Compute Express Link Memory Devices

A Compute Express Link Memory Device is a CXL component that implements the CXL.mem protocol. It contains some amount of volatile memory, persistent memory, or both. It is enumerated as a PCI device for configuration and passing messages over an MMIO mailbox. Its contribution to the System Physical Address space is handled via HDM (Host Managed Device Memory) decoders that optionally define a device's contribution to an interleaved address range across multiple devices underneath a host-bridge or interleaved across host-bridges.

CXL Bus: Theory of Operation

Similar to how a RAID driver takes disk objects and assembles them into a new logical device, the CXL subsystem is tasked to take PCIe and ACPI objects and assemble them into a CXL.mem decode topology. The need for runtime configuration of the CXL.mem topology is also similar to RAID in that different environments with the same hardware configuration may decide to assemble the topology in contrasting ways. One may choose performance (RAID0) striping memory across multiple Host Bridges and endpoints while another may opt for fault tolerance and disable any striping in the CXL.mem topology.

Platform firmware enumerates a menu of interleave options at the "CXL root port" (Linux term for the top of the CXL decode topology). From there, PCIe topology dictates which endpoints can participate in which Host Bridge decode regimes. Each PCIe Switch in the path between the root and an endpoint introduces a point at which the interleave can be split. For example platform firmware may say at a given range only decodes to 1 one Host Bridge, but that Host Bridge may in turn interleave cycles across multiple Root Ports. An intervening Switch between a port and an endpoint may interleave cycles across multiple Downstream Switch Ports, etc.

Here is a sample listing of a CXL topology defined by 'cxl_test'. The 'cxl_test' module generates an emulated CXL topology of 2 Host Bridges each with 2 Root Ports. Each of those Root Ports are connected to 2-way switches with endpoints connected to those downstream ports for a total of 8 endpoints:

```
# cxl list -BEMPu -b cxl_test
{
  "bus":"root3",
  "provider":"cxl_test",
  "ports:root3":[
    {
      "port":"port5",
      "host":"cxl_host_bridge.1",
      "ports:port5":[
        {
          "port":"port8",
          "host":"cxl_switch_uport.1",
          "endpoints:port8":[
            {
              "endpoint":"endpoint9",
              "host":"mem2",
              "memdev":{
                "memdev":"mem2",
                "pmem_size":"256.00 MiB (268.44 MB)",
                "ram_size":"256.00 MiB (268.44 MB)",
                "serial":"0x1",
                "numa_node":1,
                "host":"cxl_mem.1"
              }
            },
            {
              "endpoint":"endpoint15",
              "host":"mem6",
              "memdev":{
                "memdev":"mem6",
                "pmem_size":"256.00 MiB (268.44 MB)",
                "ram_size":"256.00 MiB (268.44 MB)",
                "serial":"0x5",
                "numa_node":1,
                "host":"cxl_mem.5"
              }
            }
          ]
        },
        {
          "port":"port12",
          "host":"cxl_switch_uport.3",
          "endpoints:port12":[
            {
              "endpoint":"endpoint17",
              "host":"mem8",
              "memdev":{
```

```
              "memdev":"mem8",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
              "serial":"0x7",
              "numa_node":1,
              "host":"cxl_mem.7"
            }
          },
          {
            "endpoint":"endpoint13",
            "host":"mem4",
            "memdev":{
              "memdev":"mem4",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
              "serial":"0x3",
              "numa_node":1,
              "host":"cxl_mem.3"
            }
          }
        ]
      }
    ]
  },
  {
    "port":"port4",
    "host":"cxl_host_bridge.0",
    "ports:port4":[
      {
        "port":"port6",
        "host":"cxl_switch_uport.0",
        "endpoints:port6":[
          {
            "endpoint":"endpoint7",
            "host":"mem1",
            "memdev":{
              "memdev":"mem1",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
              "serial":"0",
              "numa_node":0,
              "host":"cxl_mem.0"
            }
          },
          {
            "endpoint":"endpoint14",
            "host":"mem5",
            "memdev":{
              "memdev":"mem5",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
              "serial":"0x4",
              "numa_node":0,
              "host":"cxl_mem.4"
            }
          }
        ]
      },
      {
        "port":"port10",
        "host":"cxl_switch_uport.2",
        "endpoints:port10":[
          {
            "endpoint":"endpoint16",
            "host":"mem7",
            "memdev":{
              "memdev":"mem7",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
              "serial":"0x6",
              "numa_node":0,
              "host":"cxl_mem.6"
            }
          },
          {
            "endpoint":"endpoint11",
            "host":"mem3",
            "memdev":{
              "memdev":"mem3",
              "pmem_size":"256.00 MiB (268.44 MB)",
              "ram_size":"256.00 MiB (268.44 MB)",
```

```
                "serial":"0x2",
                "numa_node":0,
                "host":"cxl_mem.2"
              }
            }
          ]
        }
      ]
    }
```

In that listing each "root", "port", and "endpoint" object correspond a kernel 'struct cxl_port' object. A 'cxl_port' is a device that can decode CXL.mem to its descendants. So "root" claims non-PCIe enumerable platform decode ranges and decodes them to "ports", "ports" decode to "endpoints", and "endpoints" represent the decode from SPA (System Physical Address) to DPA (Device Physical Address).

Continuing the RAID analogy, disks have both topology metadata and on device metadata that determine RAID set assembly. CXL Port topology and CXL Port link status is metadata for CXL.mem set assembly. The CXL Port topology is enumerated by the arrival of a CXL.mem device. I.e. unless and until the PCIe core attaches the cxl_pci driver to a CXL Memory Expander there is no role for CXL Port objects. Conversely for hot-unplug / removal scenarios, there is no need for the Linux PCI core to tear down switch-level CXL resources because the endpoint ->remove() event cleans up the port data that was established to support that Memory Expander.

The port metadata and potential decode schemes that a give memory device may participate can be determined via a command like:

```
# cxl list -BDMu -d root -m mem3
{
  "bus":"root3",
  "provider":"cxl_test",
  "decoders:root3":[
    {
      "decoder":"decoder3.1",
      "resource":"0x8030000000",
      "size":"512.00 MiB (536.87 MB)",
      "volatile_capable":true,
      "nr_targets":2
    },
    {
      "decoder":"decoder3.3",
      "resource":"0x8060000000",
      "size":"512.00 MiB (536.87 MB)",
      "pmem_capable":true,
      "nr_targets":2
    },
    {
      "decoder":"decoder3.0",
      "resource":"0x8020000000",
      "size":"256.00 MiB (268.44 MB)",
      "volatile_capable":true,
      "nr_targets":1
    },
    {
      "decoder":"decoder3.2",
      "resource":"0x8050000000",
      "size":"256.00 MiB (268.44 MB)",
      "pmem_capable":true,
      "nr_targets":1
    }
  ],
  "memdevs:root3":[
    {
      "memdev":"mem3",
      "pmem_size":"256.00 MiB (268.44 MB)",
      "ram_size":"256.00 MiB (268.44 MB)",
      "serial":"0x2",
      "numa_node":0,
      "host":"cxl_mem.2"
    }
  ]
}
```

...which queries the CXL topology to ask "given CXL Memory Expander with a kernel device name of 'mem3' which platform level decode ranges may this device participate". A given expander can participate in multiple CXL.mem interleave sets simultaneously depending on how many decoder resource it has. In this example mem3 can participate in one or more of a PMEM interleave that spans to Host Bridges, a PMEM interleave that targets a single Host Bridge, a Volatile memory interleave that spans 2 Host Bridges, and a Volatile memory interleave that only targets a single Host Bridge.

Conversely the memory devices that can participate in a given platform level decode scheme can be determined via a command like

the following:

```
# cxl list -MDu -d 3.2
[
  {
    "memdevs":[
      {
        "memdev":"mem1",
        "pmem_size":"256.00 MiB (268.44 MB)",
        "ram_size":"256.00 MiB (268.44 MB)",
        "serial":"0",
        "numa_node":0,
        "host":"cxl_mem.0"
      },
      {
        "memdev":"mem5",
        "pmem_size":"256.00 MiB (268.44 MB)",
        "ram_size":"256.00 MiB (268.44 MB)",
        "serial":"0x4",
        "numa_node":0,
        "host":"cxl_mem.4"
      },
      {
        "memdev":"mem7",
        "pmem_size":"256.00 MiB (268.44 MB)",
        "ram_size":"256.00 MiB (268.44 MB)",
        "serial":"0x6",
        "numa_node":0,
        "host":"cxl_mem.6"
      },
      {
        "memdev":"mem3",
        "pmem_size":"256.00 MiB (268.44 MB)",
        "ram_size":"256.00 MiB (268.44 MB)",
        "serial":"0x2",
        "numa_node":0,
        "host":"cxl_mem.2"
      }
    ]
  },
  {
    "root decoders":[
      {
        "decoder":"decoder3.2",
        "resource":"0x8050000000",
        "size":"256.00 MiB (268.44 MB)",
        "pmem_capable":true,
        "nr_targets":1
      }
    ]
  }
]
```

...where the naming scheme for decoders is "decoder<port_id>.<instance_id>".

Driver Infrastructure

This section covers the driver infrastructure for a CXL memory device.

CXL Memory Device

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 322)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/pci.c
   :doc: cxl pci
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 325)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/pci.c
   :internal:
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 328)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/mem.c
   :doc: cxl mem
```

CXL Port

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 333)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/port.c
   :doc: cxl port
```

CXL Core

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 338)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/cxl.h
   :doc: cxl objects
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 341)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/cxl.h
   :internal:
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 344)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/port.c
   :doc: cxl core
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 347)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/port.c
   :identifiers:
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 350)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/pci.c
   :doc: cxl core pci
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 353)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/pci.c
   :identifiers:
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 356)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/pmem.c
   :doc: cxl pmem
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 359)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/regs.c
   :doc: cxl registers
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 362)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: drivers/cxl/core/mbox.c
   :doc: cxl mbox
```

External Interfaces

CXL IOCTL Interface

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 371)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: include/uapi/linux/cxl_mem.h
   :doc: UAPI
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\driver-api\cxl\[linux-master][Documentation][driver-api][cxl]memory-devices.rst, line 374)

Unknown directive type "kernel-doc".

```
.. kernel-doc:: include/uapi/linux/cxl_mem.h
   :internal:
```