Microsoft Azure

FreeBSD on Azure ARM64

Souradeep Chakrabarti (schakrabarti@microsoft.com)
Wei Hu (weh@microsoft.com)

Introduction

- We are from Microsoft Linux Systems Group

- Souradeep Chakrabarti:
  Worked on AIX Unix, Linux and FreeBSD for last 11 years.
  Currently working in Microsoft Linux Systems Group.
- Wei Hu:
  Worked on Solaris, VmWare ESXi for 15+ years.
  Currently also in Microsoft Linux System Group

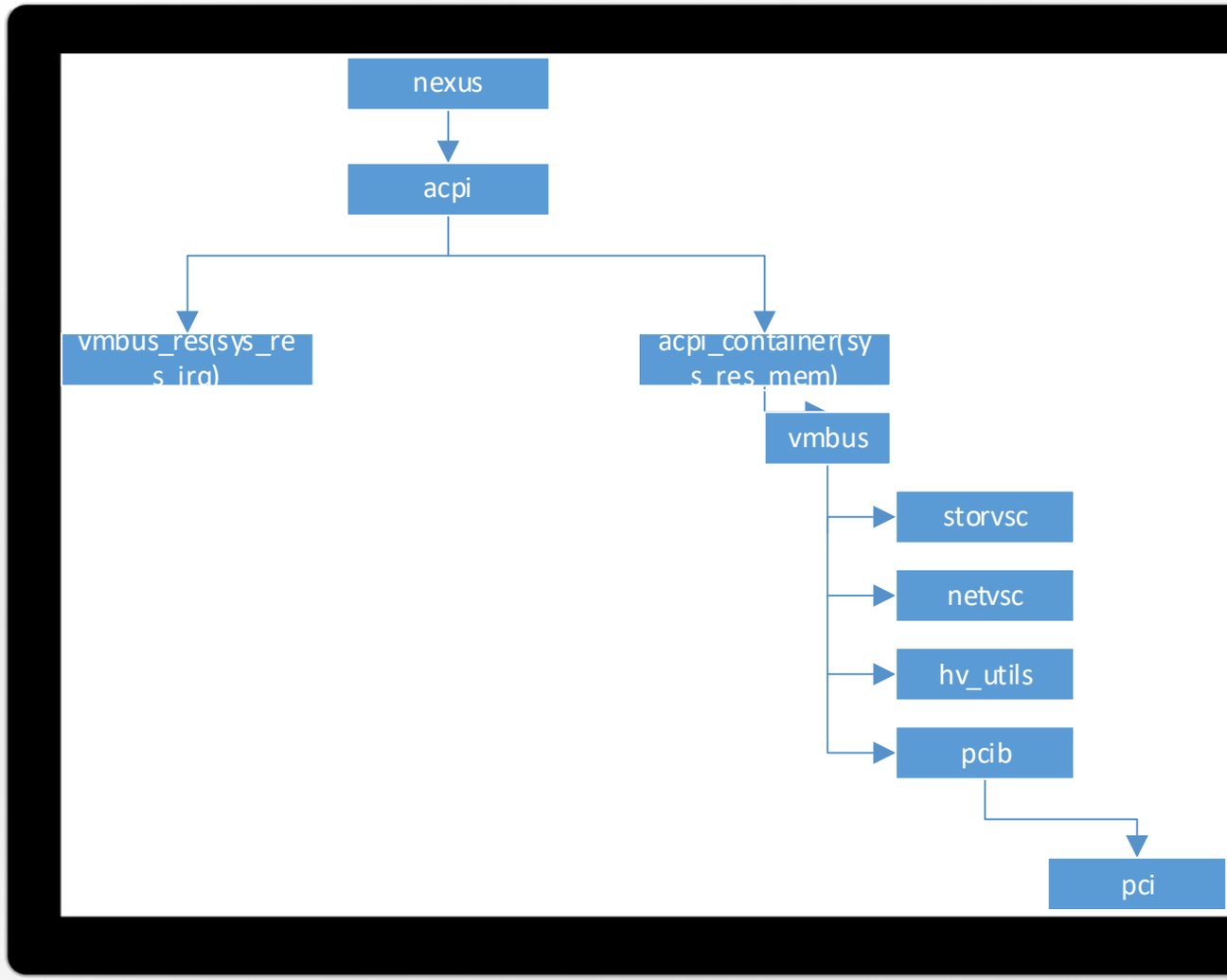
Preface

- Microsoft is currently offering Linux in ARM64 SKUs of Azure.
- FreeBSD is available for x86 SKUs in Azure.
- Major appliance vendors use FreeBSD on Azure.
- We are working on enabling FreeBSD on ARM64 SKUs of Azure.
- The following slides are on the major changes done to make it happen.

Hyper-V driver in FreeBSD

X86 Hyper-V device driver layout

- vmbus/. The parent of all Hyper-V devices. It also contains code for early initialization, i.e. before any drivers are loaded.
- vmbus/amd64/ and vmbus/i386/. Contains vmbus IDT vector entry and hypercall.
- storvsc/. Synthetic SCSI controller driver.
- netvsc/. Synthetic network controller driver.
- pcib/. PCI bridge driver for SR-IOV/pass-through.
- input/. Synthetic keyboard driver.
- utilities/. Drivers for KVP, VSS, time synchronization etc.
- include/. Shared header files; exposed by the vmbus.



New Hyper-V driver layout

- vmbus/aarch64/. Contains vmbus_aarch64.c, hyperv_reg.h, hyperv_machdep.h, hyperv_machdep.c, hyperv_aarch64.c : These files are specific for ARM64 Hyper-V.
- vmbus_aarch64.c : Contains new interrupt handler setup and teardown code.
- hyperv_aarch64.c : Contains Hyper-V identify.
- hyperv_machdep.c : Contains new hypercalls for ARM64 Hyper-V.
- hyperv_reg.h : Contains ARM64 specific synthetic MSR values.

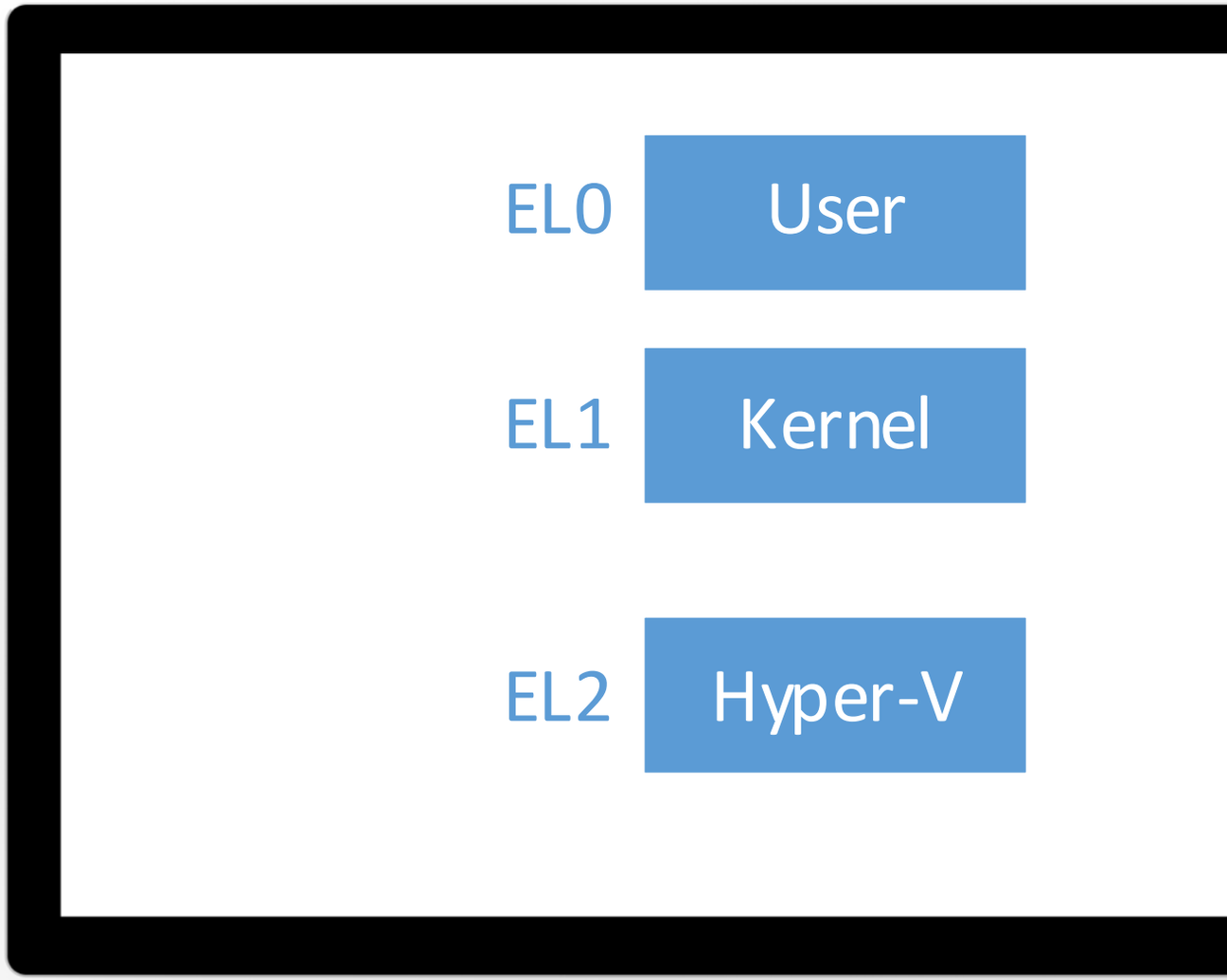
Contd

- vmbus/x86/. Contains vmbus_x86.c, hyperv_x86.c,hyperv_reg.h, hyperv_machdep.h . These are for both i386 and amd64.
- Also new file introduced hyperv_common_reg.h, which contains common synthetic MSR values for Hyper-V.
- This approach to avoid redundancy of the code.

Use of ARM SMCCC HVC

- To implement writing of MSR and reading of MSR in ARM64 HvCallSetVpRegisters hypercall and HvCallGetVpRegisters hypercall is used.
- To have the Hypercalls from El1 to El2, ARM SMCCC HVC is used
- HvCallGetVpRegisters accesses registers beyond a0 to a3. For that SMCCC 1.2 is implemented.
- Code :

sys/dev/psci/smccc_arm64.S
sys/dev/psci/smccc.h



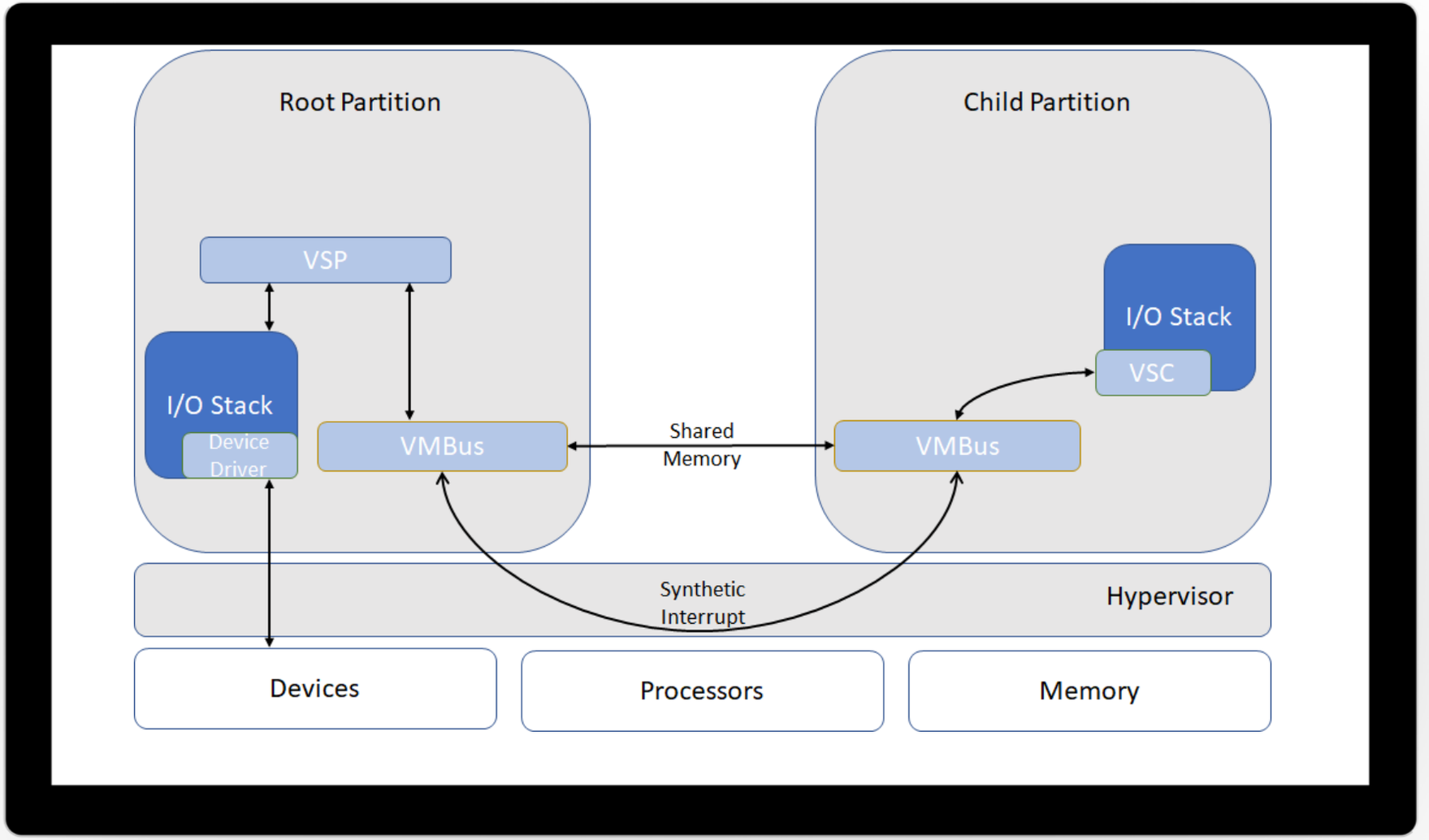
Hyper-v identify and LOAD

- Azure ARM64 hosts virtualizes the system counter and timer defined by the ARM64 architecture.
- Hyper-V synthetic timer counter initialization is not required here.
- hypercall page setup is moved from hyperv.c to x86 specific hyperv_x86.c, along with hyperv timer counter initialization.
- hyperv_et.c is also not required for ARM64, it is now for x86.
- Have used ACPI FADT to identify Hyper-V, which was done using CPUID in x86.
- Have used ARM SMCCC HVC to identify certain features of Hyper-V and to set the guest OS id.

EFI Serial console

- EFI serial console was not working and was causing hang during loading.
- Upon investigation, it was found the problem is coming from efi comconsole setAttribute().
- https://bugs.freebsd.org/bugzilla/show_bug.cgi?id=266248
- The fix of the same has been committed :
- https://cgit.freebsd.org/src/commit/?id=4b2322bba19d26f91d0f1a993798c52ebf45d41b

Hyper-V Vmbus



VMBUS Interrupt handling

x86 VMBus was using Free IDT vector for Hyper-V ISR.
In ARM64 VMBus uses Interrupt mentioned in the _CRS of the HID VMBus.

This resource is currently owned by vmbus_res as a direct child of ACPI.
To access this resource from vmbus_res,
we have used :

```
devclass_get_device(devclass_find("vmbus_res"),0)
```

Also introduced new attributes in vmbus_softc:
ires, icookie and vector.

Hyper-V acpi table

```
Name (_HID, "VMBus")  // _HID: Hardware ID
    Name (_UID, Zero)  // _UID: Unique ID
    ...
    Name (_CRS, ResourceTemplate ()  // _CRS:
Current Resource Settings
    {
      Interrupt (ResourceConsumer, Edge,
ActiveHigh, Exclusive, ,, )
      {
        0x00000012,
      }
```

Contd.

- From the successful allocated ires resource, we are getting the irq number using rman_get_virtual(), which we are using then for synthetic interrupt controller setup.
  sc->vmbus_idtvec = irq_data->irq;
- These changes are in vmbus_aarch64.c and the lapic based IDT vector setup has been moved in vmbus_x86.c

vmbus pcib

- Enabled vmbus_pcib for to use accelerated networking feature of Hyper-V in Azure.
- Hyper-V does not emulate a full-fledged PCI bridge.
- A cooperative PCI bridge driver is needed on FreeBSD.
  - Handle PCI configuration space accessing.
  - Setup BARs for SR-IOV/passed-through devices.
  - Remap MSI/MSI-X data and address.

- This is to enable SR-IOV, AN, NVME enabled for FreeBSD on ARM64 Hyper-V.

Enable SPI-MSIX mapping

- Azure HCI in ARM64 uses SPI to map MSIX, as it is not supporting ITS and LPI.
- FreeBSD ACPI did not had support for MBI ranges, and gicv3 driver only support SPIs under FDT.
- In this course of work, gic_v3_acpi_attach() has been changed to address mbi start and mbi end and to register with intr_msi_register().

CHANGE in VMBUS_PCIB

- In Azure ARM64 HCI, PCI protocol version1.4 was required to communicate with the host.
- New message structures were required to be supported by host side, for succesful VF attachment.
- Also, in x86 nexus used to take care of msix allocation, release and mapping. Those have changed here to use intr_alloc/intr_release/intr_map functions.

Current performance

- This gives a performance boost on I/O by avoiding the synthetic devices, and using Hyper-V DDA.
- As of today, the network performance of FreeBSD on Azure ARM64 per with Linux.

Current upstream changes

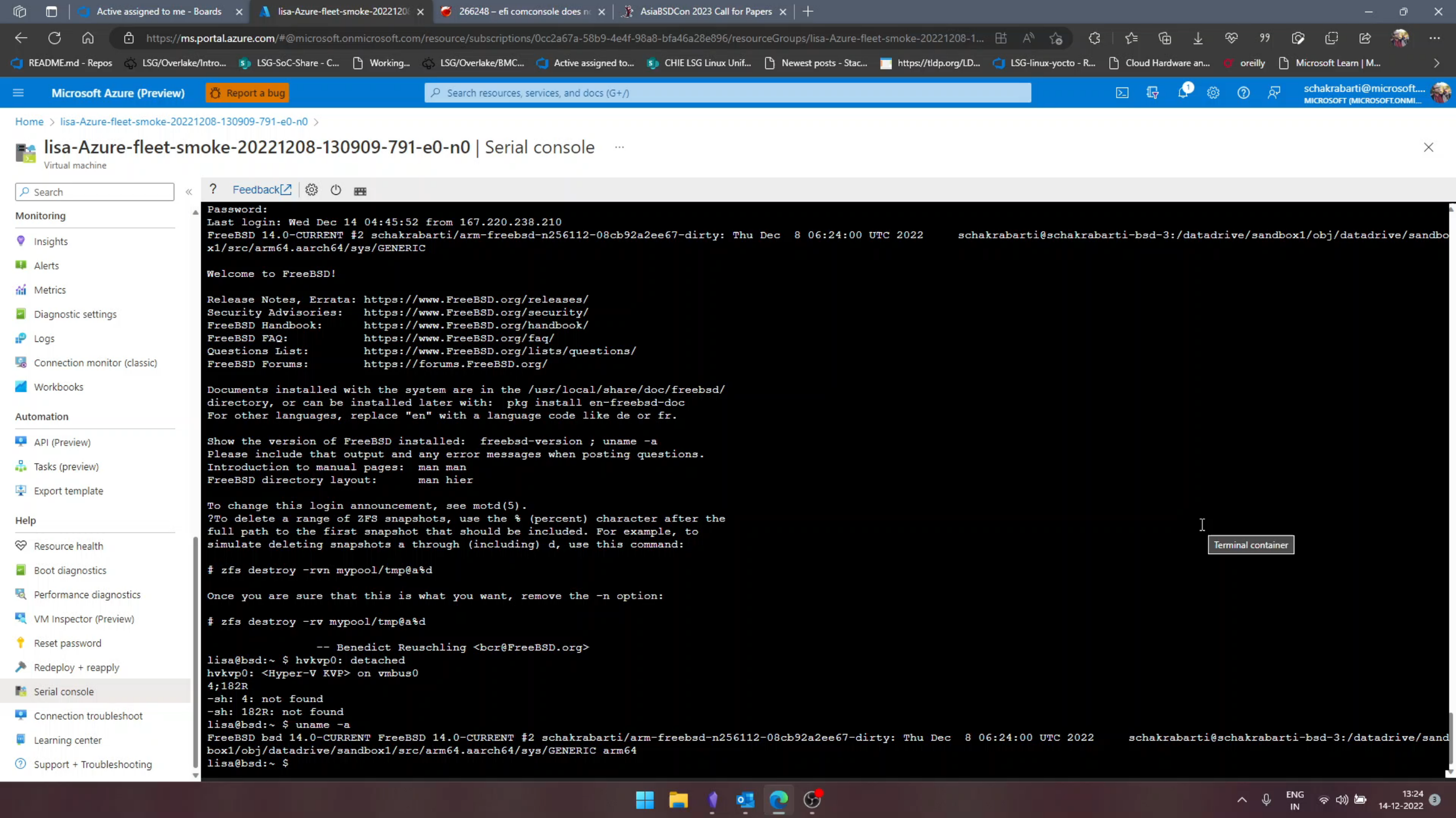
about summary refs log tree commit diff

log msg schakrabarti sea

	Commit message (Expand)	Author	Age	Files	Lines
*	arm64: Hyper-V: Add vPCI and Mellanox driver modules into build	Wei Hu	7 days	2	-1/+12
*	efiserialio: use port settings (sio->Mode) for initial setup	Toomas Soome	2023-02-03	1	-16/+22
*	arm64: Hyper-V: vPCI: Enabling v-PCI in FreeBSD in ARM64 Hyper-V	Wei Hu	2023-02-01	1	-57/+214
*	arm64: Hyper-V: vPCI: Adding Hyper-V PCI protocol 1.4	Wei Hu	2023-02-01	1	-3/+80
*	arm64: Hyper-V: vPCI: SPI MSI mapping for gic v3 acpi in arm64	Wei Hu	2023-02-01	2	-1/+24
*	arm64: Hyper-V: enablement for ARM64 in Hyper-V (Part 3, final)	Souradeep Chakrabarti	2022-10-27	20	-552/+164
*	arm64: Hyper-V: fix couple more commit errors caused by duplicated lines	Wei Hu	2022-10-24	2	-288/+0
*	arm64: Hyper-V: fix a commit error caused duplicated lines in vmbus_aarch64.c	Wei Hu	2022-10-21	1	-157/+0
*	arm64: Hyper-V: enablement for ARM64 in Hyper-V (Part 2)	Souradeep Chakrabarti	2022-10-21	4	-0/+670
*	arm64: Hyper-V: vmbus: use the IRQ resource from vmbus_res	Souradeep Chakrabarti	2022-10-21	1	-1/+3
*	arm64: enablement for ARM64 in Hyper-V (Part 1)	Souradeep Chakrabarti	2022-09-29	5	-0/+1052
*	arm64: Enabling new hypercalls using HvCallSetVpRegisters and HvCallGetVpRegi...	Wei Hu	2022-09-26	2	-0/+60
*	Hyper-V: storvsc: Call bus_dmamap_sync() for dma operations	Wei Hu	2022-08-15	1	-0/+25

What is next...

- During provisioning on Azure, following issues were seen intermittently:
  - panic: ram_attach: resource 7 failed to attach.
  - The VM boots up fine for the first time. But the second boot with 'reboot' command in guest caused either panic or filesystem inconsistency error.
  - VM boot hangs with certain error in CAM layer when the VM has more than 4 synthetic nics
  - VM boots up fine with ZFS root filesystem. But at certain stage it panic in certain ZFS routines.
  - UFS checksum error when the data disk is UFS


Active assigned to me - Boards
lisa-Azure-fleet-smoke-2022120
266248 – efi comconsole does n
AsiaBSDCon 2023 Call for Papers
https://ms.portal.azure.com/#@microsoft.onmicrosoft.com/resource/subscriptions/0cc2a67a-58b9-4e4f-98a8-bfa46a28e896/resourceGroups/lisa-Azure-fleet-smoke-20221208-1...
README.md - Repos
LSG/Overlake/Intro...
LSG-SoC-Share - C...
Working...
LSG/Overlake/BMC...
Active assigned to...
CHIE LSG Linux Unif...
Newest posts - Stac...
https://tldp.org/LD...
LSG-linux-yocto - R...
Cloud Hardware an...
oreilly
Microsoft Learn | M...
Microsoft Azure (Preview)
Report a bug
Search resources, services, and docs (G+/)
schakrabarti@microsoft....
MICROSOFT (MICROSOFT.ONMI...
Home
lisa-Azure-fleet-smoke-20221208-130909-791-e0-n0
lisa-Azure-fleet-smoke-20221208-130909-791-e0-n0 | Serial console
Virtual machine
Search
Monitoring
Insights
Alerts
Metrics
Diagnostic settings
Logs
Connection monitor (classic)
Workbooks
Automation
API (Preview)
Tasks (preview)
Export template
Help
Resource health
Boot diagnostics
Performance diagnostics
VM Inspector (Preview)
Reset password
Redeploy + reapply
Serial console
Connection troubleshoot
Learning center
Support + Troubleshooting
Feedback
Password:
Last login: Wed Dec 14 04:45:52 from 167.220.238.210
FreeBSD 14.0-CURRENT #2 schakrabarti/arm-freebsd-n256112-08cb92a2ee67-dirty: Thu Dec 8 06:24:00 UTC 2022 schakrabarti@schakrabarti-bsd-3:/datadrive/sandbox1/obj/datadrive/sandbox1/src/arm64.aarch64/sys/GENERIC
Welcome to FreeBSD!
Release Notes, Errata: https://www.FreeBSD.org/releases/
Security Advisories: https://www.FreeBSD.org/security/
FreeBSD Handbook: https://www.FreeBSD.org/handbook/
FreeBSD FAQ: https://www.FreeBSD.org/faq/
Questions List: https://www.FreeBSD.org/lists/questions/
FreeBSD Forums: https://forums.FreeBSD.org/
Documents installed with the system are in the /usr/local/share/doc/freebsd/
directory, or can be installed later with: pkg install en-freebsd-doc
For other languages, replace "en" with a language code like de or fr.
Show the version of FreeBSD installed: freebsd-version ; uname -a
Please include that output and any error messages when posting questions.
Introduction to manual pages: man man
FreeBSD directory layout: man hier
To change this login announcement, see motd(5).
?To delete a range of ZFS snapshots, use the % (percent) character after the
full path to the first snapshot that should be included. For example, to
simulate deleting snapshots a through (including) d, use this command:
zfs destroy -rvn mypool/tmp@a%d
Once you are sure that this is what you want, remove the -n option:
zfs destroy -rv mypool/tmp@a%d
-- Benedict Reuschling <bcr@FreeBSD.org>
lisa@bsd:~ $ hvkvp0: detached
hvkvp0: <Hyper-V KVP> on vmbus0
4;182R
-sh: 4: not found
-sh: 182R: not found
lisa@bsd:~ $ uname -a
FreeBSD bsd 14.0-CURRENT FreeBSD 14.0-CURRENT #2 schakrabarti/arm-freebsd-n256112-08cb92a2ee67-dirty: Thu Dec 8 06:24:00 UTC 2022 schakrabarti@schakrabarti-bsd-3:/datadrive/sandbox1/obj/datadrive/sandbox1/src/arm64.aarch64/sys/GENERIC arm64
lisa@bsd:~ $
Terminal container
ENG
IN
13:24
14-12-2022

\\.\pipe\Testpipe - PuTTY
schakrabart...
lqqqqqqqqqquCompletetqqqqqqqqqqqqqk
x Installation of FreeBSD complete! x
x Would you like to reboot into the x
x installed system now?   x
tqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqu
x [Reboot] [Shutdown] [Live CD] x
mqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqqj
Shutdown NOW!
shutdown: [pid 1459]
System shutdown time has arrived
Dec 14 07:59:18 shutdown[1459]: power-down by root:
Dec 14 07:59:18 syslogd: exiting on signal 15
Waiting (max 60 seconds) for system process `vnlru' to stop... done
Waiting (max 60 seconds) for system process `syncer' to stop...
Syncing disks, vnodes remaining... 0 0
ENG
IN
11:59 PM
12/13/2022

```
    inet6 fe80::20d:3aff:fe6c:c061/64 scope link
       valid_lft forever preferred_lft forever
3: enP10554s1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdis
c mq master eth0 state UP group default qlen 1000
    link/ether 00:0d:3a:6c:c0:61 brd ff:ff:ff:ff:ff:ff
    altname enP10554p0s2
lisa@Ubuntu-x86-scha:~$ iperf -s
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size:  128 KByte (default)
------------------------------------------------------------
^Clisa@Ubuntu-x86-scha:~$ iperf -s
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size:  128 KByte (default)
------------------------------------------------------------
[  4] local 10.0.0.5 port 5001 connected with 10.0.0.4 port 12299 (p
eer 2.1.8)
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  13.0 GBytes  11.1 Gbits/sec
^Clisa@Ubuntu-x86-scha:~$ iperf -s
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size:  128 KByte (default)
------------------------------------------------------------
[  4] local 10.0.0.5 port 5001 connected with 10.0.0.6 port 37510 (p
eer 2.1.5)
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  12.9 GBytes  11.1 Gbits/sec
lisa@Ubuntu-x86-scha:~$
lisa@Ubuntu-x86-scha:~$
lisa@Ubuntu-x86-scha:~$ uname -a
Linux Ubuntu-x86-scha 5.15.0-1023-azure #29~20.04.1-Ubuntu SMP Wed Oct 26 19:18:25 UTC 20
22 x86_64 x86_64 x86_64 GNU/Linux
lisa@Ubuntu-x86-scha:~$ iperf -s
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size:  128 KByte (default)
------------------------------------------------------------
[  4] local 10.0.0.5 port 5001 connected with 10.0.0.6 port 40456 (peer 2.1.5)
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  13.9 GBytes  11.9 Gbits/sec
[  4] local 10.0.0.5 port 5001 connected with 10.0.0.4 port 47829 (peer 2.1.8)
[  4]  0.0-10.0 sec  12.8 GBytes  11.0 Gbits/sec
[  4] local 10.0.0.5 port 5001 connected with 10.0.0.4 port 54636 (peer 2.1.8)
[  4]  0.0-10.0 sec  12.8 GBytes  11.0 Gbits/sec
lisa@Ubuntu-x86-scha:~$
```

```
    acpi_ged0
    cpu0
    cpu1
    cpu2
    cpu3
    cpu4
    cpu5
    cpu6
    cpu7
    psci0
    gic0
    generic_timer0
    pmu0
  efirtc0
  armv8crypto0
  cryptosoft0
lisa@bsd:~ $ pciinfo -l
-sh: pciinfo: not found
lisa@bsd:~ $ pciconf -l
mlx5_core0@pci1:0:2:0:  class=0x020000 rev=0x80 hdr=0x00 vendor=0x15b3 device=0x1018 subv
endor=0x15b3 subdevice=0x0080
lisa@bsd:~ $ una
```

```
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 10
00
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen
1000
    link/ether 00:0d:3a:6f:11:47 brd ff:ff:ff:ff:ff:ff
    inet 10.0.0.6/24 metric 100 brd 10.0.0.255 scope global eth0
       valid_lft forever preferred_lft forever
    inet6 fe80::20d:3aff:fe6f:1147/64 scope link
       valid_lft forever preferred_lft forever
3: enP8783s1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master eth0 state
 UP group default qlen 1000
    link/ether 00:0d:3a:6f:11:47 brd ff:ff:ff:ff:ff:ff
    altname enP8783p0s2
lisa@test-ubuntu-arm64-scha:~$ lspci
224f:00:02.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5 Virtua
l Function] (rev 80)
lisa@test-ubuntu-arm64-scha:~$
```

Microsoft Azure

Thank you.