Multi-Root Resource Discovery and Allocation

Michael Krause (HP, co-chair)

Renato Recio (IBM, co-chair)

Work In Progress

NOTE: The information in this presentation refers to a specification still in the development process. This presentation reflects the current thinking of the workgroup, but all material is subject to change before the specification is released.

Contents

- Overviews
  - ✓ Adapter Sharing Approaches: Intermediary vs Native
  - ✓ Function types
  - ✓ Terms
  - ✓ Initialization Flows
  - ✓ Mapping of (new) PCI IOV Capabilities to Function types
- New PCIe® IO Virtualization (IOV) Capabilities register set.
  - ✓ MR-IOV Capabilities Structure

Resource Discovery and Allocation

Overviews

Function types

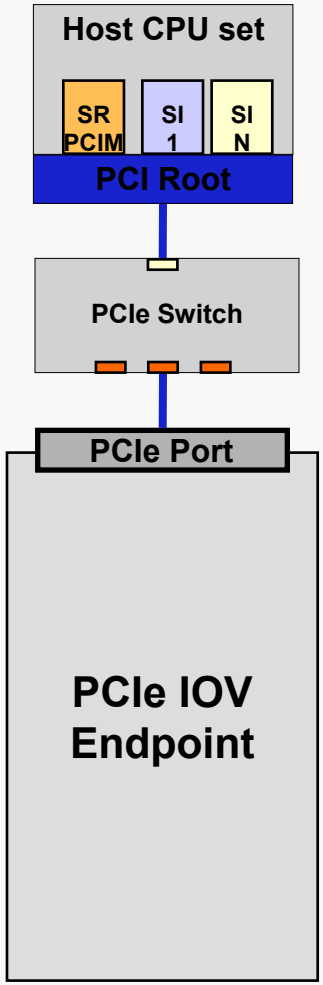
Terms

Initialization Flows

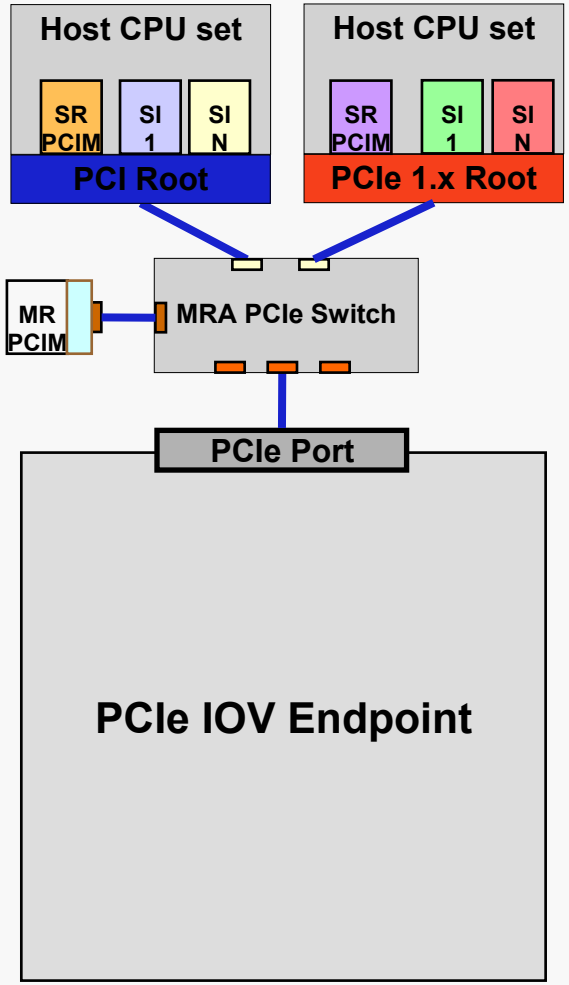
Mapping of (new) PCI IOV Capabilities to Function types

Topology Overview and Terms

SR Topology



Multi-Root Topology



Terms

Single Root (SR) IOV Overview

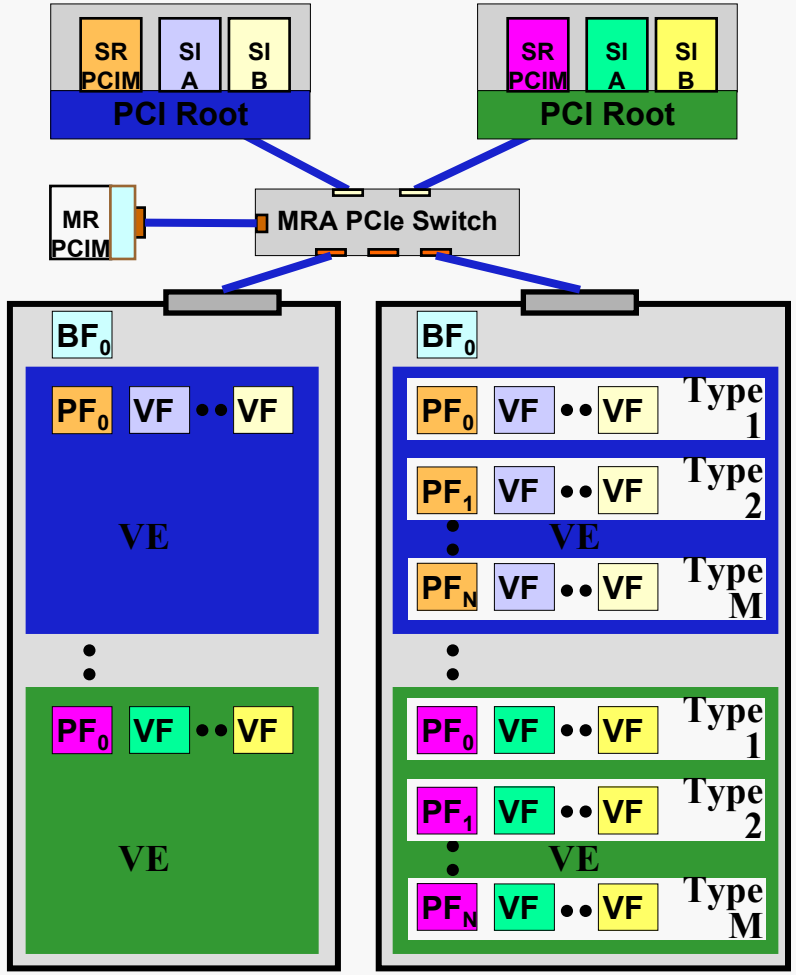
- Only has one Root.
- Switches only need to support PCIe base functionality.
- To make full use of IOV, EP must support SR-IOV capabilities.
- SR-PCIM configures the EP.

Multi-Root (MR) IOV Overview,

- One or more Roots.
- Switches with Multi-Root Aware (MRA) functionality are needed.
- To make full use of IOV, EP must support SR & MR-IOV capabilities.
- MR-PCIM assigns Virtual Endpoints (VEs) to RCs and manages PCIe components.
- SR-PCIM configures its VEs.

Multi-Root IOV Function Types and Terms

MR Topology



MR Topology Terms

Virtual Endpoint (VE) is the set of physical and virtual functions assigned to an RC.

Each VE is assigned to a **Virtual Hierarchy** (VH).

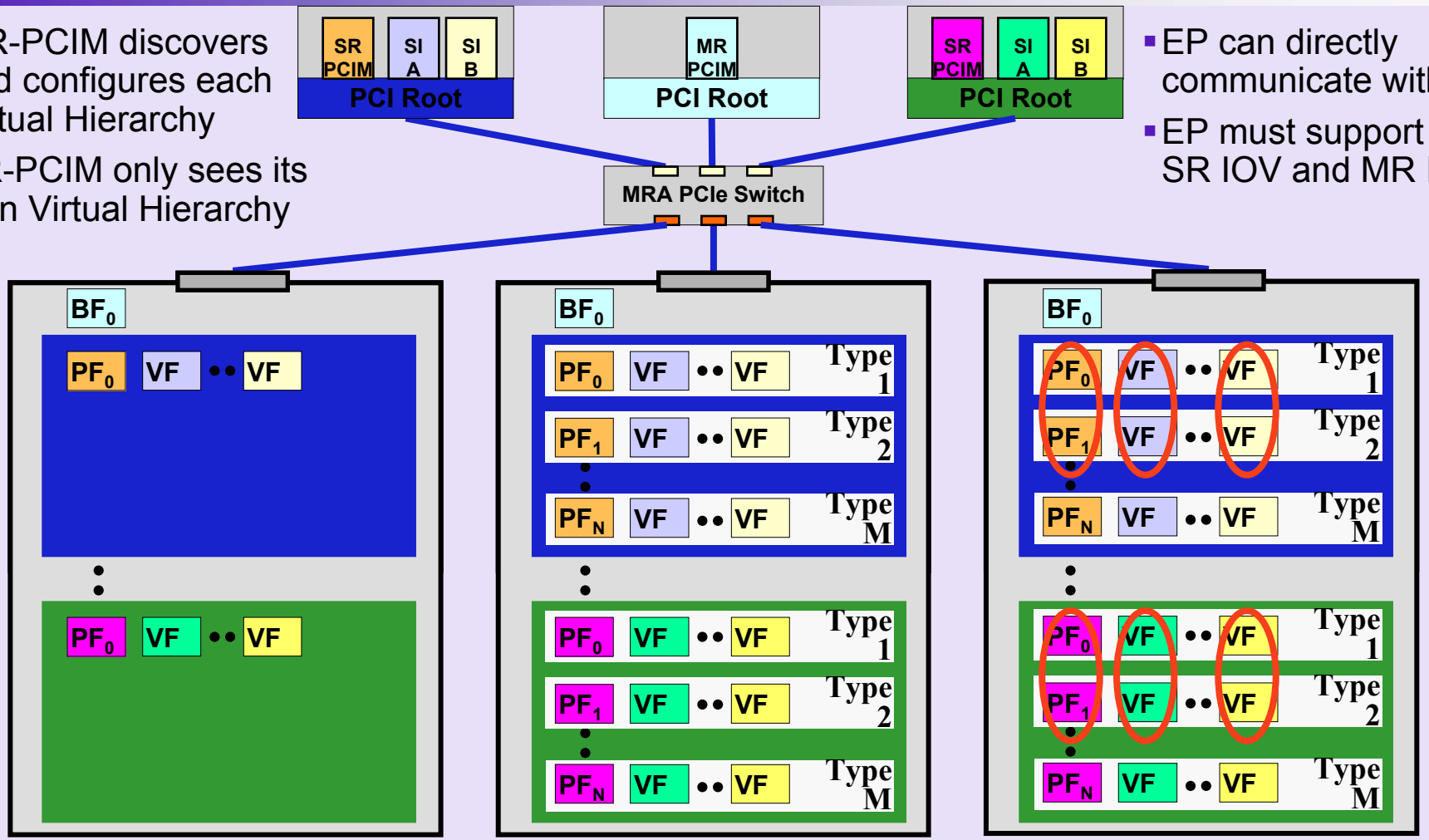
Virtual Hierarchy is a fully functional PCIe hierarchy that is assigned to an RC or MR-PCIM. Note, all PFs and VFs in a VE are assigned the same VH.

Base Function (BF) only 1 per EP and is used by MR-PCIM to manage an MR aware EP (e.g. assigning functions to Virtual Endpoints).

More details later …

EP enabled for SR and MR IO Virtualization

- MR-PCIM discovers and configures each Virtual Hierarchy
- SR-PCIM only sees its own Virtual Hierarchy

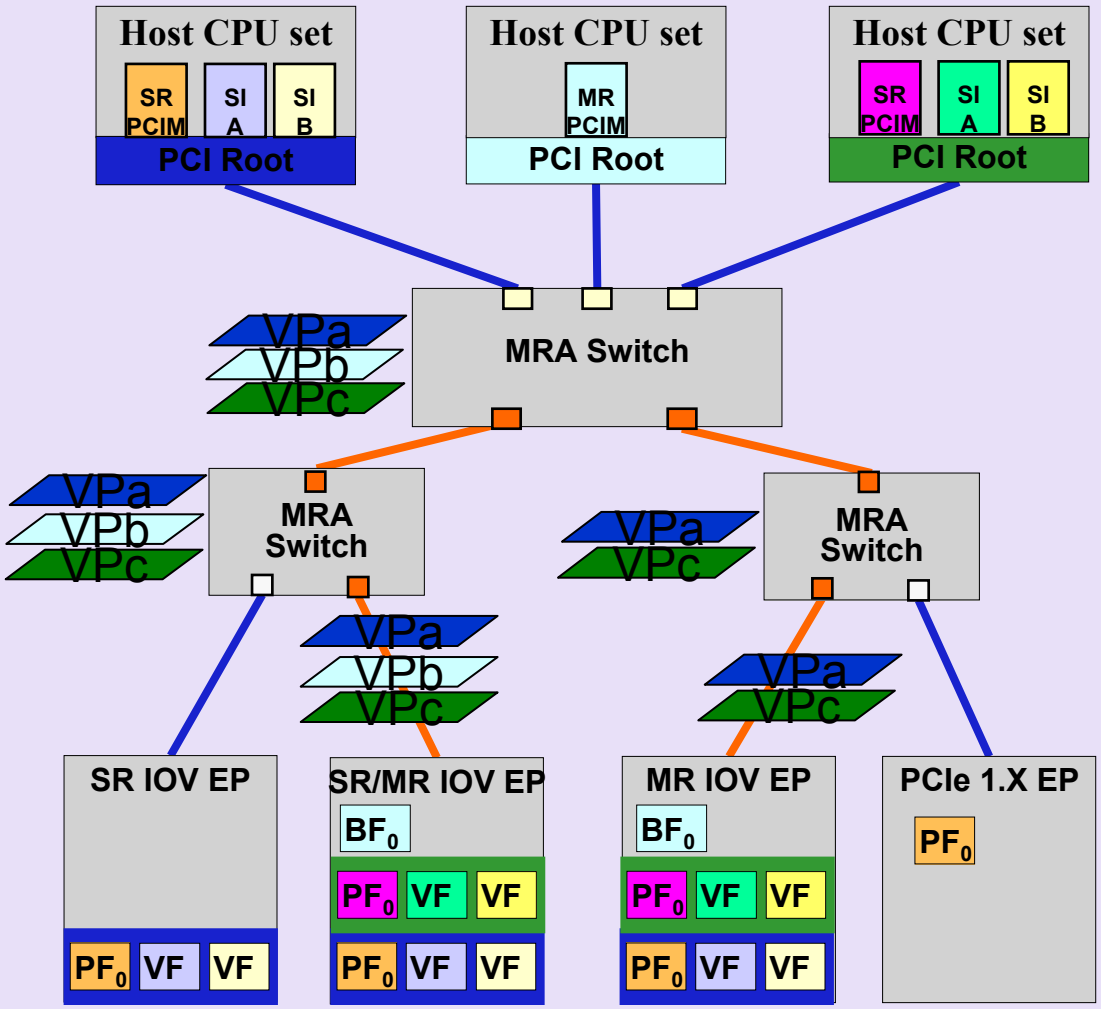


- EP can directly communicate with SI
- EP must support SR IOV and MR IOV

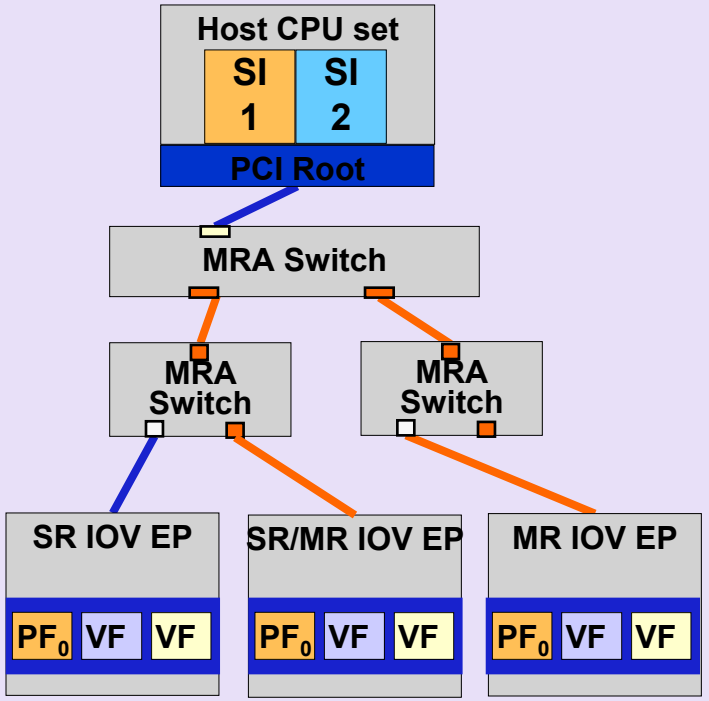
- Use by EP with one PF
- Use by EP with multiple independent functions
- Use by EP with multiple dependent functions

Mixed MR Topology Example

Physical Topology



Virtual Hierarchy

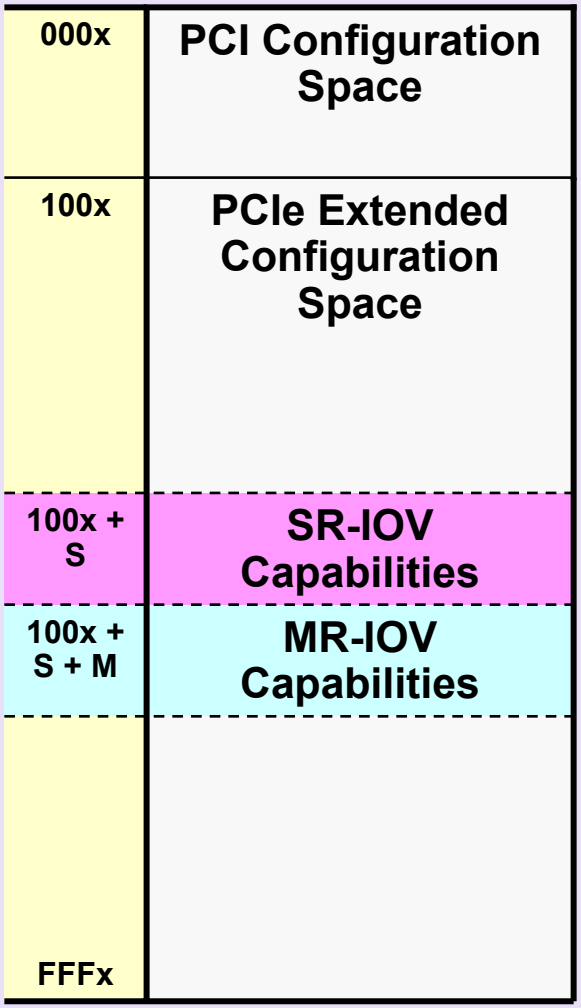


Each RC sees a virtual hierarchy

- Full PCIe functionality within a virtual hierarchy
- RC doesn't see EPs that are not part of its virtual hierarchy

IOV Capabilities

Function 0's Configuration Space Layout



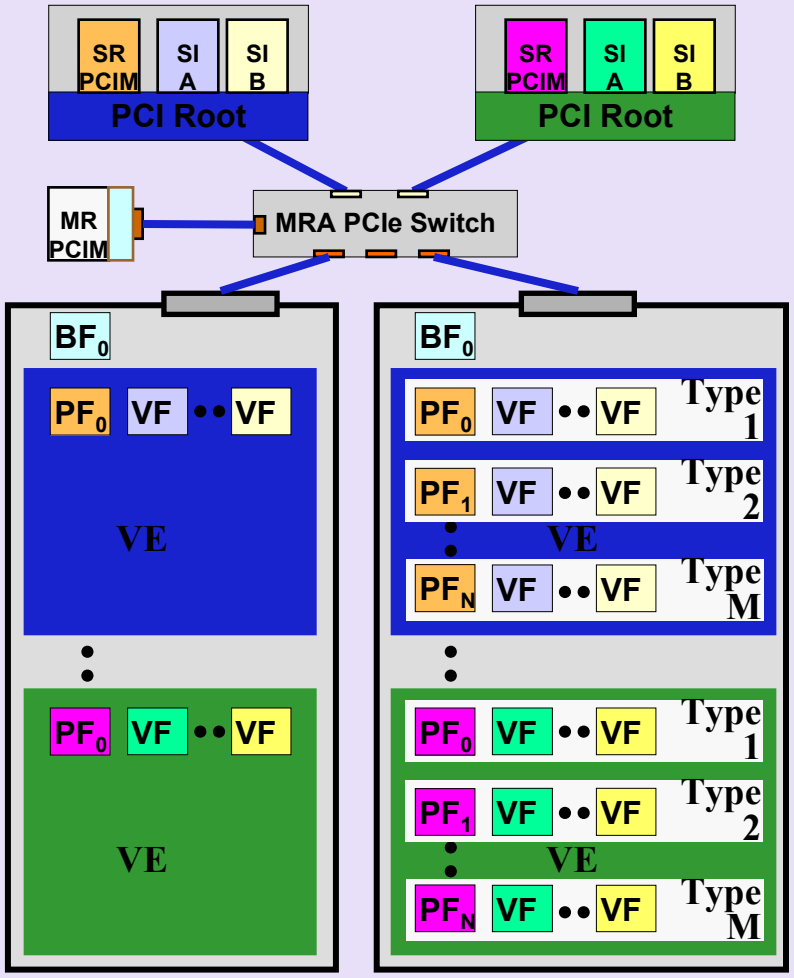
- IOV adds two new, optional PCIe Extended Capabilities to Function 0.
  - ✓ SR-IOV Capability - used by SR-PCIM to:
    - Determine EP is SR-IOV capable
    - Enable/disable and configure the EP's SR-IOV capability
  - ✓ MR-IOV Capability - used by MR-PCIM to:
    - Determine EP is MR-IOV capable (implies SR-IOV capable as well)
    - Enable/disable and configure the EP's MR-IOV capability
- Both new capabilities are located in the PCIe Extended Configuration Space (offset 256 or greater).

Overview of Function Types

Function type	Who owns it for Single Root	Who owns it for Multiple Root
Base function		MR-PCIM
Physical function	SR-PCIM	SR-PCIM
Virtual function	System Image	System Image

- MR-PCIM uses **Base Function** to discover and set-up MR-IOV and SR-IOV capabilities, such as:
  - ✓ Assign PFs to VHs
  - ✓ Assign VFs to PFs
- SR-PCIM uses **Physical Function** to discover and set-up SR-IOV capabilities, such as:
  - ✓ Assign VFs to SIs
  - ✓ Configure BARs for VFs

MR EP Initialization Overview



- MR-PCIM discovers and enumerates PCIe components.
- MR-PCIM discovers MR-IOV EP by reading MR-IOV Capabilities Register in BF_0.
- MR-PCIM assigns resources to RCs by writing MR-IOV Capabilities Register in BF_0.
- After MR-PCIM configures the MR-IOV PCIe Switches and the MR-IOV EPs, each RC has assigned its own Virtual Hierarchy (VH).
  - ✓ The SR-PCIM in each RC now performs its Root Topology Initialization Flow.

How IOV Capabilities Map to Function Types

Function type	Single Root	Multiple Root
Base function		PCI Configuration Space PCIe Extended Config Space MR-IOV Capabilities
Physical function	PCI Configuration Space PCIe Extended Config Space SR-IOV Capabilities	PCI Configuration Space PCIe Extended Config Space SR-IOV Capabilities
Virtual function	PCI Configuration Space PCIe Extended Config Space	PCI Configuration Space PCIe Extended Config Space

Multi-Root Resource Discovery and Allocation

New PCIe IOV Capabilities register set

MR-IOV Capabilities Structure

MR-IOV Capabilities

- Same EP architecture approaches as covered in Single-Root:
  - ✓ EPs with a single function type
  - ✓ EPs with multiple function types
  - ✓ EPs with dependent function types - *not covered in this deck*

- Please note:
  - ✓ Single-bit fields are fully defined.
  - ✓ Size of multi-bit fields has not been set … yet
    (WG still discussing scaling of the field over time).
  - ✓ Following slides represent our current direction…
    … which is subject to change as complete the specification.

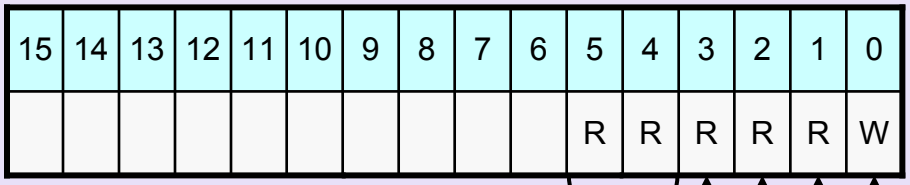
MR-IOV Capabilities

MR-IOV PCIe Extended Configuration Space		
MR-IOV Capabilities	Next Cap Ptr	Version
RID Space Allocation Register(s) (TBD)		
Mapping Dependent Fields (covered on the slides that follow)		

- IOV Capabilities; Next Capabilities Pointer; Version
- RID Space Allocation Register(s) to be defined by RID sub-team
- Mapping Dependent Fields - Contents differ depending on which EP architecture approach is used:
  - ✓ EPs with a single PF type
  - ✓ EPs with multiple (independent) PF types
  - ✓ EPs with dependent PF combinations - *not covered in this deck*

MR-IOV Capabilities Register

R = Read only
W = Write/Read



Mapping of PFs and VFs to VHs	bits
EPs with one PF type	00b
EPs with multiple PF types	01b
EPs with multiple dependent PF types	10b
Reserved	11b

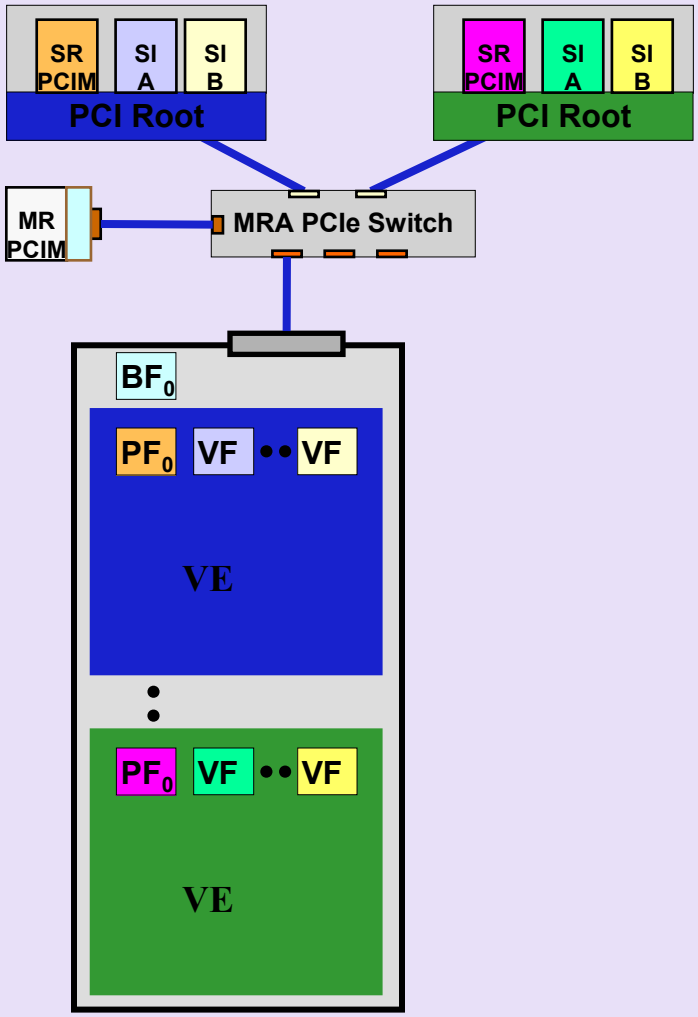
VF Assignment Mechanism (set if MR-PCIM assigned, reset if EP assigned)

VH Migration Supported (set if supported)

VF Migration Supported (set if supported)

VH Enable/Disable (set by MR-PCM)

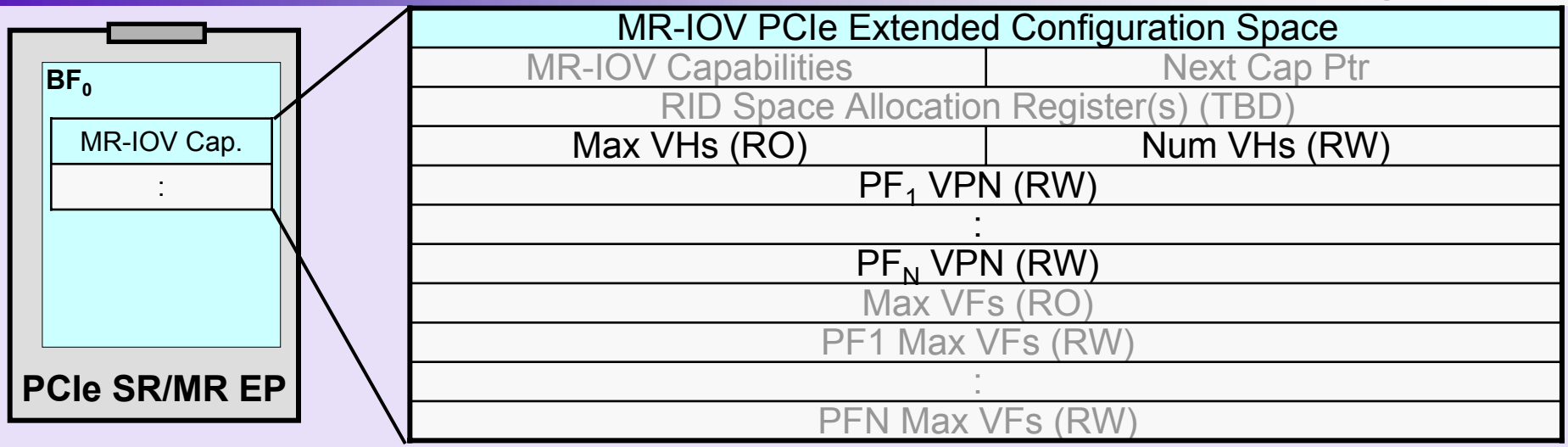
EPs with one function type



Overview of MR-PCIM responsibilities, MR-PCIM:

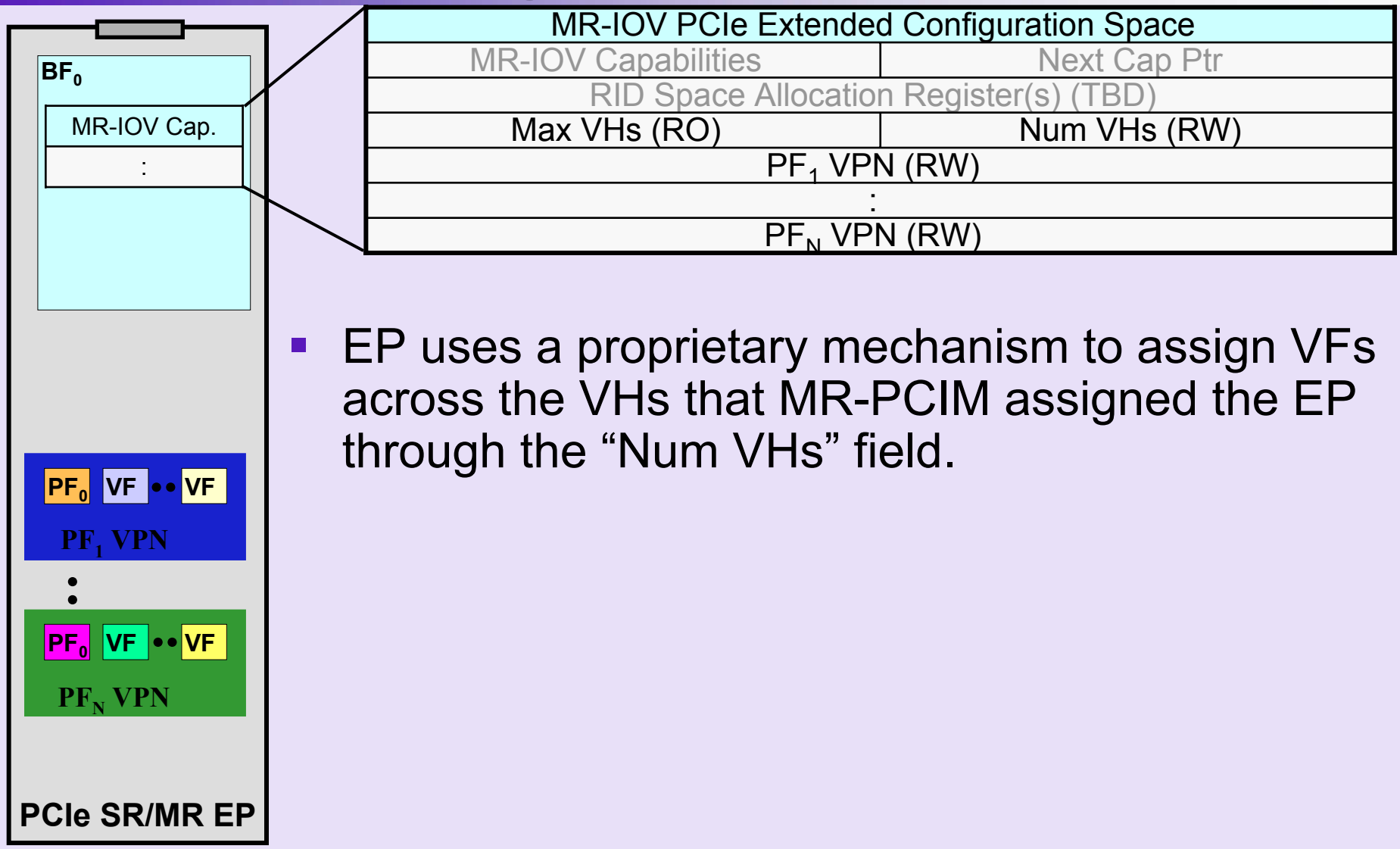
- Assigns the number of Virtual Hierarchies for the EP.
- Assigns PFs to VHs.
- Sets Max VFs for each PF.
  - ✓Two options:
    - MR-PCIM assigns Max VFs
    - EP uses proprietary assignment mechanism
- After above assignments are done, SR-PCIM can determine the maximum number of VFs associated with a PF through the SR-IOV “Max VFs” field.

VH Assignment for EPs with one function type



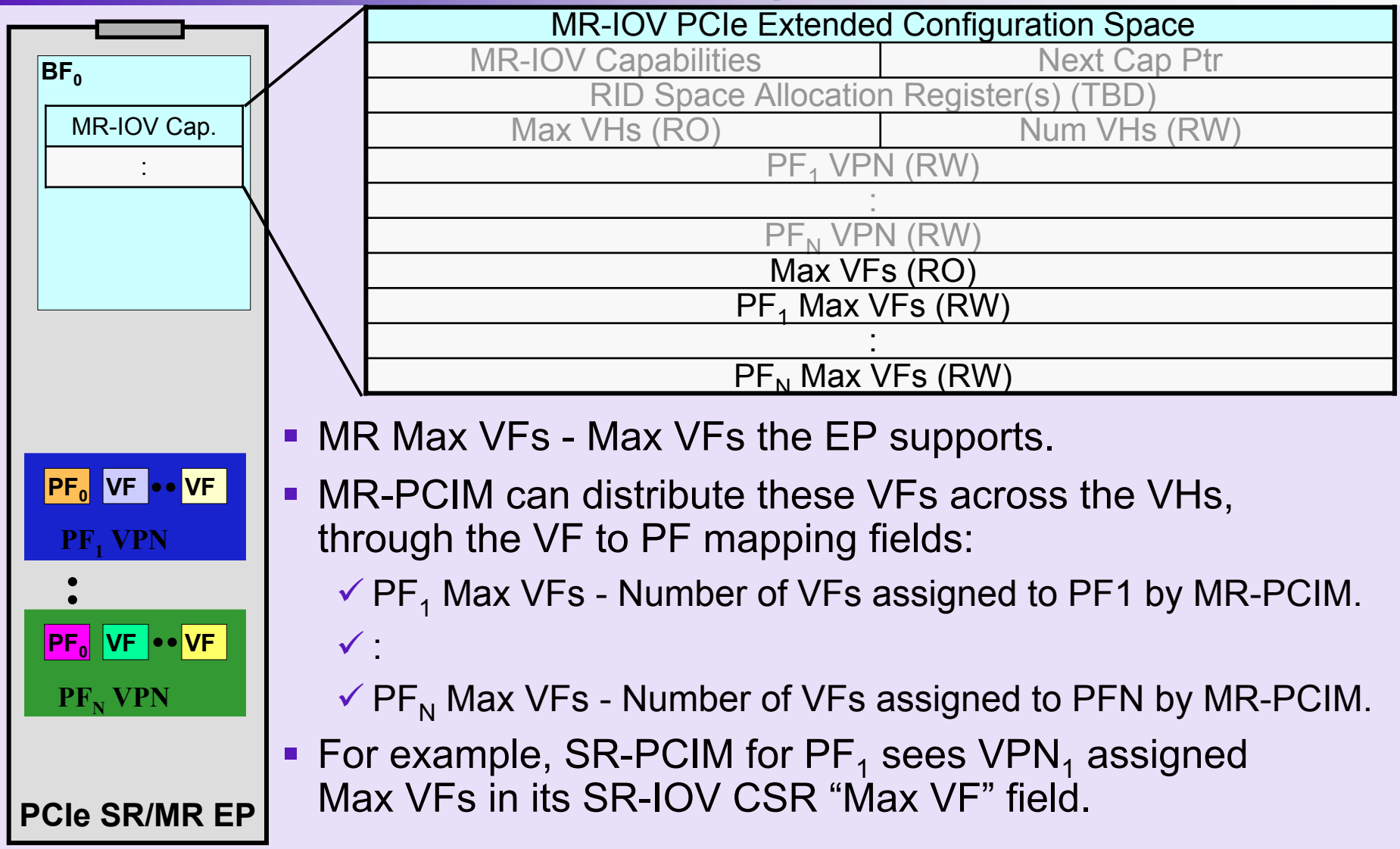
- Max VHs - Maximum number of Virtual Hierarchies the EP supports.
  - ✓ Used by MR-PCIM to determine number of VHs.
- Num VHs - The number of VHs assigned to the EP by MR-PCIM (Num VHs => Max VHs).
  - ✓ Used by MR-PCIM to assign the actual number of VHs EP will use.
- Mapping of Physical Functions to Virtual Hierarchies:
  - ✓ PF_1 VPN - Virtual Plane Number assigned to PF1 by MR-PCIM.
  - ✓ :
  - ✓ PF_N VPN - Virtual Plane Number assigned to PFN by MR-PCIM.

EP Assigned VFs



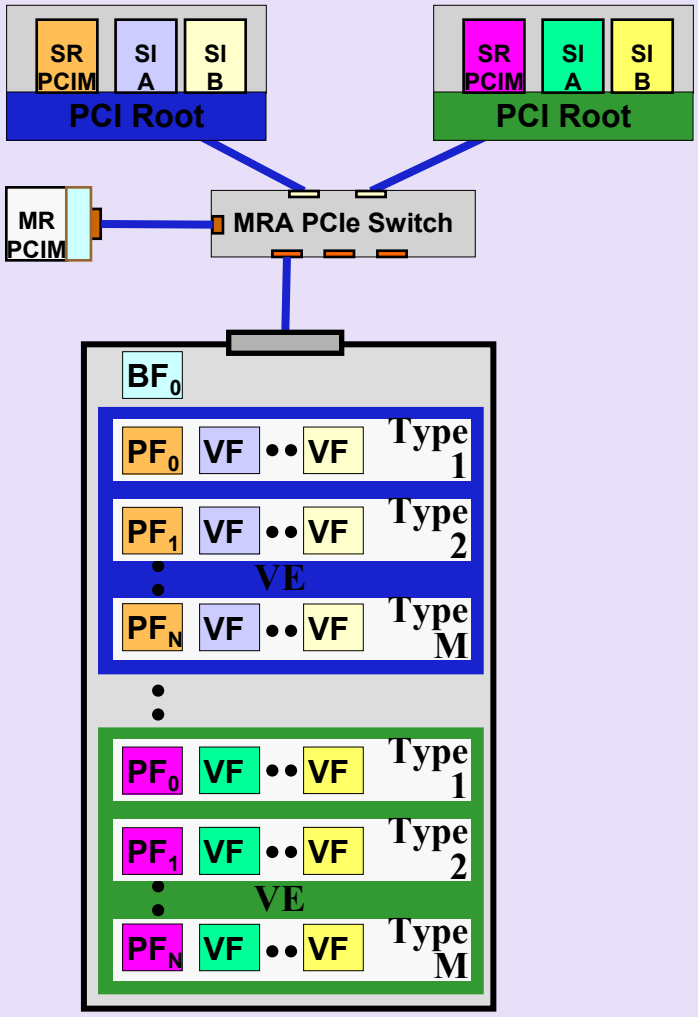
- EP uses a proprietary mechanism to assign VFs across the VHs that MR-PCIM assigned the EP through the “Num VHs” field.

MR-PCIM Assigned VFs



- MR Max VFs - Max VFs the EP supports.
- MR-PCIM can distribute these VFs across the VHs, through the VF to PF mapping fields:
  - ✓ PF_1 Max VFs - Number of VFs assigned to PF1 by MR-PCIM.
  - ✓ :
  - ✓ PF_N Max VFs - Number of VFs assigned to PFN by MR-PCIM.
- For example, SR-PCIM for PF_1 sees VPN_1 assigned Max VFs in its SR-IOV CSR “Max VF” field.

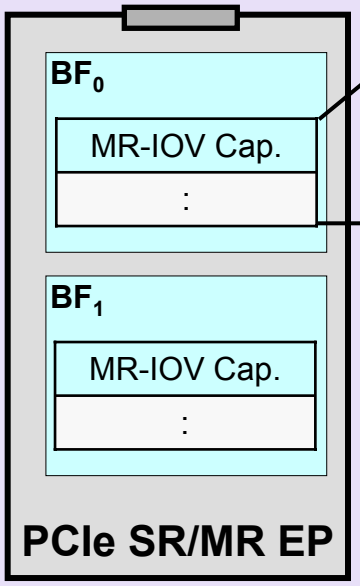
EPs with multiple function types



Overview of MR-PCIM responsibilities

- For each PF type, MR-PCIM
  - ✓Assigns the number of VHs.
    - MR-PCIM assigns PFs to VHs.
  - ✓Sets Max VFs for each PF.
    - Two options:
      - MR-PCIM assigns Max VFs
      - EP uses proprietary assignment mechanism
- After above assignments are done, SR-PCIM can determine the maximum number of VFs associated with each PF through the SR-IOV “Max VFs” field.

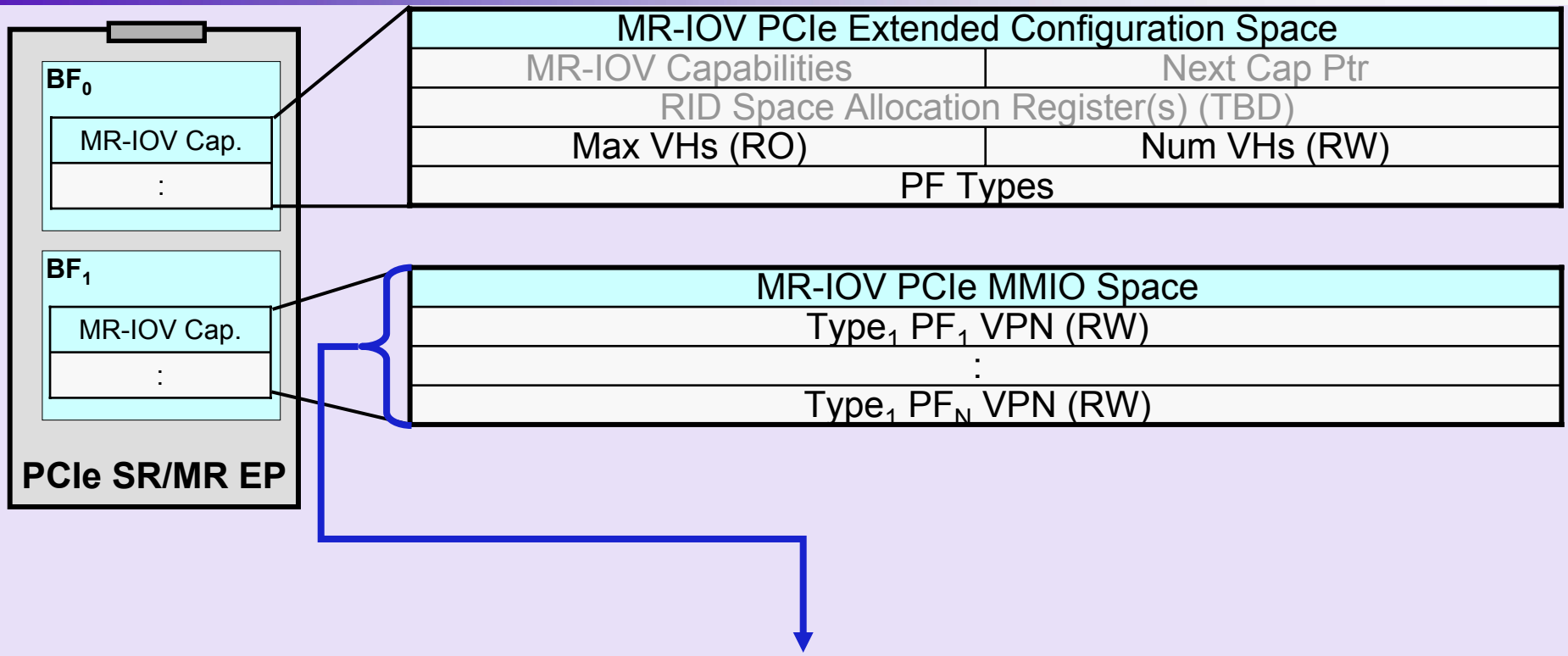
VH and PF Type Discovery for EPs with multiple function types



MR-IOV PCIe Extended Configuration Space	
MR-IOV Capabilities	Next Cap Ptr
RID Space Allocation Register(s) (TBD)	
Max VHs (RO)	Num VHs (RW)
PF Types	

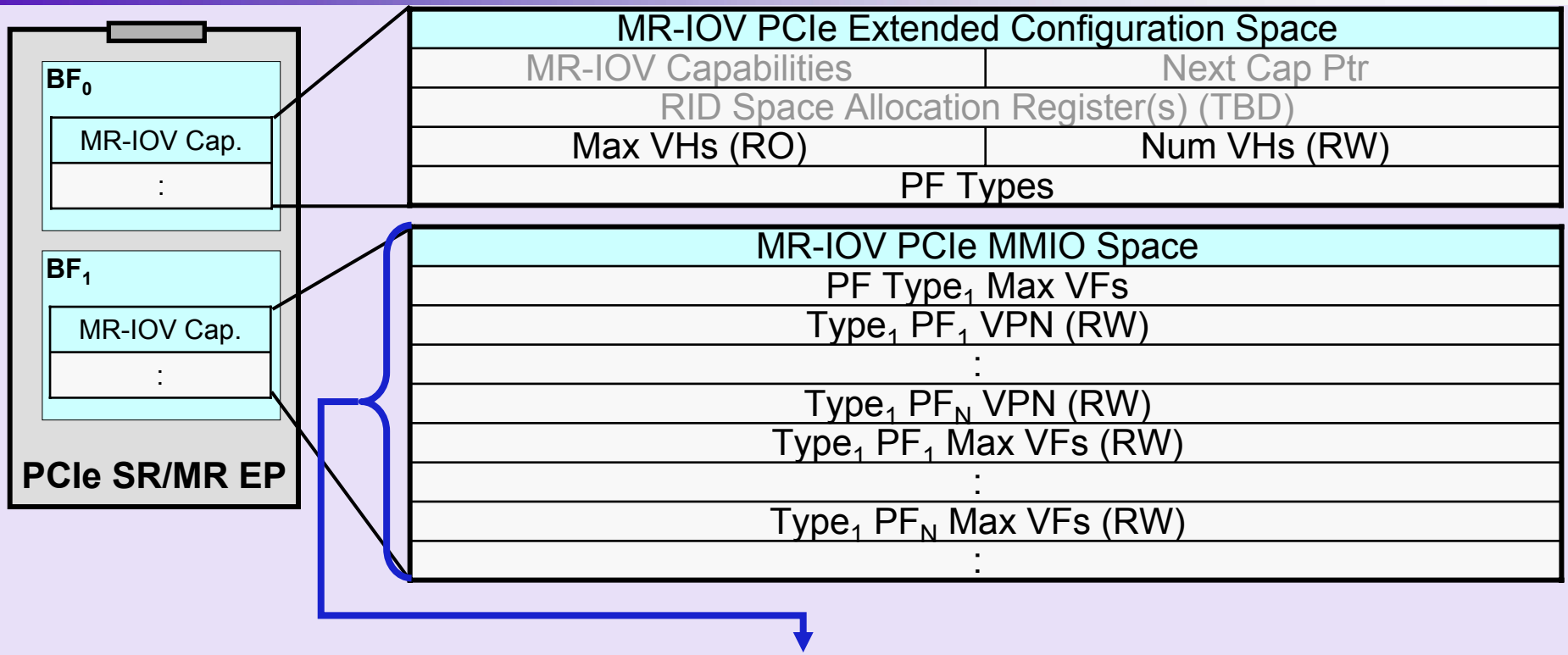
- Two configuration areas used by MR-PCIM:
- **BF_0** - Contains fields that define the number of VHs and PF Types supported by the EP:
  - ✓ Max VHs - Maximum number of VHs the EP supports.
    - Used by MR-PCIM to determine number of VHs.
  - ✓ Num VHs - The number of VHs assigned to the EP by MR-PCIM (Num VHs => Max VHs).
    - Used by MR-PCIM to assign the actual number of VHs EP will use.
  - ✓ PF Types - the number of PF Types supported by the EP.
- **BF_1** - Contains fields that describe each PF Type and are used by MR-PCIM to configure each PF Type.
  - ✓ All BF1 CSR fields can be used for MR-PCIM capabilities.

EP based VF assignment for EPs with multiple function types



- For each PF Type, the following fields are used by MR-PCIM:
  - ✓ $Type_1$ PF_1 VPN - Virtual Plane Number assigned to $Type_1$ PF_1 by MR-PCIM
- Note: EP assigns VFs to each PF Type using a proprietary mechanism.

MR-PCIM based VF assignment for EPs with multiple function types



- For each PF Type, the following fields are used by MR-PCIM:
  - ✓ PF $Type_1$ Max VFs - Max VFs the EP supports, used by MR-PCIM to determine how many VFs it can distribute across the $Type_1$ PFs.
  - ✓ $Type_1$ PF_1 VPN - Virtual Plane Number assigned to $Type_1$ PF_1 by MR-PCIM
  - ✓ $Type_1$ PF_1 Max VFs - Number of VFs assigned to $Type_1$ PF_1 by MR-PCIM.

Questions

Thank you for attending.

For more information please go to
www.pcisig.com

Multi-Root Resource Discovery and Allocation

Michael Krause (HP, co-chair)

Renato Recio (IBM, co-chair)