  - Very low intercept rate and performance overhead
  - VirtIO block, console and network
  - No channel subsystem
  - Up to 64 virtual cpus per guest
  - Nested 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



https://lnxhmc5 - LNXHMC5: Hardware Management Console Workplace (Versio

Hardware Management Console

Systems Management > Servers > T63

View: Tree

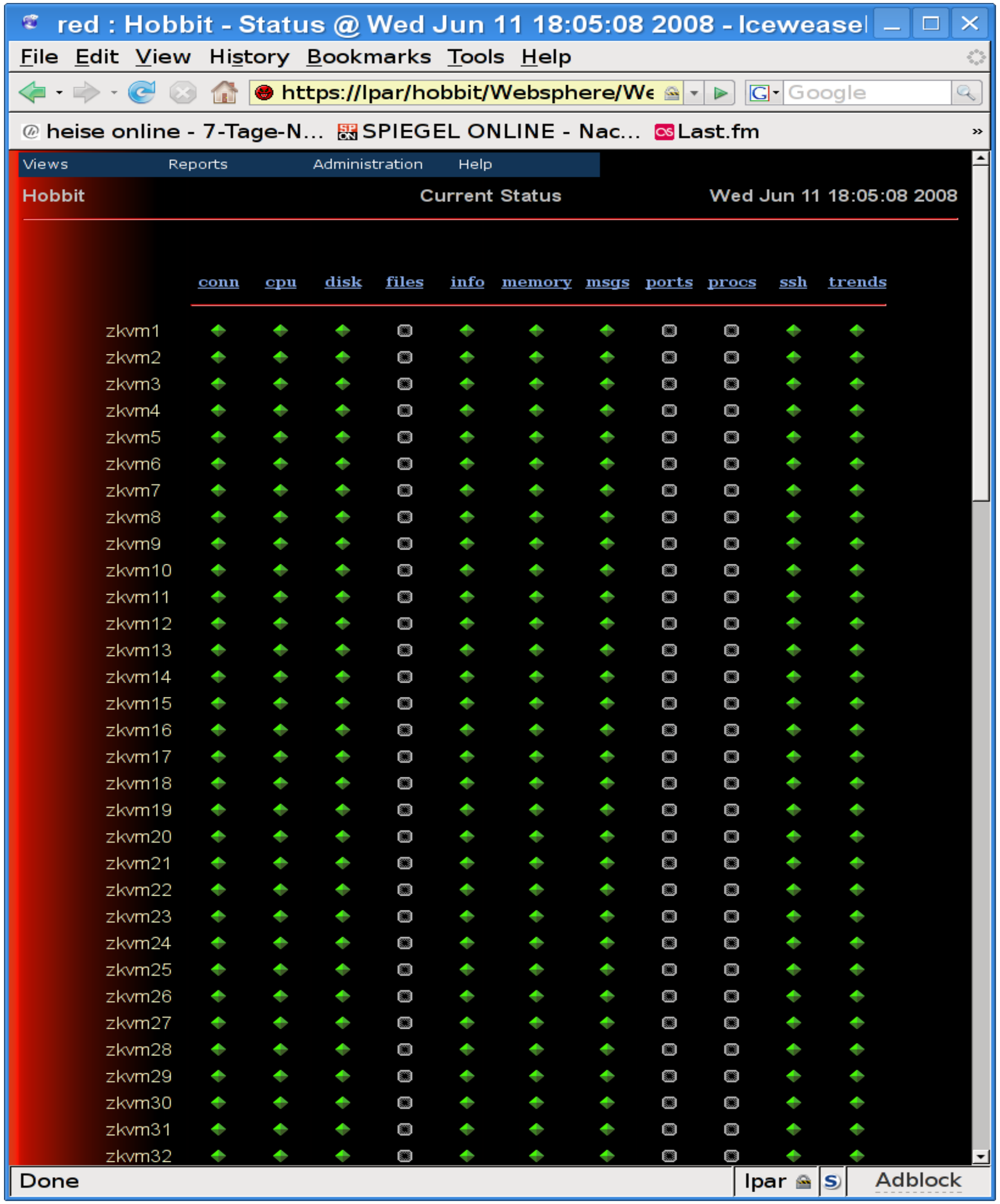
Select	Name	Status	Activation Profile	Last Used Profile	OS Name	OS Type	OS Level
□	T63LP29	🟩 Operating	T63LP29		BOET6329	z/VM	5.3.0 - 0701
□	T63LP30	🟩 Operating	T63LP30		BOET6330	z/VM	5.3.0 - 0702
□	T63LP31	⊗ Exceptions	T63LP31				
□	T63LP32	⊗ Not Operating	T63LP32	T63LP32			
□	T63LP33	🟩 Operating	T63LP33				
□	T63LP34	🟩 Operating	T63LP34		CBORNTRA		
□	T63LP35	🟩 Operating	T63LP35				000000000002061a
□	T63LP36	🟩 Operating	T63LP36				
□	T63LP37	🟩 Operating	T63LP37				
□	T63LP38	🟩 Operating	T63LP38				
□	T63LP39	🟩 Operating	T63LP39				

Total: 60 Filtered: 60 Selected: 0

Tasks: T63

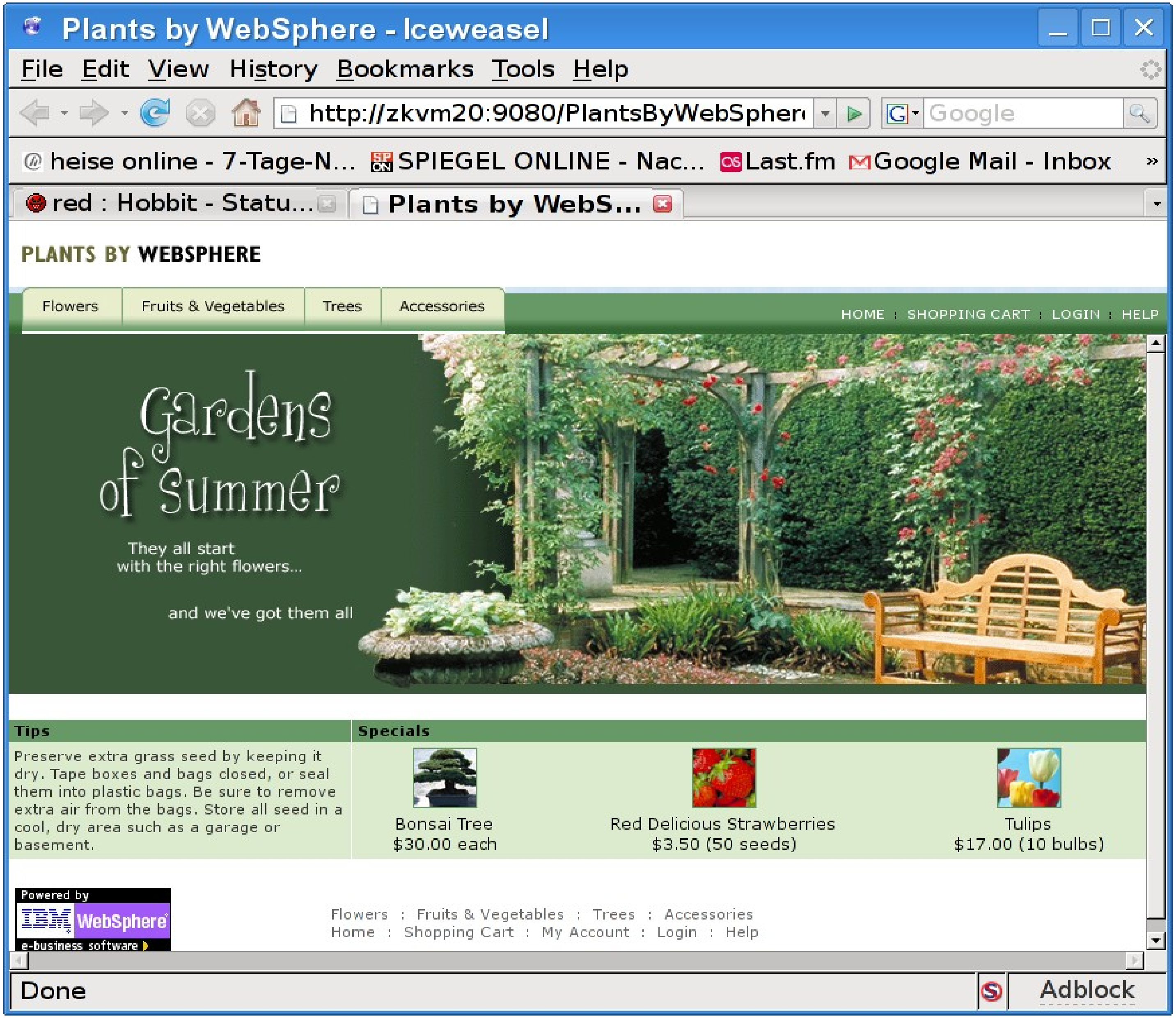
CPC Details
Toggle Lock
⊞ Daily
⊞ Recovery
⊞ Service
⊞ Change Management
⊞ Remote Customization
⊞ Operational Customization
⊞ Configuration

Transferring data from lnxhmc5...


red : Hobbit - Status @ Wed Jun 11 18:05:08 2008 - Iceweasel
File Edit View History Bookmarks Tools Help
https://lpar/hobbit/Websphere/We
Google
heise online - 7-Tage-N... SPIEGEL ONLINE - Nac... Last.fm
Views Reports Administration Help
Done
lpar
Adblock

Hobbit Current Status Wed Jun 11 18:05:08 2008

	conn	cpu	disk	files	info	memory	msgs	ports	procs	ssh	trends
zkvm1	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm2	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm3	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm4	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm5	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm6	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm7	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm8	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm9	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm10	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm11	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm12	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm13	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm14	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm15	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm16	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm17	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm18	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm19	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm20	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm21	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm22	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm23	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm24	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm25	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm26	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm27	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm28	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm29	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm30	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm31	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆
zkvm32	◆	◆	◆	○	◆	◆	◆	○	○	◆	◆


Plants by WebSphere - Iceweasel
File Edit View History Bookmarks Tools Help
http://zkvm20:9080/PlantsByWebSpher
Google
heise online - 7-Tage-N...
SPIEGEL ONLINE - Nac...
Last.fm
Google Mail - Inbox
red : Hobbit - Statu...
Plants by WebS...
PLANTS BY WEBSPHERE
Flowers
Fruits & Vegetables
Trees
Accessories
HOME : SHOPPING CART : LOGIN : HELP
Gardens of Summer
Powered by IBM WebSphere e-business software
Flowers : Fruits & Vegetables : Trees : Accessories
Home : Shopping Cart : My Account : Login : Help
Done
Adblock

They all start
with the right flowers...

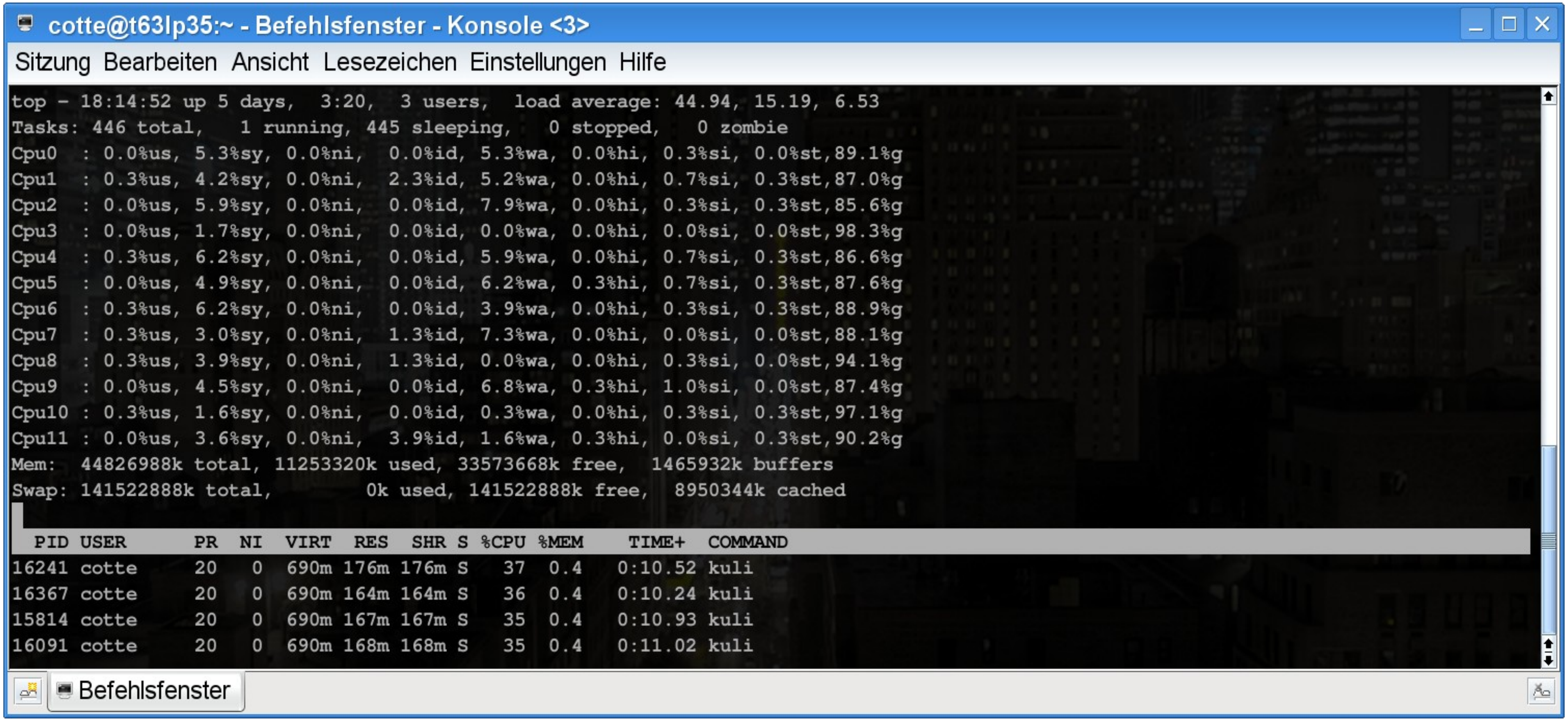
and we've got them all

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.

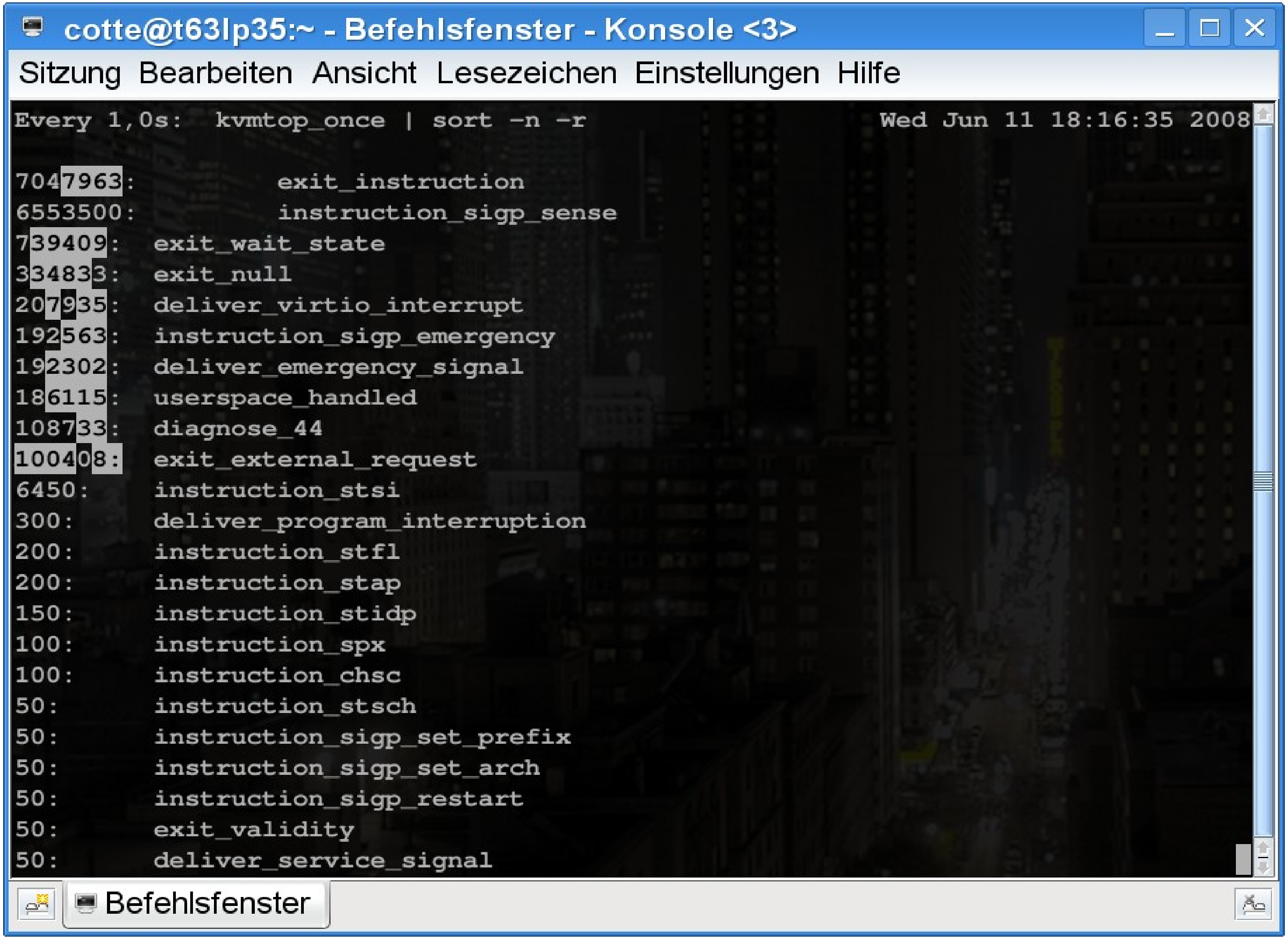
Specials

Bonsai Tree $30.00 each	Red Delicious Strawberries $3.50 (50 seeds)	Tulips $17.00 (10 bulbs)


cotte@t63lp35:~ - Befehlsfenster - Konsole <3>
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
Befehlsfenster

```
top - 18:14:52 up 5 days,  3:20,  3 users,  load average: 44.94, 15.19, 6.53
Tasks: 446 total,   1 running, 445 sleeping,   0 stopped,   0 zombie
Cpu0  : 0.0%us, 5.3%sy, 0.0%ni,  0.0%id, 5.3%wa, 0.0%hi, 0.3%si, 0.0%st,89.1%g
Cpu1  : 0.3%us, 4.2%sy, 0.0%ni,  2.3%id, 5.2%wa, 0.0%hi, 0.7%si, 0.3%st,87.0%g
Cpu2  : 0.0%us, 5.9%sy, 0.0%ni,  0.0%id, 7.9%wa, 0.0%hi, 0.3%si, 0.3%st,85.6%g
Cpu3  : 0.0%us, 1.7%sy, 0.0%ni,  0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st,98.3%g
Cpu4  : 0.3%us, 6.2%sy, 0.0%ni,  0.0%id, 5.9%wa, 0.0%hi, 0.7%si, 0.3%st,86.6%g
Cpu5  : 0.0%us, 4.9%sy, 0.0%ni,  0.0%id, 6.2%wa, 0.3%hi, 0.7%si, 0.3%st,87.6%g
Cpu6  : 0.3%us, 6.2%sy, 0.0%ni,  0.0%id, 3.9%wa, 0.0%hi, 0.3%si, 0.3%st,88.9%g
Cpu7  : 0.3%us, 3.0%sy, 0.0%ni,  1.3%id, 7.3%wa, 0.0%hi, 0.0%si, 0.0%st,88.1%g
Cpu8  : 0.3%us, 3.9%sy, 0.0%ni,  1.3%id, 0.0%wa, 0.0%hi, 0.3%si, 0.0%st,94.1%g
Cpu9  : 0.0%us, 4.5%sy, 0.0%ni,  0.0%id, 6.8%wa, 0.3%hi, 1.0%si, 0.0%st,87.4%g
Cpu10 : 0.3%us, 1.6%sy, 0.0%ni,  0.0%id, 0.3%wa, 0.0%hi, 0.3%si, 0.3%st,97.1%g
Cpu11 : 0.0%us, 3.6%sy, 0.0%ni,  3.9%id, 1.6%wa, 0.3%hi, 0.0%si, 0.3%st,90.2%g
Mem:  44826988k total, 11253320k used, 33573668k free,  1465932k buffers
Swap: 141522888k total,        0k used, 141522888k free,  8950344k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
16241 cotte     20   0  690m 176m 176m S   37  0.4   0:10.52 kuli
16367 cotte     20   0  690m 164m 164m S   36  0.4   0:10.24 kuli
15814 cotte     20   0  690m 167m 167m S   35  0.4   0:10.93 kuli
16091 cotte     20   0  690m 168m 168m S   35  0.4   0:11.02 kuli
```


cotte@t63lp35:~ - Befehlsfenster - Konsole <3>
Sitzung Bearbeiten Ansicht Lesezeichen Einstellungen Hilfe
Befehlsfenster

```
Every 1,0s:  kvmtop_once | sort -n -r                        Wed Jun 11 18:16:35 2008

7047963:         exit_instruction
6553500:         instruction_sigp_sense
739409:  exit_wait_state
334833:  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_stsch
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**
- **Circumention 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 utililize their memory, which exceeds the host memory size in total (200* 640MB = 128 GB versus 44GB)
- **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?