Build Cloud Infrastructure with ARMv8

Yu-Chiang Huang (黃宇強)  Jim Huang (黃敬群)
National Chiao Tung University, Taiwan  National Cheng Kung University, Taiwan
Open Source Summit North America / Sep 13, 2017
About this presentation

- Rethink ARM server as new choice
- BYOC™ using OpenStack [build your own cloud]
- ARM/Platform-specific efforts
- Disadvantage and Further improvements
- Disclaimer: the performance data listed is only for unoptimized build of OpenStack or Linux
ARM server as new choice
Market for ARMv8-A servers
Typical ARM Server SoC
Example: Cavium OCTEON TX™ 64-bit ARM-based SOCs

- high density of cores on a chip
- Server Grade Hypervisor & Container support for Virtualization
- 400Mbps to 200Gbps of security, packet throughput
- vNIC for multiple networking software paradigms

Source: http://www.cavium.com/OCTEON-TX_ARM_Processors.html
Market: Networking

- Generally very low power so possible for high density of cores on a chip
  - Cavium delivering 96 cores
- move to Network Function Virtualization (NFV)
  - Move to deliver hardware functions as virtual appliances
  - OpenStack seen as standard platform for delivering these instances.
  - Drive to Software Defined Networking (SDN) is helping
- Benefits
  - The time NFV spends on IO always longer than on CPU
  - Highly-concurrent characteristics make NFV attractive
Market: Networking
Example: NFV PicoPod

- OPNFV cluster using 6 Marvell MACCHIA TObin boards along with a Marvell Prestera DX 14 port, 10GbE switch.
- supports OpenDataPlane (ODP), Data Plane Development Kit (DPDK), and OpenFastPath (OPF) in addition to OPNFV.

Source:
https://www.picocluster.com/products/nfv-picopod
Market: Storage

• Software defined storage runs on servers
• Object stores are very complimentary to VM use
  ○ Swift & Ceph are perfect use cases
• With the high density / low cost capabilities giving cloud scale resilience
• Server vendors taking storage specific solutions using ARM to Market
  ○ Cavium, HPE
Market: Storage
Example: Mars 200 8-node ARM-based micro server

- Less than 60 Watts power consumption (exclude HDD & SSD)
- 8 x Marvell Armada 385 Dual ARMv7 Core 1.6 GHz
- Optimized Ceph software supports object, block and share file storage
- Support OpenStack storage and S3 API

When do ARM servers fit?

• Frequently concurrent data access powered by high density of cores within limited physical environments
• Due to the scale required for certain cloud services, i.e. the number of machines allocated to them, it becomes more economically feasible to optimize the hardware to the workload instead of the other way around
Key Facilities in ARMv8-A
ARMv8.x SoC

• Common base IP blocks required for a complete system with hardware-accelerated virtualization
  ○ Standard UART, Interrupt controller (GICv3), SMMUv3…

• Standard I/O and interconnects (SATA/USB, PCIe/Fabrics)
  ○ General adoption of PCIe in all current ODM/OEM designs
  ○ Increasing interest in fabrics (see also CCIX announcement)
ARM Virtualization Extension

- CPU virtualization ⇒ new mode and a new Privilege Level.
  - Hyp mode: non-secure Privilege Level 2.
- Memory virtualization ⇒ Intermediate Physical Address (IPA)
  - Guest OS cannot access physical address directly.
- In I/O virtualization ⇒ Virtual Generic Interrupt Controller (vGIC) interface
  - to deliver interrupt in more faster way.
- support I/O virtualization via a system MMU (SMMU) similar to the IOMMU
ARMv8.x SoC in a nutshell
ARMv8 System Platform

- Server Base Boot Requirements lays the “ground rules”
- All systems boot using UEFI for platform firmware
  - Abstraction, standardized boot, and (capsule) updates
  - Runtime services for common features (time, reboot, etc.)
- All systems use ACPI for hardware device enumeration
  - Abstracted generic platform with vendor managed hooks
  - Runtime method calls hide platform “glue” from OS
Linaro Enterprise Group (LEG)

- Linaro Enterprise Group (LEG) was formed in 2012
  - Initial focus on enabling 64-bit platform standards
- Provide reference software stacks
  - Base platform (OpenPlatformPkg)
    - Software Defined Infrastructure (OpenStack)
  - BigData and other examples
    Linaro Developer Cloud
Shortcoming of ARM Server Ecosystem

• Software stack lacks of effective upgrade and enhancements because of diverse SoC and peripherals
• Inconsistent encapsulation to Linux computing nodes
• Weak and semi-open community. Unlike OpenWRT
• documentation, collaboration, field-trial
Build ARM cloud using OpenStack
A cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface.

Source: http://de.wikipedia.org/wiki/OpenStack
What we expect to deliver

• Full-functioned OpenStack environments for ARMv8-powered server
• Tweaked and verified build of OpenStack
• Share the instructions and collaborate with community
Example flow\(^{(1)}\)

User Requests a VM running SUSE Linux

Source: http://getcloudify.org/2014/07/18/openstack-wiki-open-cloud.html
Example flow (2)

The Dashboard (horizon) passes the request to the Compute Component (NOVA)
Example flow (3)

The request goes to Identity Component (keystone) for authentication.
Example flow

NOVA requests the Networking Component (Neutron) for an IP Address
After getting the image, Nova (compute component), mounts it on a VM host. During the boot process of the VM, it requests Neutron (DHCP component) for an IP address.
Case study: Deploy IOTA nodes

- IOTA: first scalable Blockchain that solves issues of Scalability, Privacy and Fees.
  - billions of transactions can be sent without any transaction fees
  - No miner centralization or Block Size debates.
- can even send encrypted Data and transact Offline.
- However, we do need more nodes
  - speed up transaction
  - make IOTA more secure
  - enable micropayment

https://iota.org/
Case study: Deploy IOTA nodes

Tamper-proof

Banks
Insurance Providers
Customs / Ports
Importers/Exporters
Shipping Liners
Other M2M
Case study: Deploy IOTA nodes

- All IOTA full nodes are powered by OpenStack running on ARMv8-A (Cavium ThunderX SoC) with virtualization
- Each instance only consumes 1GB RAM and 2 vcpu
OpenStack Compute: Nova

- Component based architecture enabling quicker additions of new features
- Built on a messaging architecture and all of its components can typically be run on several servers
  - enable communications among components through message queue
- Supports virtualization technology
  - KVM, XenServer, Linux Container (LXC)
It’s not straightforward for ARM Enablement

• Data centers have many hosts and appliances with different architectures
  ○ Pending Aarch64 patches against OpenStack
  ○ Application-specific tweaks. E.g. node-ffi for Aarch64

• One network storage appliance may have dozens of hosts.
  ○ Obvious slow-down for CPU-bound operations

• One host has multiple firmware images:
  ○ Processor FW images: ARM Trusted Firmware, UEFI
  ○ Microcodes for inter--processor link, PCIe, etc.
User: Issue firmware update command from OS

OS: Are the capsule/platform good to go?

No

OS: Copy capsule to EFI System Partition; set a special OS bootloader as BootNext; reboot

Ye

OS bootloader: Deliver user interface; call UpdateCapsule()
UEFI DXE driver: Is the capsule valid?

- No
  - UEFI DXE driver: issue SMC call to enter into secure world
  - ARM TF secure service: flash new firmware image into flash, return to non-secure world
  - UEFI DXE driver: return to OS bootloader
  - OS bootloader: restore BootNext; reboot

- Yes
ARM/Platform-specific Efforts
Significant efforts done by Linaro

- Linaro continues working on OpenStack Nova for a long time
- Critical patches of Nova enablement on AArch64
  - Set SCSI as the default disk controller on AArch64
  - Add support for libvirt virtio-mmio address type
  - Set cpu-mode to host-passthrough on AArch64
  - Set virtio-scsi as the default CDROM bus for AArch64
- Others
  - Fix Nova unit tests on AArch64
  - Fixup Magnum deployment manual
  - Pick the first available disk as configure drive
  - Fix default console type name for AArch64
  - Un-define libvirt domain with "--nvram" parameter
  - Fix deletion failure of NVRAM enabled VM

Source: http://events.linuxfoundation.org/sites/events/files/slides/OpenStack%20On%20AArch64%20%28linuxcon%29.pdf
Compute component on AArch64
Standing on the shoulders of giants

- We can benefit from Linaro’s participation on OpenStack Nova
  - Don’t have to patch OpenStack Mitaka at the first glance
- However, due to incompatible convention between libvirt and QEMU-AArch64, some functionalities, such as VNC Console, could not work out of the box
- For Ocata and above, logd in libvirt was introduced to manipulate logging service listening to QEMU, but default build on AArch64 can not satisfy required dependency, thus, broken.
  - You should be aware of similar issues as well
General Instructions for OpenStack

- Instructions to install and configure AArch64 targets
  - Follow “OpenStack Install Guide” as starting point
  - {Identity, Image, Compute} service: configured for AArch64-specific
  - Networking service install and configure
- Extra tweaks are necessary though
Platform-specific Tweaks

• Environment ⇒ database
  ○ Find the MariaDB default settings in Ubuntu 16.04
  ○ Disable all the conflict settings with OpenStack guide

• Image service ⇒ Verify operation
  ○ Download the UEFI images from Ubuntu Cloud Image
  ○ image property
  ○ Boot image via uefi --property hw_firmware_type=uefi
Platform-specific Tweaks

• Compute service
  ○ `/etc/nova/nova.conf` in all nodes need to disable vnc
  ○ Remember to install qemu-efi
  ○ Disable usb tablet with `use_usb_tablet=False`
  ○ If OpenStack claims that it can not delete the instance, find out whether nova libvirt driver contains `VIR_DOMAIN_UNDELETE_NVRAM` in `%PythonPATH%/nova/virt/libvirt/guest.py`
  ○ If so, upgrade nova-compute package to solve it.
Platform-specific Tweaks

- Take Cavium ThunderX as example
- arm-platforms repository is maintained by Arm and SoC vendor
  - Branch “gicv3-cpuif-mediated-access”
  - https://git.kernel.org/pub/scm/linux/kernel/git/maz/arm-platforms.git/log/?h=kvm-arm64/gicv3-cpuif-mediated-access
- Some Cavium Thunder CPUs suffer a problem where a KVM guest may inadvertently cause the host kernel to quit receiving interrupts.
  - arm64: Add workaround for Cavium Thunder erratum 30115
- VMs can cause interrupts to be disabled on the host CPU, resulting in hangs
  - ThunderX: soft lockup on 4.8+ kernels when running qemu-efi with vhost=on
Reference Configurations

- **Controller Node**
  - Intel Xeon E5-2670
  - 8 cores, 16GB RAM
  - Ubuntu 16.04 x86_64
  - VMware ESXi 6.0

- **Compute node**
  - GIGABYTE R150-T60 (powered by dual Cavium ThunderX CN88xx; up to 96 cores)
  - 128GB RAM
  - Ubuntu 16.04 aarch64

- **Network**
  - Edge-corE Switch AS5712-54X
  - 10Gbps Copper

- **Shared Storage for VM**
  - Synology NAS DS1515+ (RAID5)
  - Network bandwidth: dual 1Gbps LACP
ARM64 VM disks are saved in Shared Storage
Viewport of OpenStack-based ARM Cloud (simplified)

Controller Node
- DHCP Namespace qdhcp
  - tap eth0
  - Metadata Process
- Linux Bridge brq
  - Port tap
  - iptables
  - Port Interface

Compute Node
- Instance
  - tap eth0
  - Port tap
  - iptables
  - Port Interface

Physical Network

Internat
Benchmark

• Network
  ○ Run `iperf` for performing cross-instance bench
  ○ Scenario: instance-to-instance, instance-to-hypervisor, outgoing packets

• Database
  ○ Use `sysbench` for performing MySQL bench
  ○ OpenStack has MariaDB dependency, and hereby identify service is affected by SQL operations.
Benchmark: networking

- Network bench using iperf

**VM-to-VM**
- ~200Mbps

**VM-to-Hypervisor**
- ~2Gbps

**VM-to-VMwareVM**
- ~500Mbps
Benchmark: Database

- **MySQL Bench** *(sysbench 1000000 rows, 60 sec test, R/W mode)*
  - Aligned to 4 cores, 4G RAM
  - AArch64 VM instance: 119.36 per second
  - x86_64 VM instance: 439.14 per second

$ sysbench --test=oltp --oltp-table-size=1000000 --mysql-db=test \\   --mysql-user=root --mysql-password=yourrootsqlpassword prepare

$ sysbench --test=oltp --oltp-table-size=1000000 --mysql-db=test \\   --mysql-user=root --mysql-password=yourrootsqlpassword --max-time=60 \\   --oltp-read-only=off --max-requests=0 --num-threads=4 run

http://www.storagereview.com/sysbench_oltp_benchmark
On-going tasks

- User experience improvements
- Allow OpenStack operated on AArch64 SPICE Console
  - Bump packages: libvirt >= 3.0, QEMU >= 2.6, SPICE >= 0.12
  - Nova libvirt driver patch to enable GPU virtio device
  - Re-build QEMU with SPICE enabled
On-going tasks

• Fix inappropriate interactions in SPICE console
  • QEMU-AArch64 lacks of default USB bus
    ■ reconfigured via virtio-kbd and virtio-tablet
  • Choose between USB input or virtio input
  • USB input device expects an valid USB bus to attach
  • GRUB could not detect virtio input device

Further reading:
Acknowledge

• We would like to thank the following organizations for long-term support during our research
  ○ Cavium, Gigabyte, Edge-corE
  ○ Aletheia University, National Central University, National Cheng Kung University, National Chiao Tung University, National Taiwan University, and National Tsing Hua University. (in alphabet order)
Conclusions

• System researchers, who are developing the techniques and software infrastructure to support cloud computing, still have trouble obtaining low-level access to such resources nowadays.

• We aim to address this problem by providing a single testbed based on a range of heterogeneous distributed data centers for systems, applications and services.
  ○ Experimenting with the combination of ARMv8 and OpenStack allows further development for high-density computing.
Contact Information

- Yu-Chiang Huang <date@dozen.cloud>
- website: https://dozencloud.org
References

• *Three reasons to read ARM’s new whitepaper on Enterprise Virtualization*, Lip-Min Khor (2017)

• *High Performance NGINX Content Delivery Using ThunderX*, Tirias Research

• *OpenStack on AArch64*, Yibo Cai, Arm Ltd.

• OpenStack Architecture
  ○ [https://www.openstack.org/software/](https://www.openstack.org/software/)
  ○ [http://docs.openstack.org/liberty/install-guide-obs/common/get_started_conceptual_architecture.htm](http://docs.openstack.org/liberty/install-guide-obs/common/get_started_conceptual_architecture.htm)