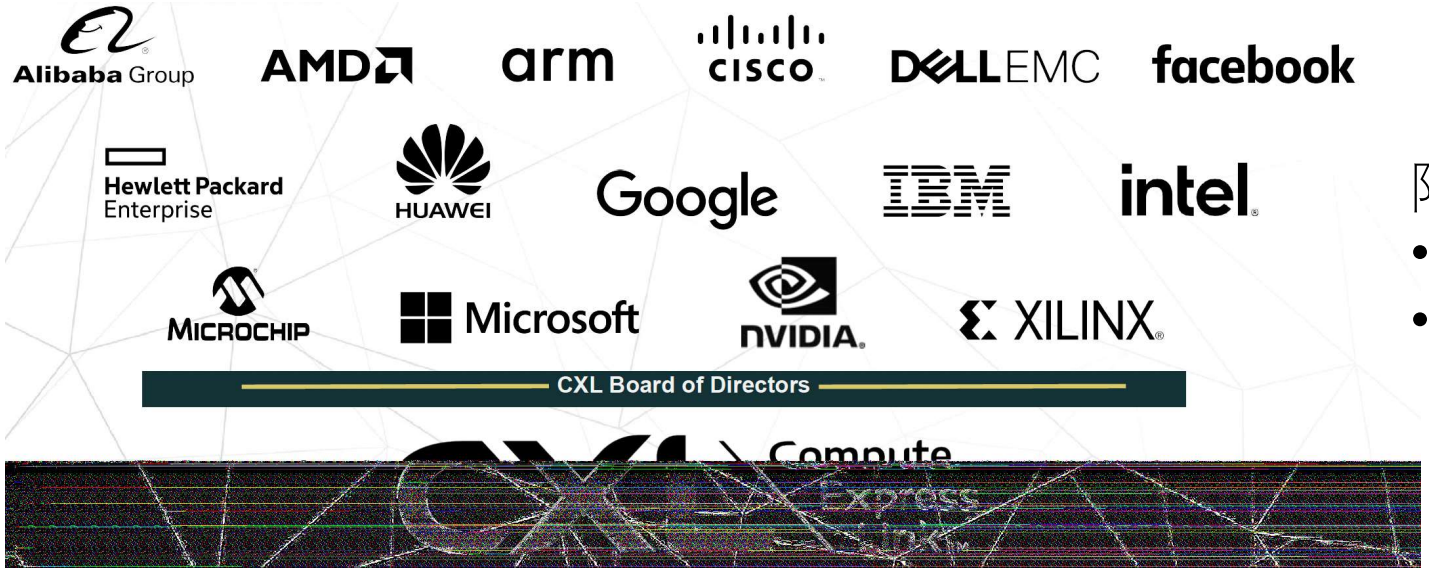
龙蜥社区 CXL 介绍

CXL背景简介

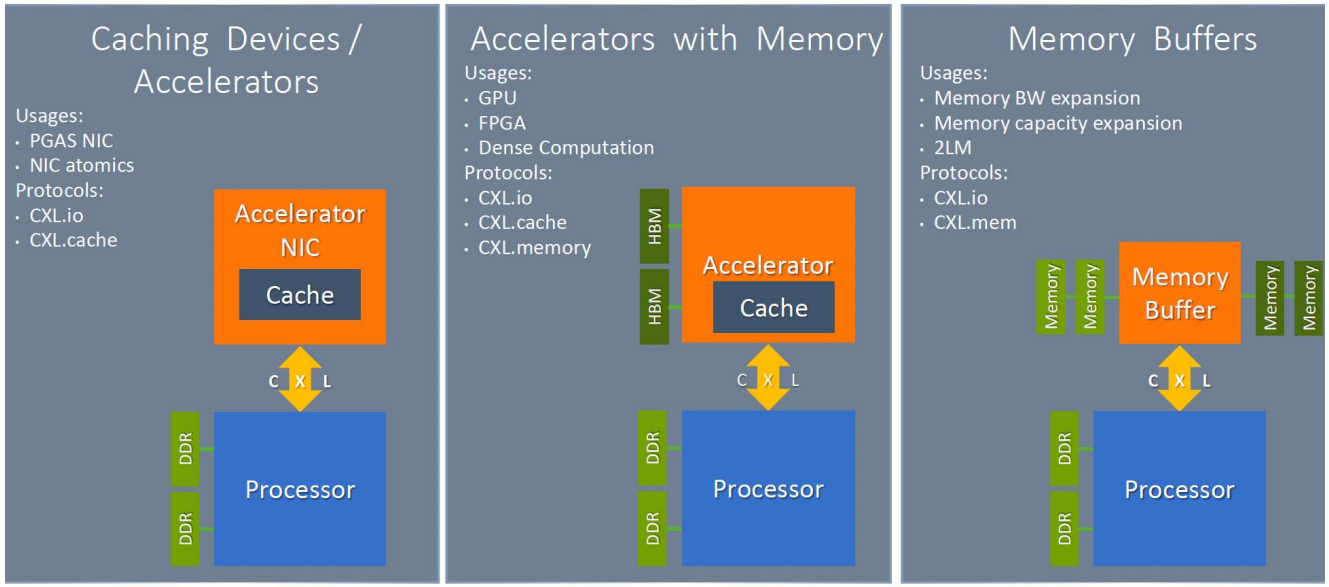


- 2019年正式成立
- 目前有224名行业成员（截止2022/10/31）
- GenZ, OpenCAPI, CCIX等互联协议都已经合并进CXL
- 针对CPU和Device之间的高速互联

阿里巴巴作为创始成员

- 深入参与标准的制定
- 在集团内部实际落地标准

CXL三种不同类型



(Type 1 Device) (Type 2 Device) (Type 3 Device)

- CXL.io: 提供CXL发现和枚举设备，设备寄存器访问，DMA，支持中断等能力
- CXL.cache: 提供CXL访问处理器内部缓存/寄存器的能力，并保证缓存一致性
- CXL.memory: 提供访问CXL上内存的能力

CXL协议规范

OpenAnolis 龙蜥社区

Features	CXL 1.0 / 1.1	CXL 2.0	CXL 3.0
Release date	2019	2020	1H 2022
Max link rate	32GTs	32GTs	64GTs
Flit 68 byte (up to 32 GTs)	✓	✓	✓
Flit 256 byte (up to 64 GTs)			✓
Type 1, Type 2 and Type 3 Devices	✓	✓	✓
Memory Pooling w/ MLDs		✓	✓
Global Persistent Flush		✓	✓
CXL IDE		✓	✓
Switching (Single-level)		✓	✓
Switching (Multi-level)			✓
Direct memory access for peer-to-peer			✓
Enhanced coherency (256 byte flit)			✓
Memory sharing (256 byte flit)			✓
Multiple Type 1/Type 2 devices per root port			✓
Fabric capabilities (256 byte flit)			✓

CXL1.0 演进 CXL2.0

- Switching：一个CXL Switch能够支持多个设备（1 USP to multiple DSP）

CXL2.0 演进 CXL3.0

- Memory Sharing: 跨主机共享，一致性由硬件实现 (Directory + Back-Invalidation)
- Multi-level Switching：支持各类设备和多级CXL Switch，支持Port-based Routing

Linux主线对于CXL支持

OpenAnolis 龙蜥社区

组件/驱动	说明	支持版本
CXL 1.1 RCEC	支持CXL识别为RCIEP	5.10
CXL 2.0 Type-3 Memory Device	CXL 2.0 PCIe 设备驱动	5.12
ACPI CEDT	ACPI CEDT	5.12
CXL PMEM	支持PMEM	5.12
CXL Host Bridge	支持CXL host bridge	5.16
CXL Memory Raw Commands	支持内存设备的原始指令接口	5.16
CXL Memory Hotplug	支持热插拔内存	5.17
CXL mem	cxl_mem driver	5.17
CXL region	支持Region粒度管理内存资源	5.17

- Linux 5.19完整支持CXL2.0驱动 (目录: driver/cxl)，下一步计划支持CXL3.0驱动
- 物理硬件适配CXL1.1，通过dax (目录: driver/dax) 驱动进行支持，daxctl工具进行管理
- 通过QEMU（`https://gitlab.com/bwidawsk/qemu`）对CXL2.0设备进行模拟，cxl工具进行管理

OpenAnolis

OpenAnolis
龙 蜥 社 区

Anolis OS:

- 5.10 release 支持CXL1.1
- Backport RCEC特性

RCEC特性

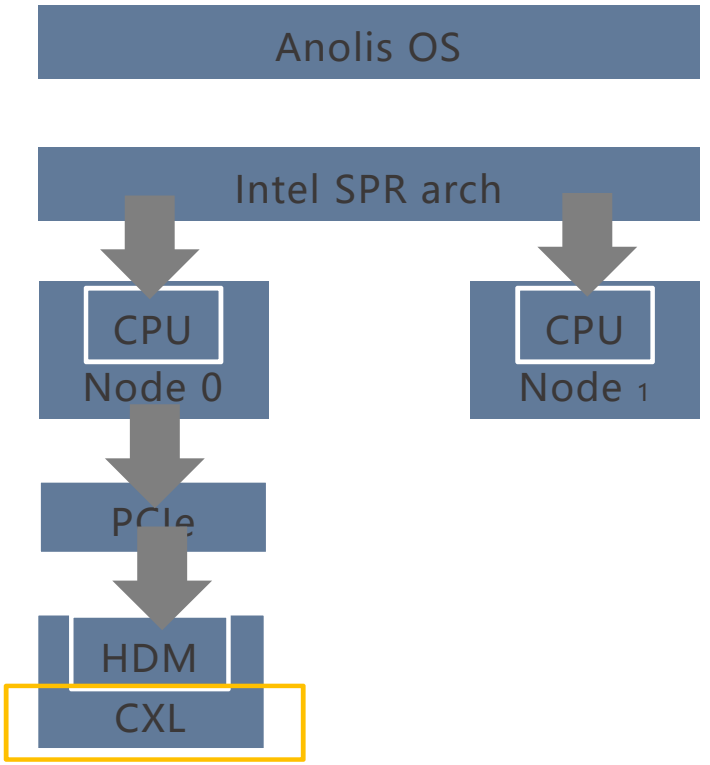
- PCIe协议的扩展
- 支持CXL上的RCiEP (Root Complex Integrated Endpoints)设备
- PME message & terminating error

CXL 1.1

- 目前能够支持FPGA模拟形式的CXL硬件设备（见下页）
- 当前为“I-Config”这一代
- CXL连接到 Node 0 的PCIe之上，虚拟成Node 2

现状和未来计划

- 主流硬件厂商Intel为龙蜥社区理事单位
- 支持CXL2.0/3.0，Y-config，虚拟化支持



CXL硬件配置

OpenAnolis 龙蜥社区

初始化为System RAM:

- CPUless NUMA node，通过设定NUMA亲和性使用
- daxctl reconfigure-device --mode=system-ram [DEV]
- daxctl list –U 显示设备

```
[
  {
    "chardev":"dax0.0",
    "size":17179869184,
    "target_node":3,
    "align":2097152,
    "mode":"system-ram",
    "movable":true
  }
]
```

初始化为devdax (原生模式):

- 在boot时，从ACPI表中注册
- Conventional memory/special purpose memory （EFI_MEMORY_SP/WB）
- Commandline配置：efi=nosoftreserve
- 通过mmap映射到虚拟内存

驱动依赖顺序:

- acpi driver => dax driver => hmem driver

QEMU模拟CXL

OpenAnolis 龙蜥社区

配置KVM/QEMU/Virsh

- 安装https://gitlab.com/bwidawsk/qemu分支(cxl-2.0v4)
- 使用CentOS镜像启动，更换5.14.10内核
- 添加源 https://repos.fedorapeople.org/repos/thl/kernel-vanilla.repo
- dnf --enablerepo=kernel-vanilla-stable update

关键启动参数:

```
-object memory-backend-file,id=cxl-mem1,share=on,mem-path=cxl-window1,size=512M \
-object memory-backend-file,id=cxl-label1,share=on,mem-path=cxl-label1,size=1K \
-object memory-backend-file,id=cxl-label2,share=on,mem-path=cxl-label2,size=1K \
-device pxb-cxl,id=cxl.0,bus=pcie.0,bus_nr=52,uid=0,len-window-base=1,window-base[0]=0x4c00000000,memdev[0]=cxl-mem1 \
-device cxl-rp,id=rp0,bus=cxl.0,addr=0.0,chassis=0,slot=0,port=0 \
-device cxl-rp,id=rp1,bus=cxl.0,addr=1.0,chassis=0,slot=1,port=1 \
-device cxl-type3,bus=rp0,memdev=cxl-mem1,id=cxl-pmem0,size=256M,lsa=cxl-label1 \
-device cxl-type3,bus=rp1,memdev=cxl-mem1,id=cxl-pmem1,size=256M,lsa=cxl-label2 \
```

QEMU内查看/配置CXL设备

- cxl list –BDMu –d root –m mem0

```
"memdevs:root0":[
  {
    "memdev":"mem0",
    "pmem_size":"256.00 MiB (268.44 MB)",
    "ram_size":0,
    "serial":"0",
    "host":"0000:0d:00.0"
  }
]
```

CXL性能

OpenAnolis 龙蜥社区

- 延迟

Node	0	1	2 (CXL)
0	116 ns	198 ns	411 ns
1	198 ns	118 ns	607 ns

- 带宽

Node	0	1	2 (CXL)
顺序	41.6 GBps	34.5 GBps	11.9 GBps
随机	4.5 GBps	2.8 GBps	2.0 GBps

- Memcached

类型	DIMM	CXL
1	137559	133763
2	138577	132511
3	135613	133588

- Redis

	Tput	<=2ms
Node 0	116615	99.98
Node 1	99072	99.94
PMEM	105638	99.96
CXL	99748	99.94

OpenAnolis
龙蜥社区

谢 谢