SNIA DEVELOPER CONFERENCE

SDC 25

By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

CXL Memory in Windows

Chet Douglas – Principal Software Engineer, Microsoft
Scott Lee – Principal Software Engineer Manager, Microsoft

Agenda

- CXL Memory in Windows
- CXL Memory Virtualization
- Windows CXL Memory Architecture
- CXL Memory Reliability Availability and Serviceability (RAS)
- Current Status
- Futures

CXL Memory in Windows

- View CXL Memory as part of the general problem of multi-tier memory support in the OS
- Two approaches for CXL memory usage
  - General Purpose (GP) Memory
    - Memory-only Non-Uniform Memory Architecture (NUMA) node for memory that have different characteristics than memory on NUMA node that has CPU
    - Memory available to any requester so can create unexpected issues if performance is slower
    - Default OS memory management prioritizes memory allocation from nearest NUMA node
    - No software changes required
  - Specific Purpose (SP) Memory
    - Each memory tier in memory-only NUMA node
    - Memory dedicated for specific usages and allocated through new APIs (Application Programming Interface)
    - Software stores data in different memory tiers based on its characteristics
    - Requires software code changes
    - Called Dedicated Memory in Windows
    - OS Memory Manager can use SP memory for various usages (e.g. pagefile, compressed and standby pages)
- Prioritizing OS-first RAS over FW-first RAS handling
- Will have a built-in CXL Type 3 driver
- Windows and Linux are aligned at a high level but differences in how SP memory is supported

Dedicated (SP) Memory in Windows

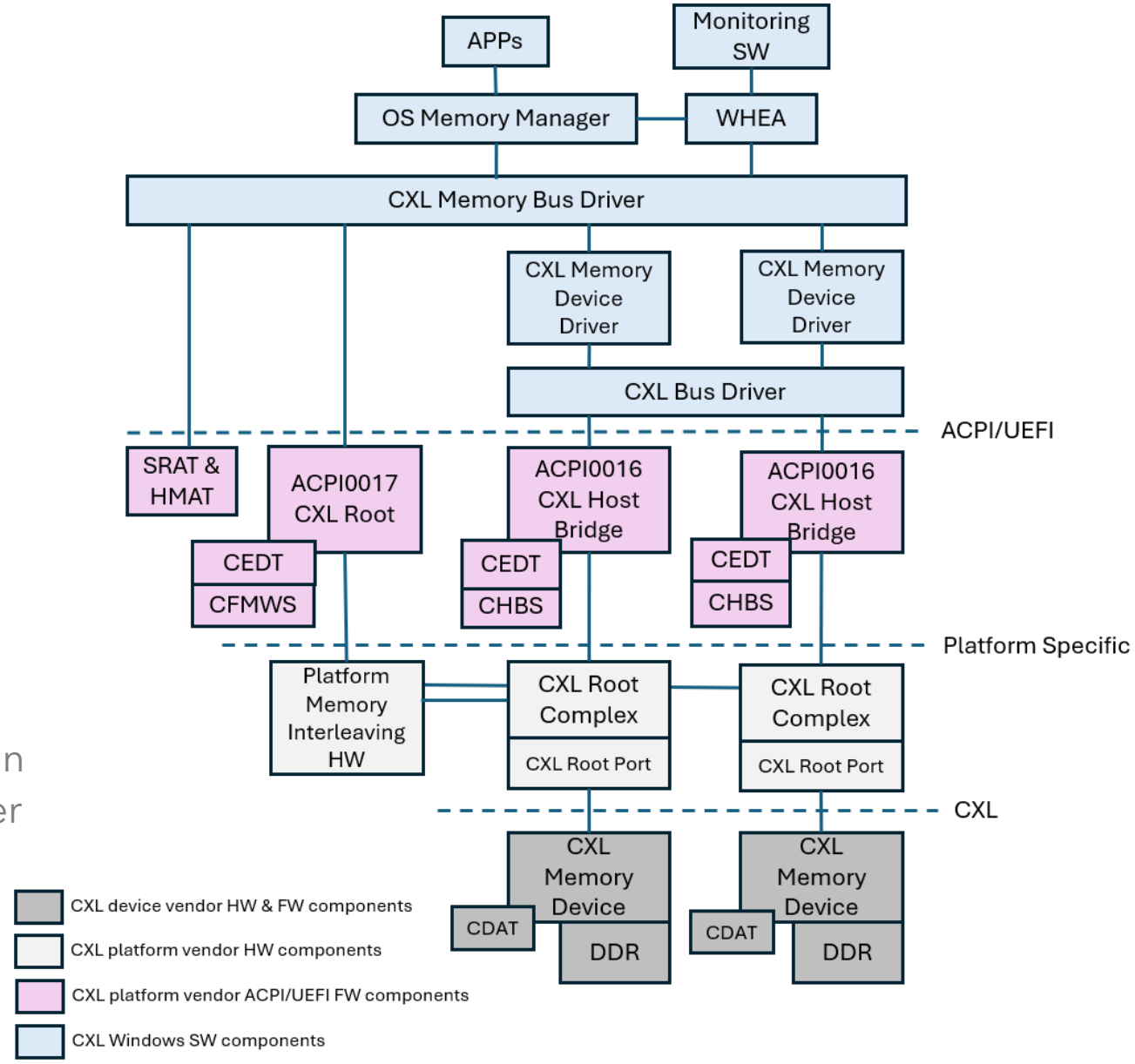
- System can have many different Dedicated Memory types
- Each Dedicated Memory type has a unique combination of attributes
  - Current attributes: read, write latency and bandwidth
  - Number of attributes to grow over time
- Performance characteristics reported will be the speed of the media outside of any intermediate caches and from the closest CPU node
- Dedicated Memory support through existing memory partition and VirtualAlloc APIs in Windows
  - A memory partition can have both regular memory and Dedicated Memory
  - APIs available to enumerate, allocate and free Dedicated Memory
- New Win32 memory-only NUMA node APIs
  - Retrieve Advanced Configuration and Power Interface (ACPI) Heterogeneous Memory Attribute Table (HMAT) related info
  - Discover closest initiator node

CXL Memory Virtualization

- Do not see any current needs to virtualize CXL memory devices to virtual machine (VM)
- Can use CXL memory in VMs using generic reporting of memory as either GP or SP Memory
- At VM creation time, can specify
  - GP or SP attribute of the memory
  - Memory only NUMA node

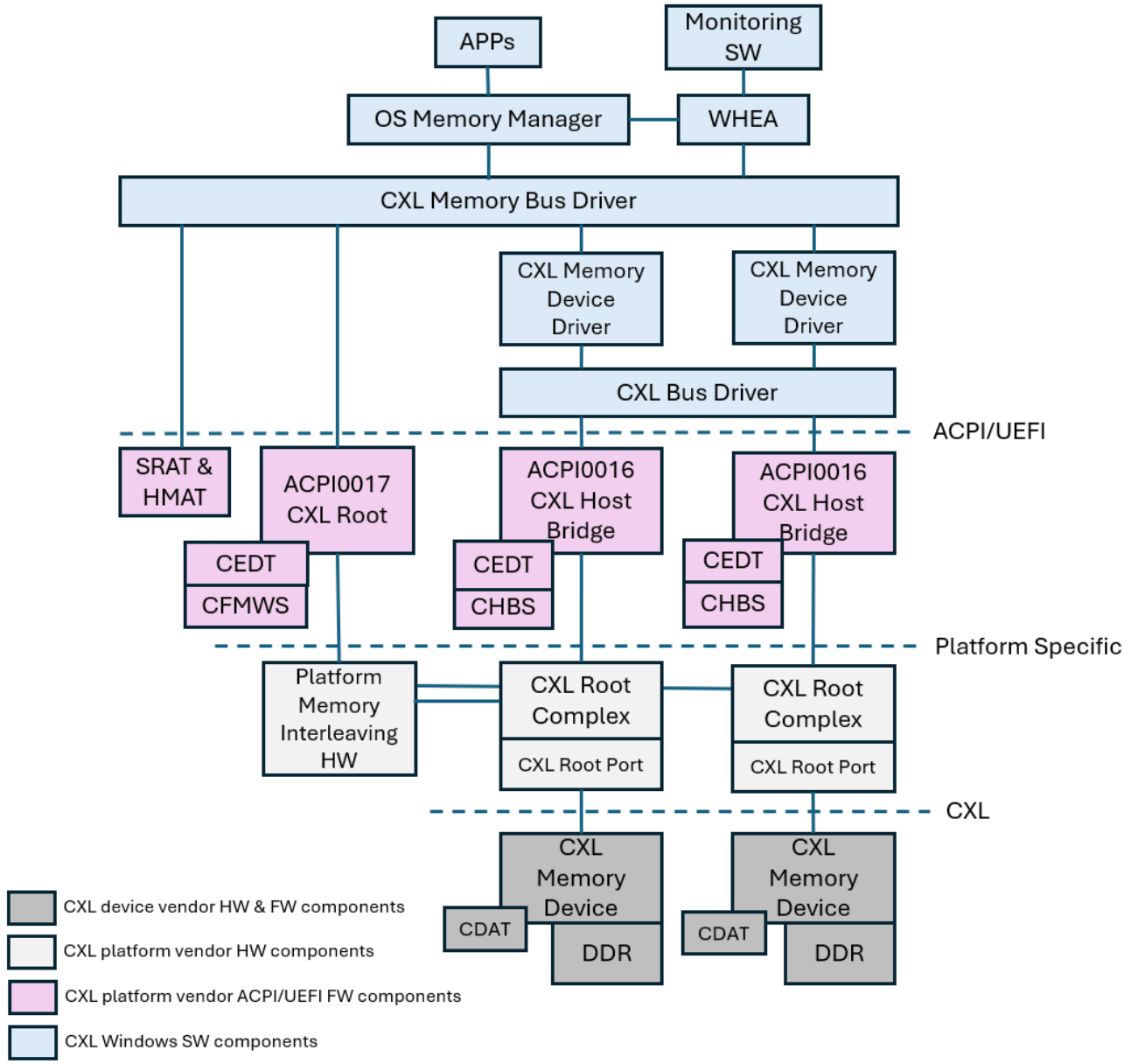
CXL in Windows –
CXL architecture in Windows

- CXL Bus Driver
  - CXL hierarchy enumeration
  - CXL root port level functionalities
  - CXL protocol & link error handling – Advanced Error Reporting (AER)
- CXL Memory Device Driver
  - Loaded for each Type 3 device enumerated
  - Utilizes CXL Type 3 Mailbox Component Command Interface (CCI)
  - Alert Configuration
  - Event record handling and interrupt configuration
  - Event record reporting to CXL Memory Bus Driver
  - Poison list harvesting
  - Device Status register handling (e.g. FW Halt)

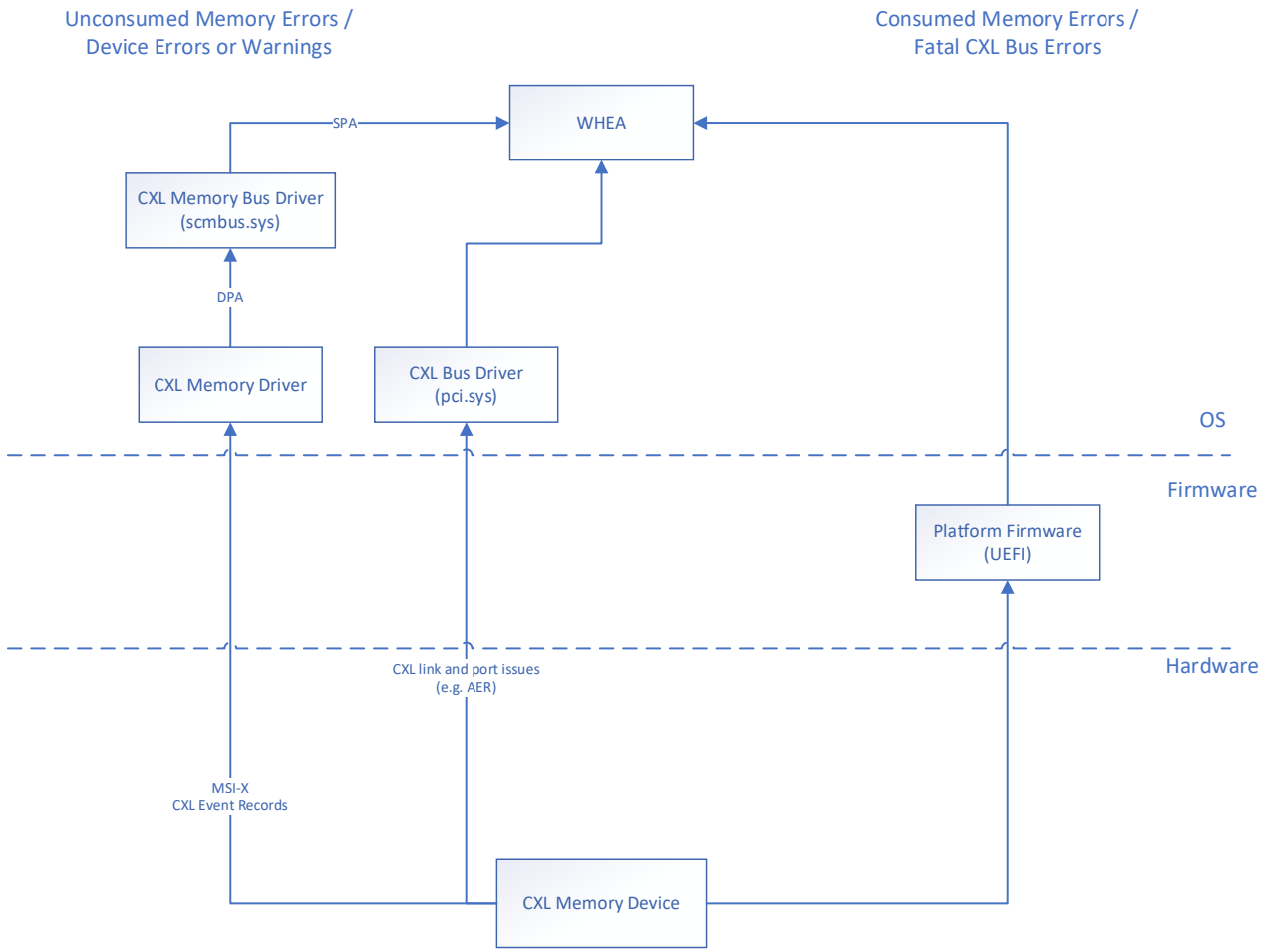


CXL in Windows –
CXL architecture in Windows

- CXL Memory Bus Driver
  - System level CXL memory functionalities
    - Memory interleaving support
    - SP memory ranges reporting to Memory Manager
    - Address translation (e.g., DPA to SPA)
  - Loaded for the ACPI0017 object
  - Consumes CXL Early Discovery Table (CEDT) CXL Fixed Memory Window Structures (CFMWS)
  - Reports error through Common Platform Error Record (CPER) payload to WHEA
- Windows Hardware Error Architecture (WHEA)
  - Windows kernel subsystem for hardware event handling
  - Acts on CPER
  - Off-lines memory ranges for specific errors through Memory Manager
  - Generates BMC System Event Log (SEL) entries



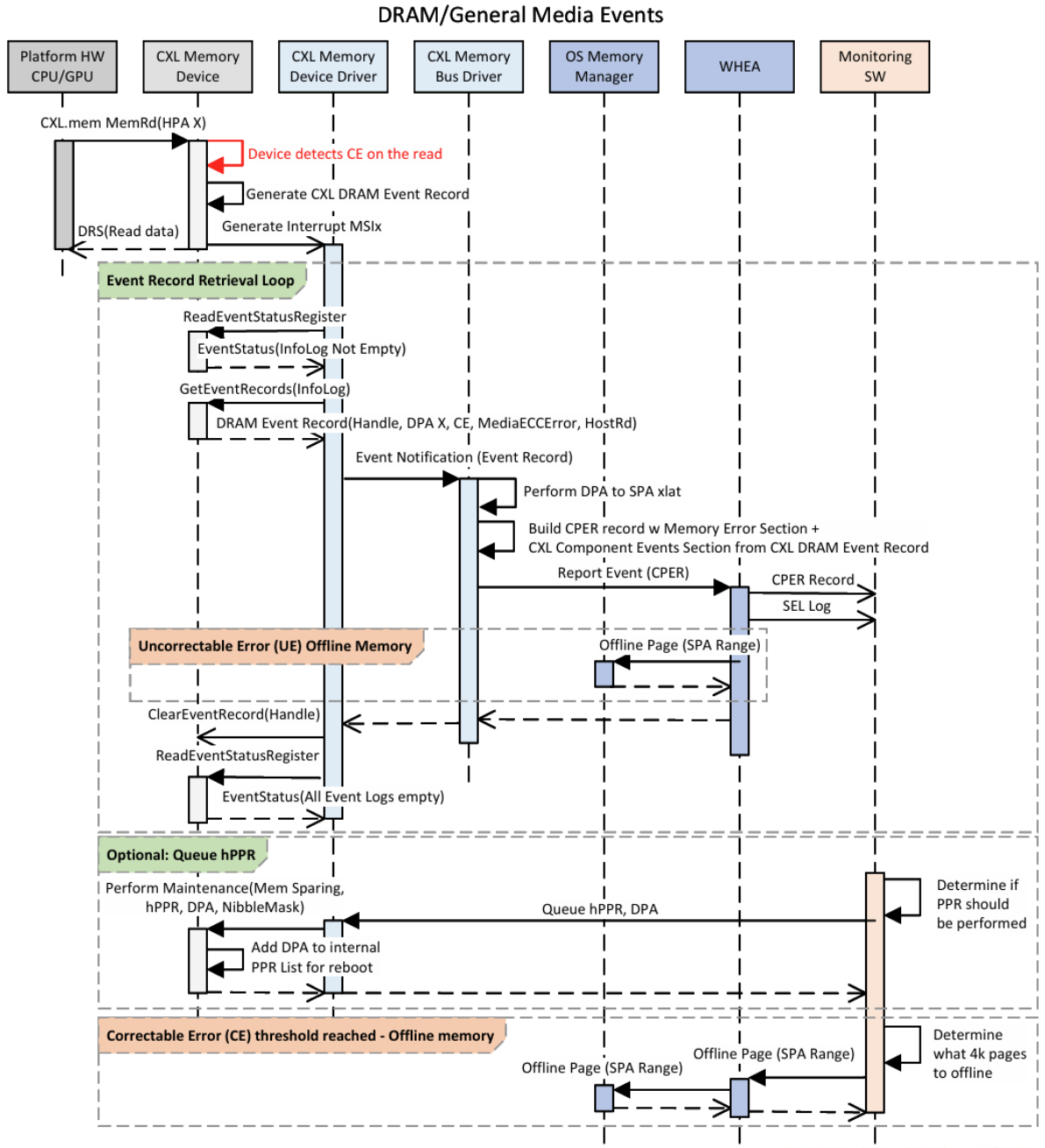
OS-First CXL Memory RAS



- CXL Memory Driver detected through interrupt and CXL event records
  - Unconsumed memory errors
  - CXL memory module level events
- CXL Bus Driver detected through interrupt
  - CXL bus and root port related events, protocol errors (e.g. AER)
- Platform Firmware detected through CPU/System on Chip (SoC)
  - Consumed uncorrectable memory error
  - Viral
- WHEA handles the hardware event
  - Events convey through CPER
  - Offline memory at 4K page size or greater size depending on the error
  - Crash the system if can't offline affected memory

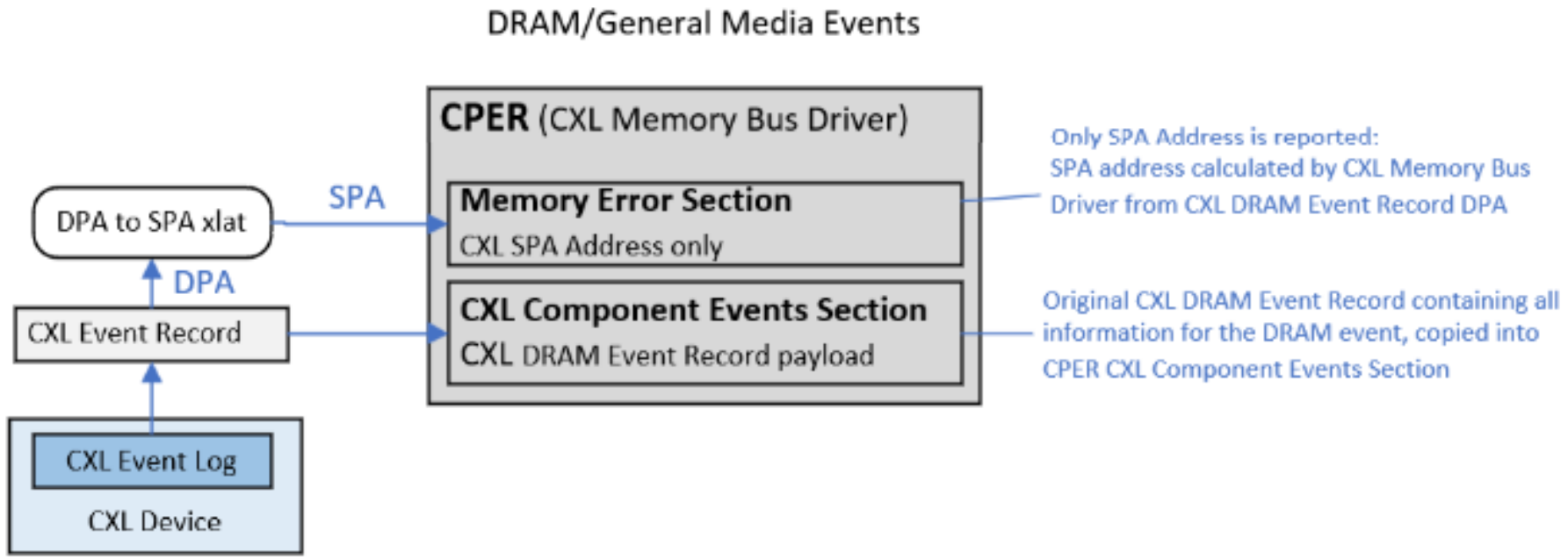
CXL in Windows – CXL OS First RAS

- DRAM/General Media Events
  - Memory events that have a known Device Physical Address (DPA)
  - CXL Device:
    - Detects Correctable Error (CE) or Uncorrectable Error (UE) memory errors
    - Adds memory location to Poison List
    - Adds DRAM Event Record to Event Log
    - Generates Message Signal Interrupt (MSIx)
  - CXL Memory Device Driver:
    - Retrieves event record from device
    - Reports event to CXL Memory Bus Driver
  - CXL Memory Bus Driver:
    - Utilizes Host Device Memory (HDM) decoder information to translate the DPA containing the error into a System Physical Address (SPA)
    - Builds CPER payload containing the CXL Event Record and the SPA address
    - Reports CPER to WHEA
  - WHEA
    - May add additional CPER sections
    - Generates SEL event(s)
    - For UE errors, off-lines SPA address with OS Memory Manager
  - OS Memory Manager
    - Removes off-lined memory from address space
    - Prevents applications from accessing off-lined memory

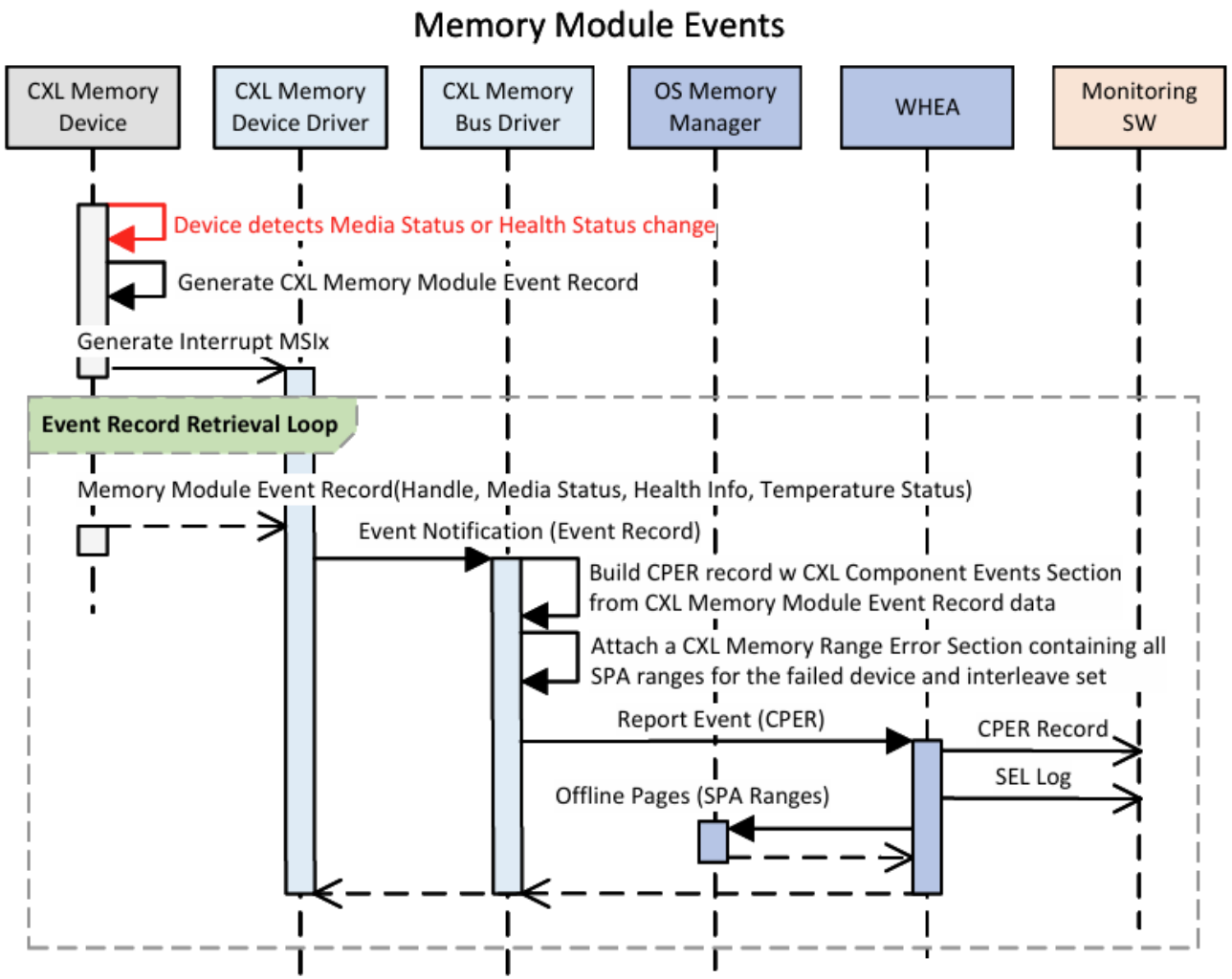


CPER sections for DRAM/General Media Events

- Page off-lining from DRAM/General Media Events
  - For events where a specific cacheline of memory is specified
  - CPER payload is created with a Memory Error Section containing a single SPA address where the event occured



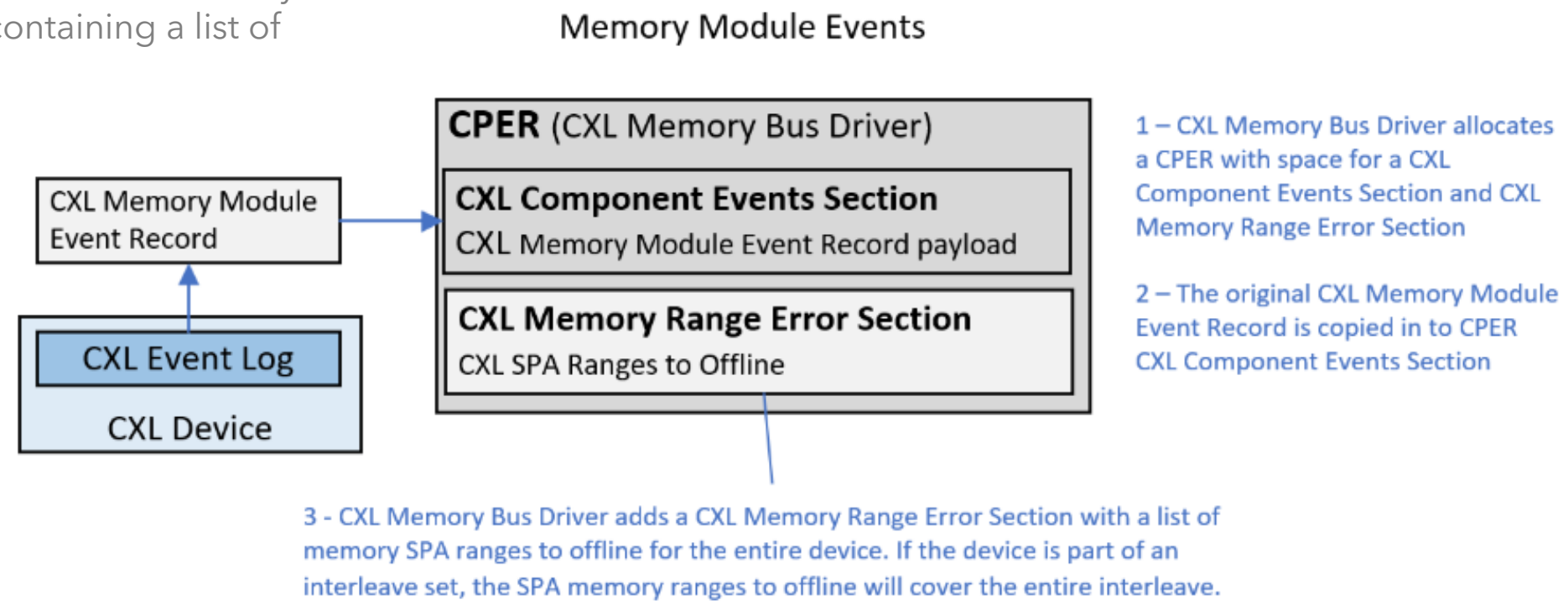
CXL in Windows –
CXL OS First RAS



- Memory Module Events
  - Events that effect the entire CXL device
  - CXL Device:
    - Detects memory module event: Temperature change, media status change, data path errors, power management faults, performance degraded, etc
    - Adds Memory Module Event Record to Event Log
    - Generates MSIx interrupt
  - CXL Memory Device Driver:
    - Retrieves event record from device
    - Reports event to CXL Memory Bus Driver
  - CXL Memory Bus Driver:
    - Filters events looking for errors requiring device memory to be removed
    - Determines all SPA address ranges effected by removal of devices' memory
    - Builds CPER payload containing the CXL Event Record and the SPA address ranges to off-line
    - Reports CPER to WHEA
  - WHEA
    - May add additional CPER sections
    - Offline memory ranges if applicable
    - Generates SEL event(s)

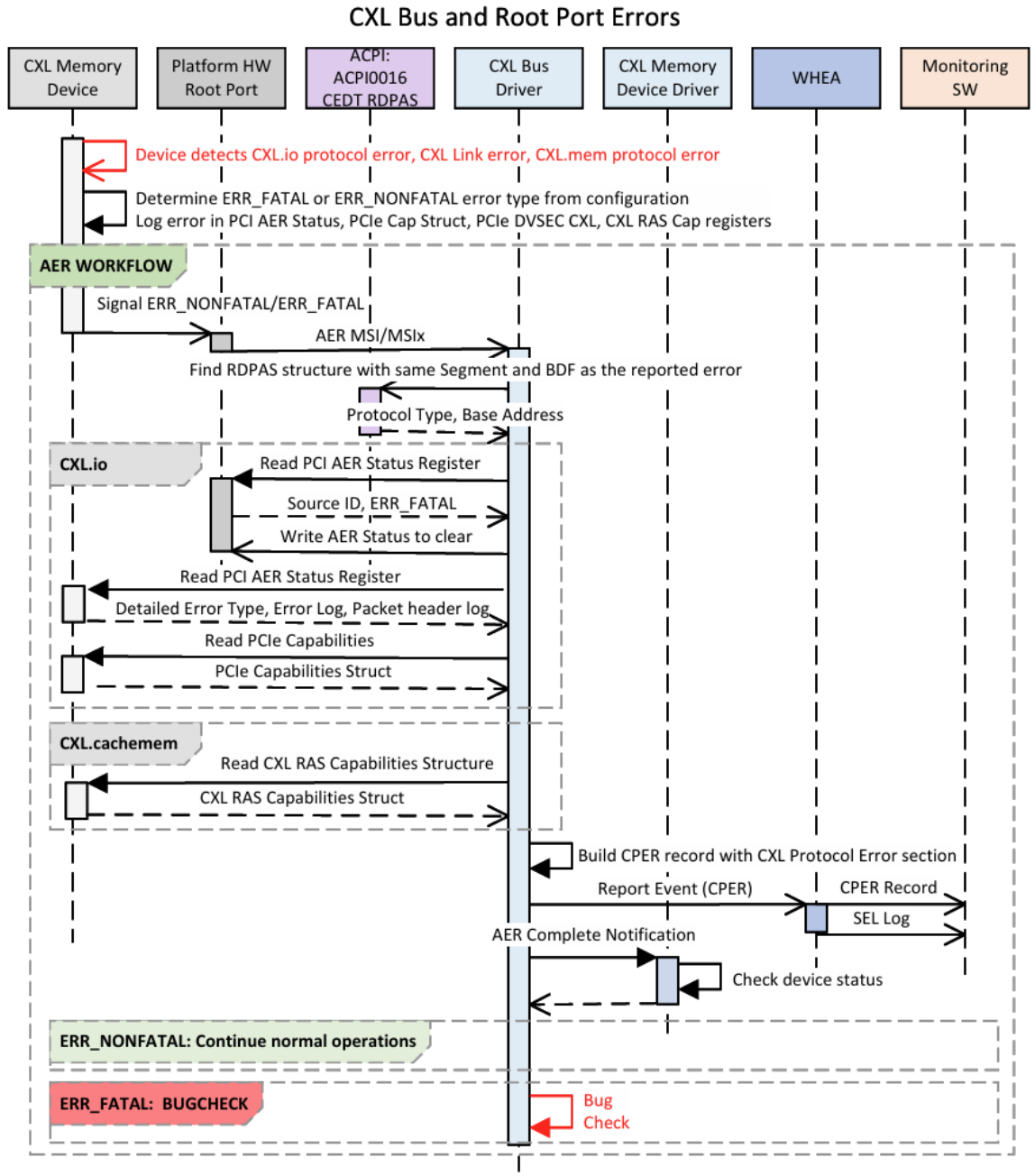
CPER sections for Memory Module Events

- Device off-lining from Memory Module Events
  - For events where device memory is not usable
  - CPER payload is created with a Memory Range Error Section containing a list of affected SPA ranges



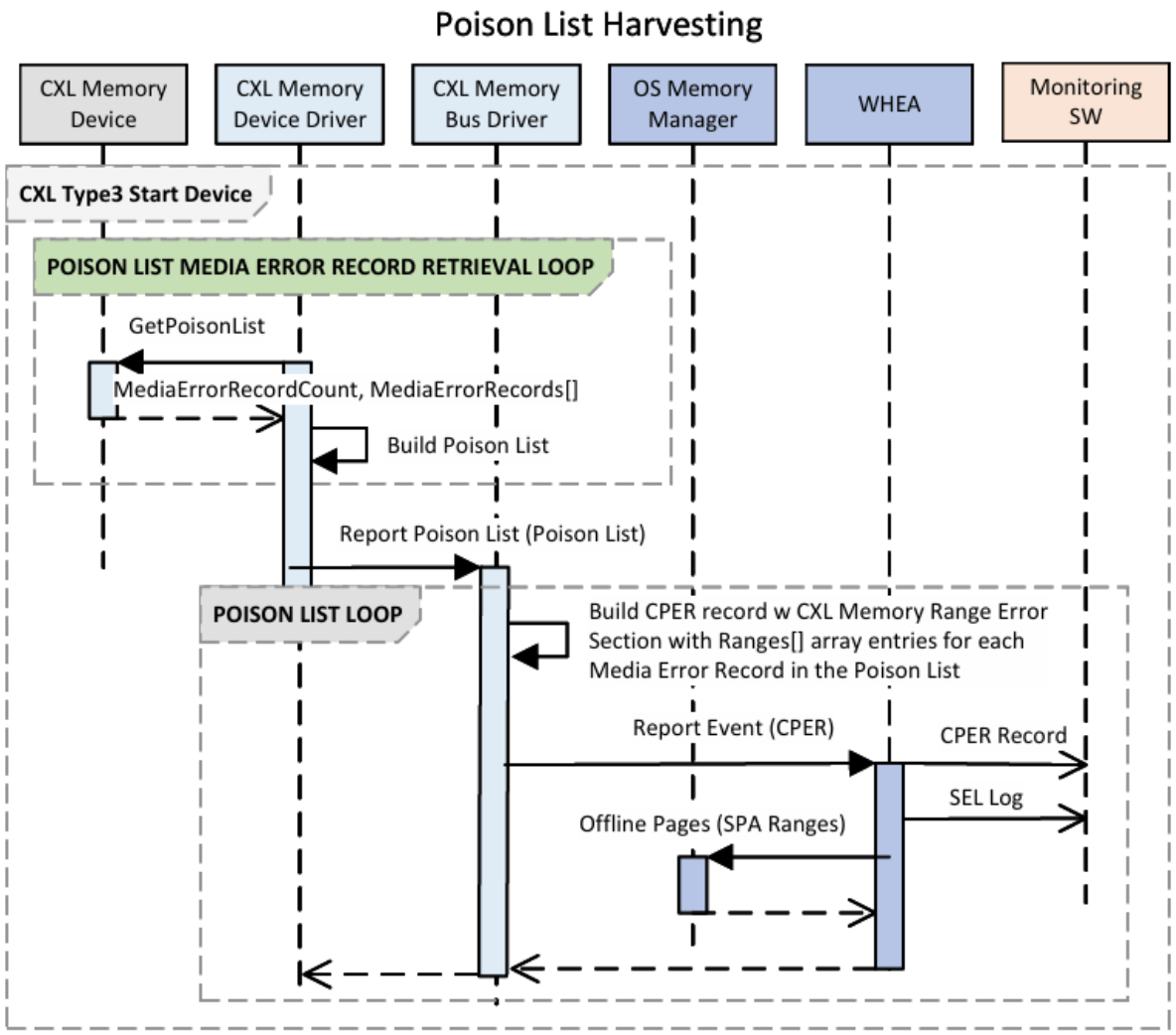
CXL in Windows –
CXL OS First RAS

- CXL Protocol Errors
  - CXL Bus, Link and Root Port protocol errors
  - CXL Device:
    - Detects CXL.io or .mem protocol or link error
    - Logs error information in PCI and CXL registers
    - Signals CXL root port ERR_NONFATAL or ERR_FATAL error
  - CXL Root Port:
    - Generates AER MSIx interrupt to CXL Bus Driver
  - CXL Bus Driver
    - Utilizes ACPI RDPAS table to find PCI and CXL error registers
    - Read/clear PCI AER Status in Root Port
    - CXL.io: Read PCI AER Status and PCIe Capabilities from the device
    - CXL.mem: Read CXL RAS Capabilities registers from device
    - Report CPER AER event to WHEA
    - Notify the CXL Memory Device Driver of AER completion
    - ERR_NONFATAL: Continue operations
    - ERR_FATAL: Crash the system
  - WHEA
    - Generates SEL event

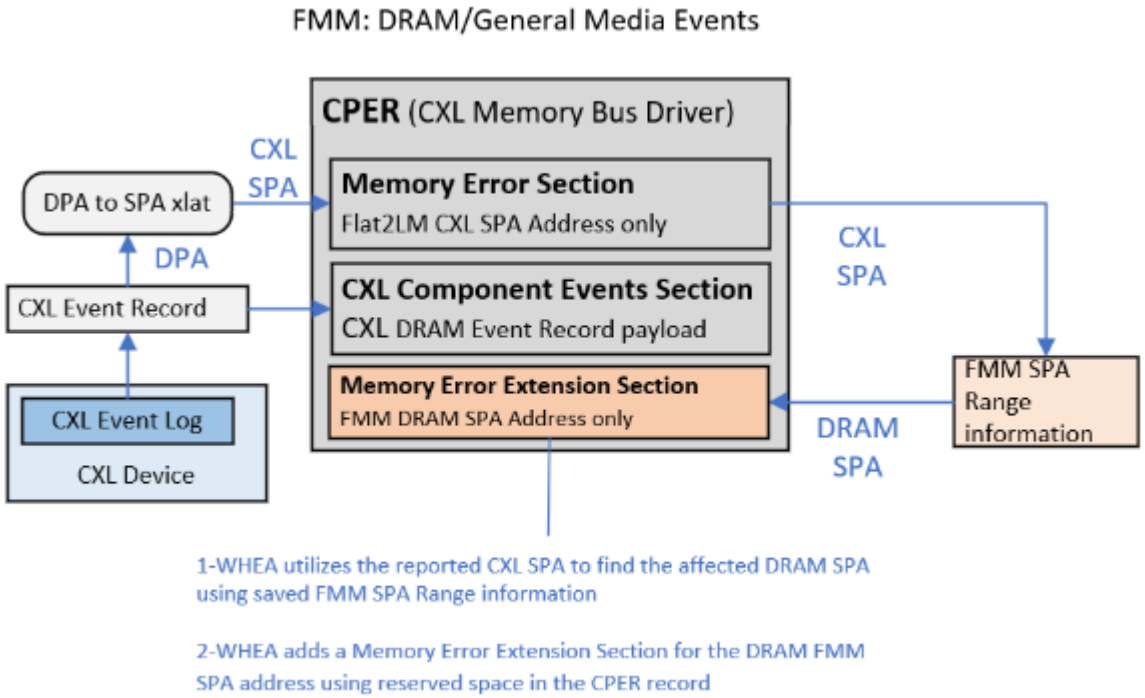


CXL in Windows –
CXL OS First RAS

- Poison List harvesting at device initialization
  - CXL Memory Device Driver:
    - At start device time the Poison List is retrieved from the CXL device
    - Each Media Error Record is reported to CXL Memory Bus Driver
  - CXL Memory Bus Driver:
    - Utilizes HDM decoder information to translate the DPA(s) from the Poison List into a list of SPA addresses
    - Builds CPER payload containing the list of affected SPA addresses
    - Reports CPER to WHEA
  - WHEA
    - May add additional CPER sections
    - Generates SEL event(s)
    - Off-lines SPA address with OS Memory Manager
  - OS Memory Manager
    - Removes off-lined memory from address space
    - Prevents applications from accessing off-lined memory



Intel Flat Memory Mode RAS



- Intel Flat Memory Mode (FMM)
  - Single address space containing equal parts DRAM and CXL attached memory
    - Platform HW controls placement of data in DRAM or CXL memory
  - Implications to RAS:
    - Data could be found at 2 possible SPA
    - Events that require CXL memory to be off-lined will also need to off-line the associated cache line in DRAM
    - Event that require DRAM memory to be off-lined will also need to off-line the associated cacheline in CXL memory
    - WHEA will determine the extra SPA address and add an additional Memory Error Extension Section to the CPER payload to report the additional address

Current Status

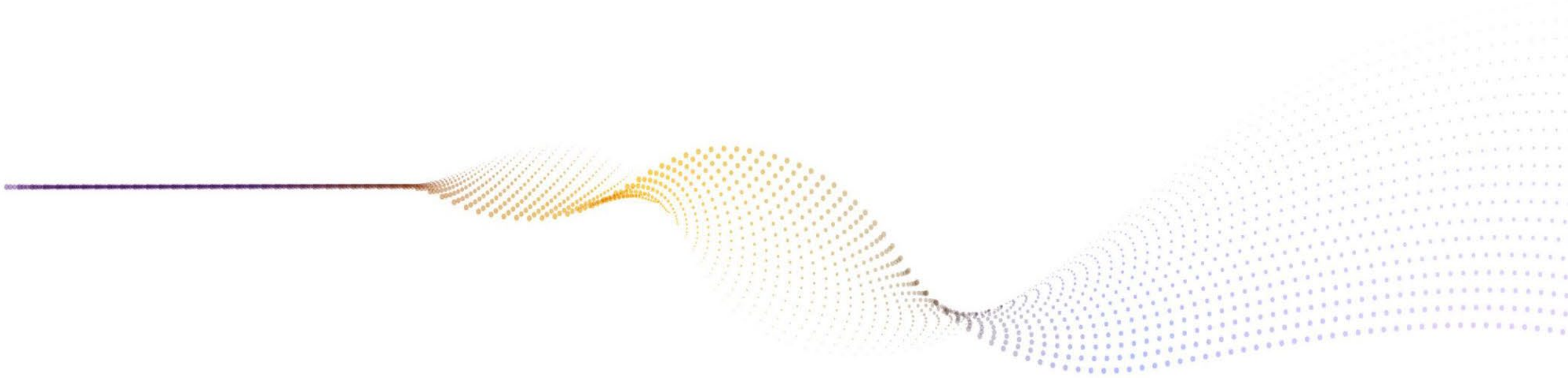
- Working on support for CXL 2.0 or greater implementations with OS-First RAS with interested applications
- Initial focus is on memory expansion (i.e. direct attached CXL memory devices) scenarios with no CXL switch
- Have preview Windows builds available to try
  - NDA (Non-Disclosure Agreement) partners have access to more collaterals
  - Non-NDA partners can gain access through Windows Insider program
  - Support available in both Windows server and client OS
- Official support is TBD (To Be Determine)

Pain Points During Enablement

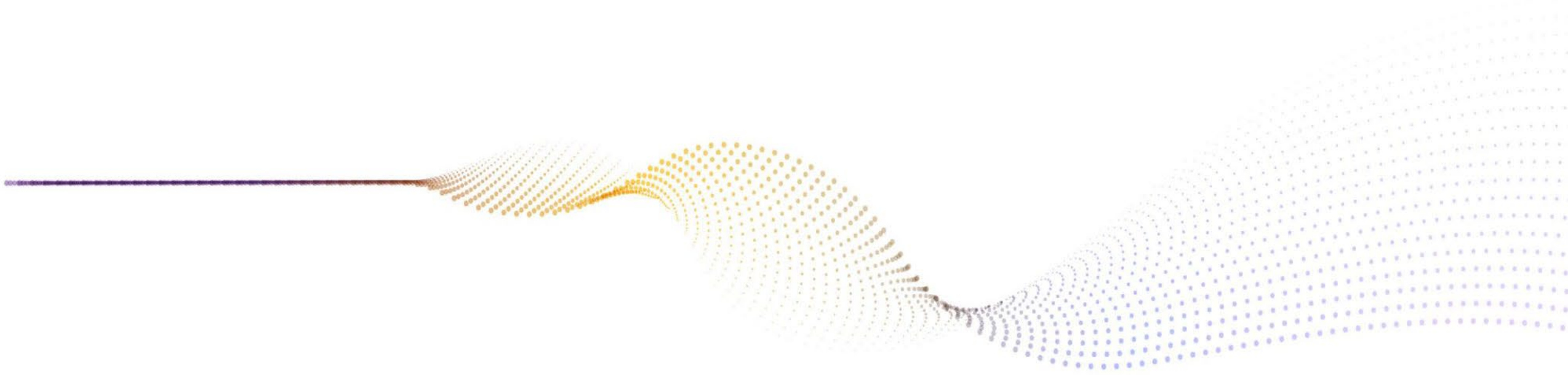
- CXL memory devices not working when installed in system
  - Extra power needed for working devices
  - Platform firmware version incompatibilities
  - Dual In-line Memory Module (DIMM) compatibility
- Implementations not fully spec compliant
- Found spec gaps for Field Replaceable Unit (FRU) identification info for DRAM components in CXL memory device
- Minimal error injection definition in CXL spec to be able to verify error handling workflows. Need to rely on vendor specific mechanism currently.

Futures

- Firmware-First RAS
- Dynamic Capacity Device (DCD)
- Non-volatile/Persistent Memory
- Trusted Execution Environment Security Protocol (TSP)
- Hot-add and remove



Questions



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

