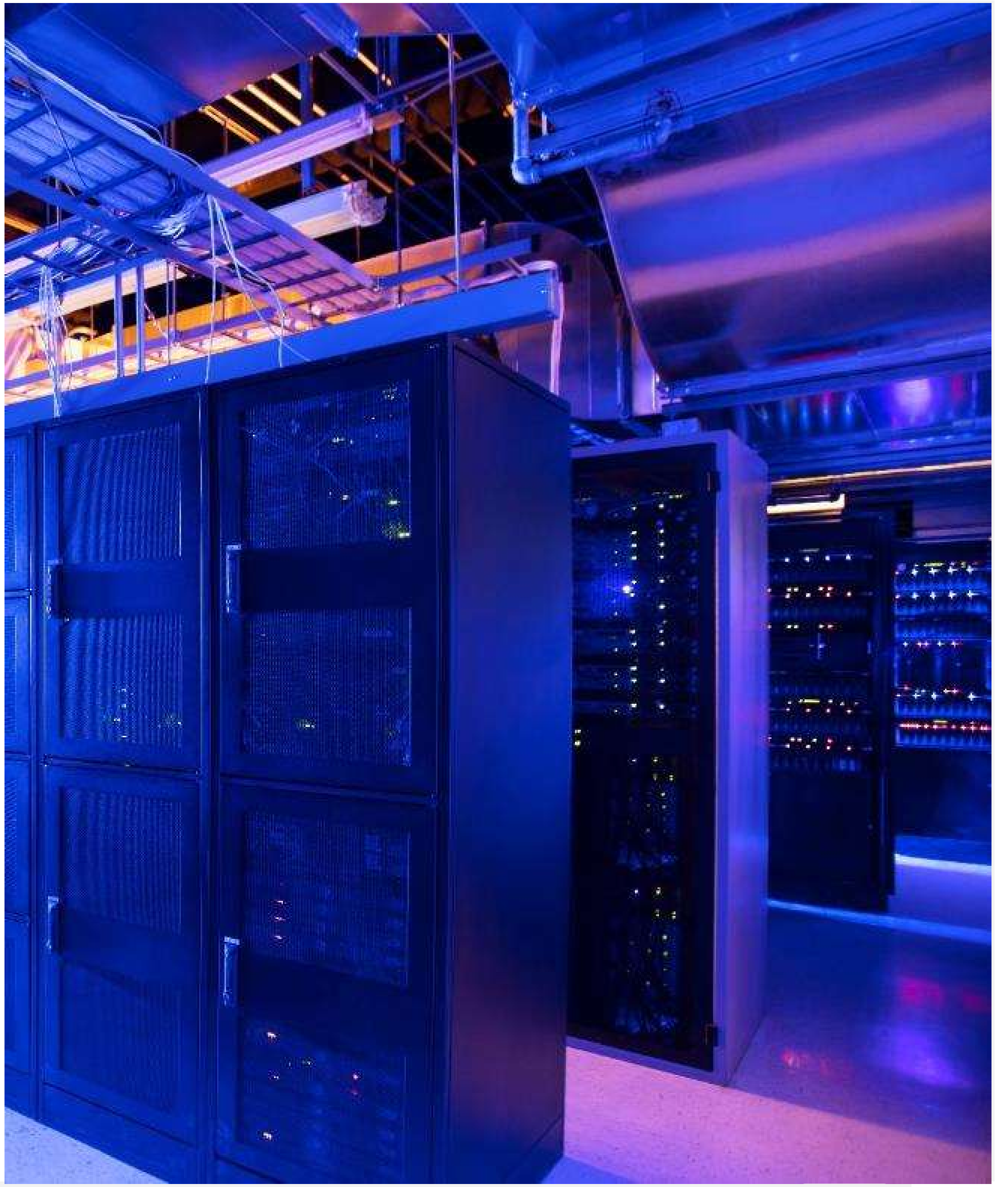
Live update of RDMA DMA memory with the help of eBPF

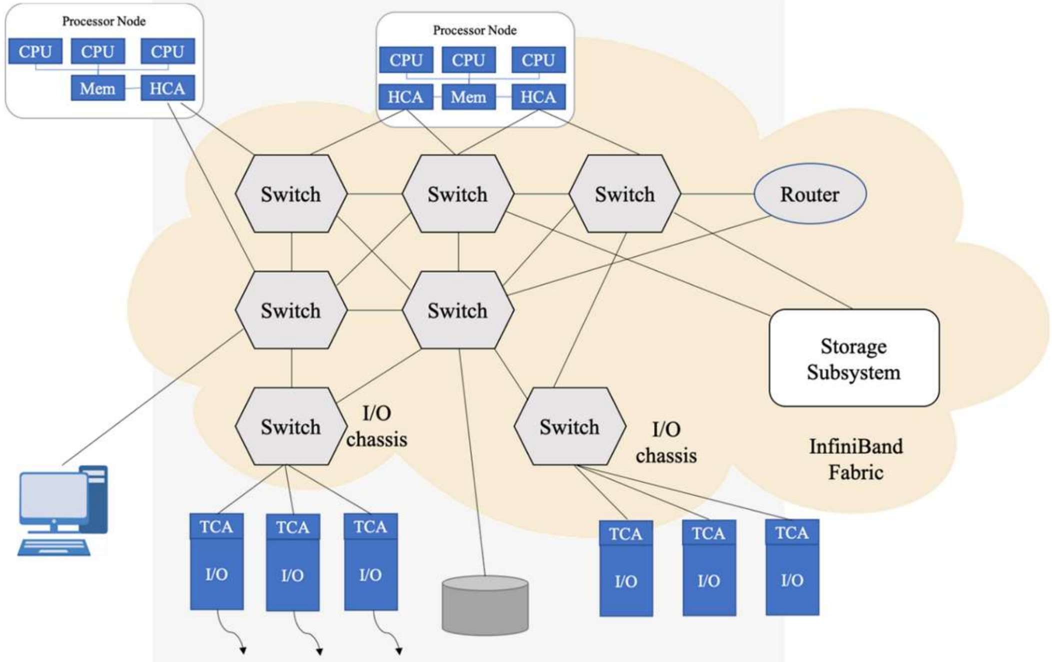
Zhu Yanjun Aug. 6, 2025

Agenda

- Background(Storage+IB network)
- History(eBPF/RDMA/DMA)
- Current architecture
- Problems
- Solutions
- Future



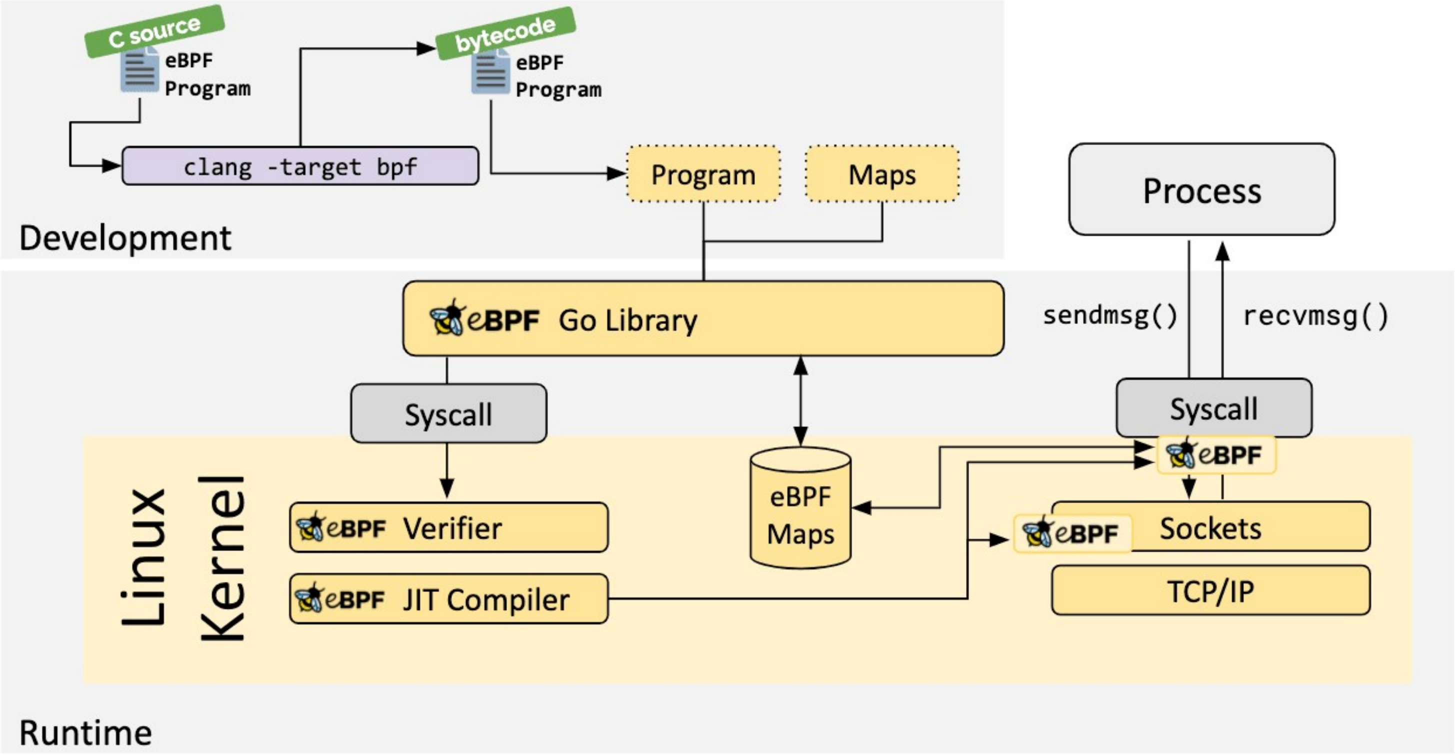
Background



A **storage + RDMA network** combines storage systems (such as NVMe drives or storage servers) with an RDMA-capable network fabric (such as **InfiniBand**, **RoCE** – RDMA over Converged Ethernet). It allows applications to perform high-speed, low-latency access to remote storage over the network.

eBPF Introduction

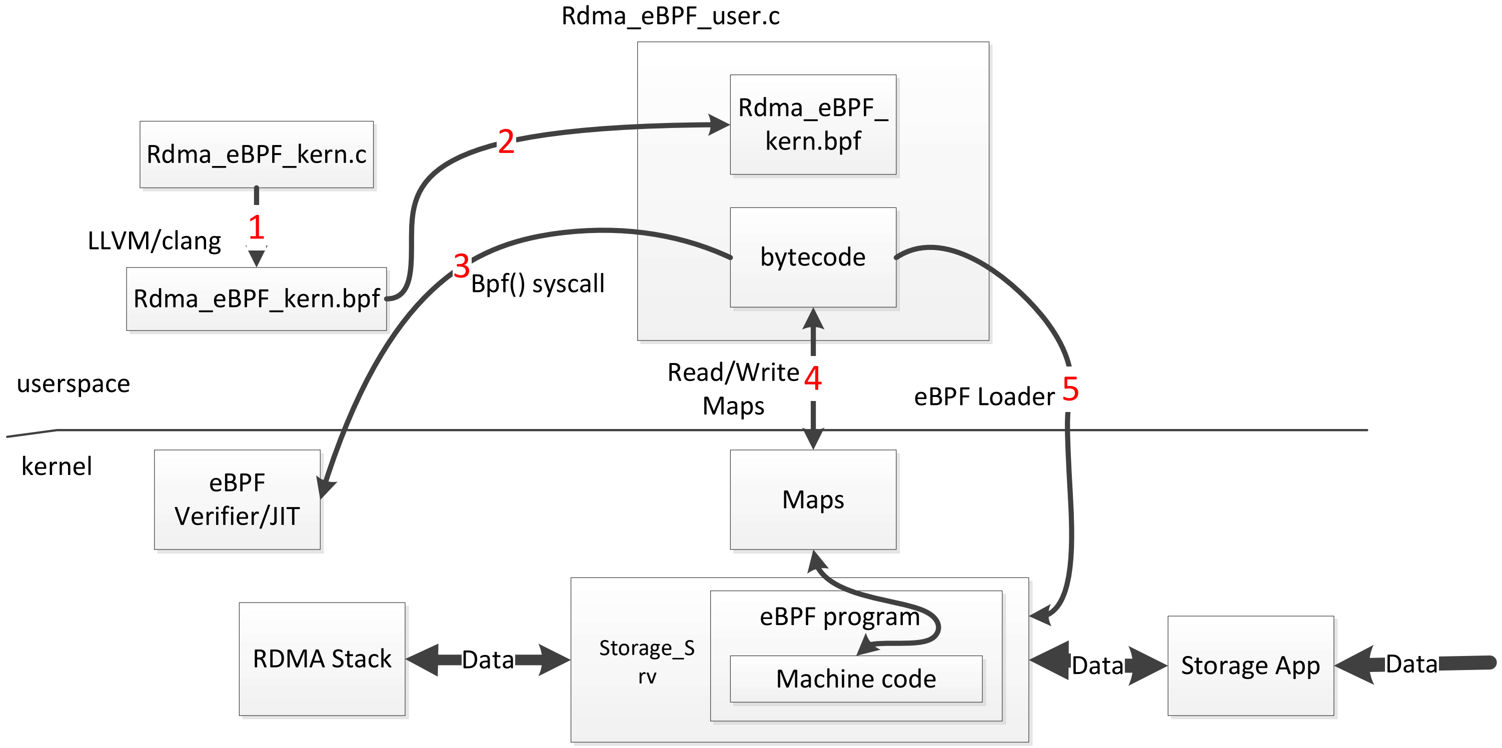
eBPF is a revolutionary technology with origins in the Linux kernel that can run sandboxed programs in a privileged context such as the operating system kernel. It is used to safely and efficiently extend the capabilities of the kernel without requiring to change kernel source code or load kernel modules.



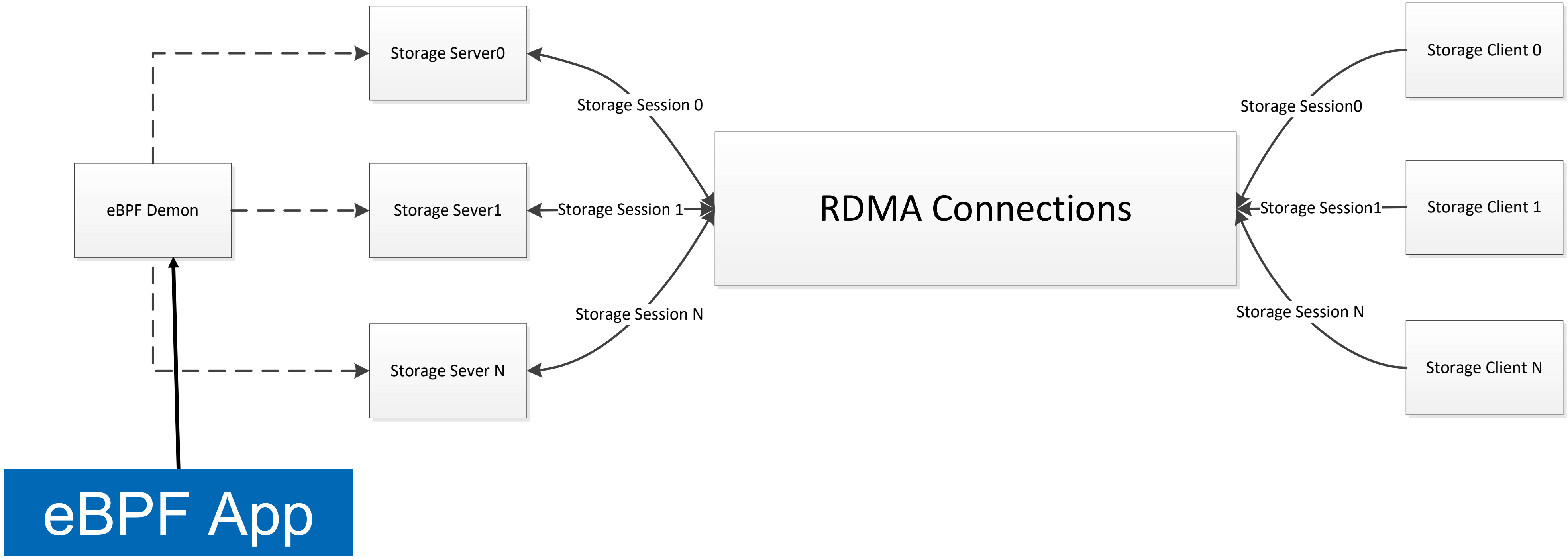
eBPF Architecture

User Space:
Path1 and 2: eBPF_Daemon user space c files are compiled into a bin file;
Path 3: This bin is checked by eBPF verifier.

Kernel:
Path 5: This bin file is attached to Storage Srv.
Path 4: Mapping rdma data between kernel and user space.



Usage of eBPF



RDMA Introduction

1. RDMA is short for Remote Direct Memory Access.
   It is A (relatively) new method for interconnecting platforms in high-speed networks that overcomes many of the difficulties encountered with traditional networks such as TCP/IP over Ethernet.

   1) new standards

   2) new protocols

   3) new hardware interface cards and switches

   4) new software

2. RDMA implementations

   Infiniband and RoCEv1/v2

Infiniband and RoCE

1. Infiniband
   Infiniband implementation needs special devices, for example HCA (Host Channel Adapter), IB switch and special cables.

2. RoCE

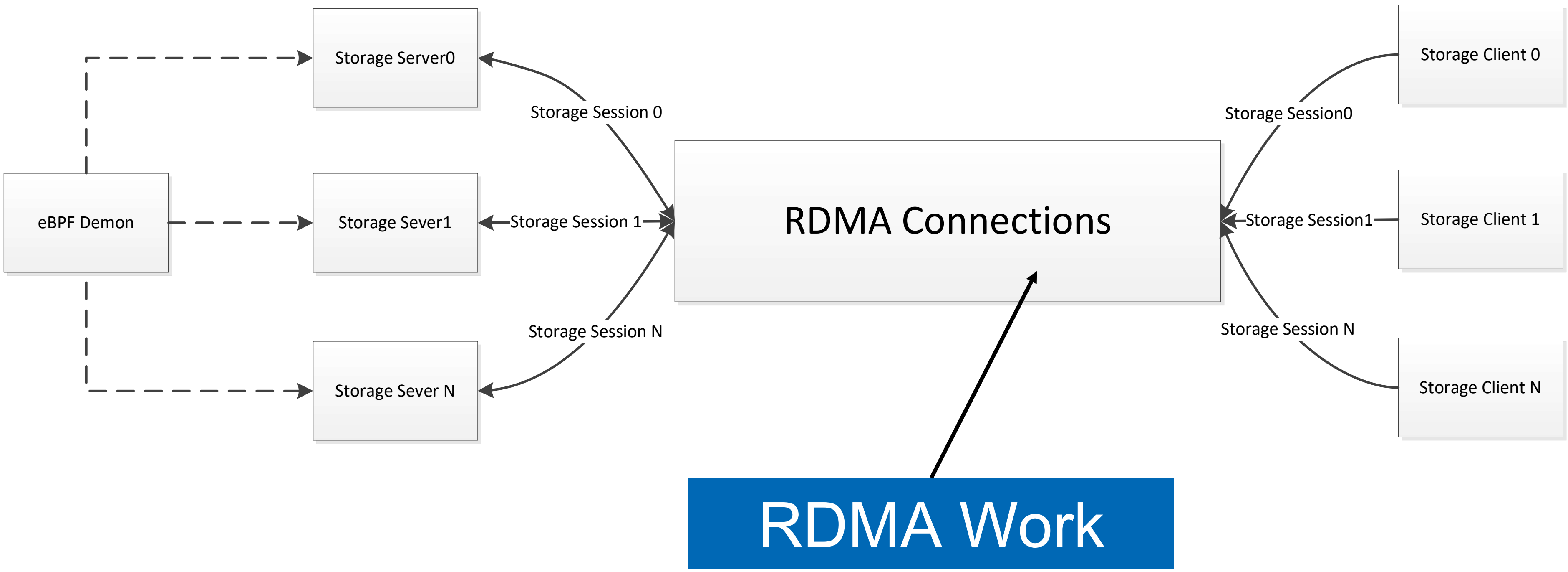
   1) RoCEv1 and RoCEv2
   2) RoCEv1 is based on TCP/IP ethernet layer. It only works in LAN (Local Area Network). It is not popular.
   3) RoCEv2 is based on UDP. It is popular in public/private Cloud and Data Center.

RoCEv2

1. RoCEv2 is based on UDP/IPv4 or UDP/IPv6. The target udp port 4791 is reserved for RoCEv2. Because RoCEv2 can be routable, RoCEv2 sometimes can be called Routable RoCE or RRoCE.

2. RoCEv2 is supported in Mellanox OFED2.3 or above. In Linux 4.5, RoCEv2 is supported. Intel E810 also supports RoCEv2.

RDMA in Storage Network



New DMA Mapping APIs

The API is split up into parts:

- **Allocate IOVA space**:

  To do any pre-allocation required. This is done based on the caller supplying some details about how much IOMMU address space it would need in worst case.

- **Map and unmap relevant structures to pre-allocated IOVA space**:

  Perform the actual mapping into the pre-allocated IOVA. This is very similar to dma_map_page().

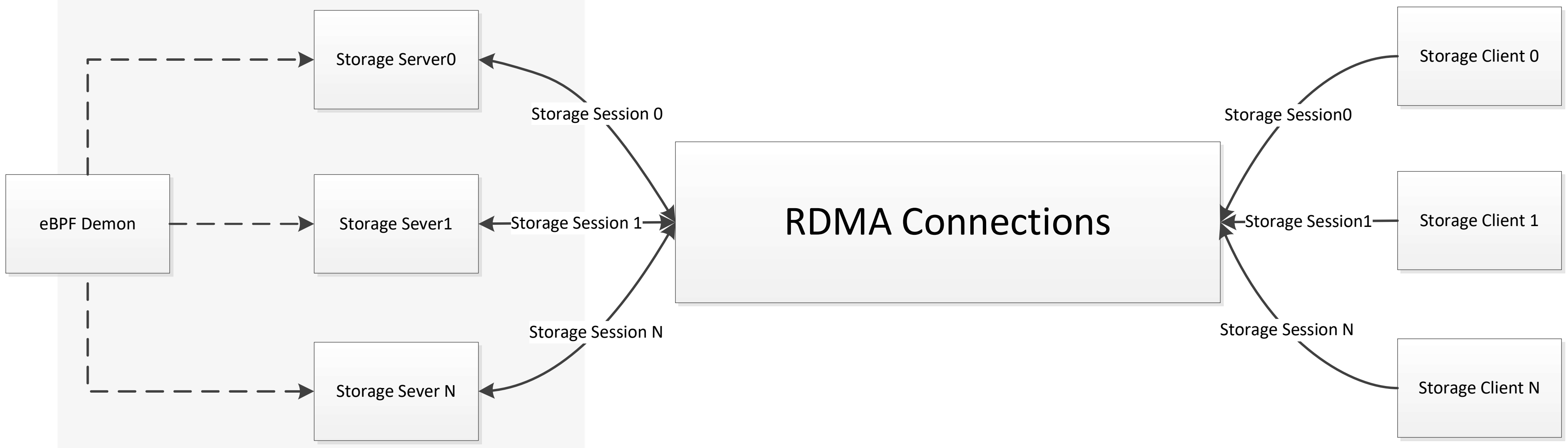
New DMA mapping examples

1. RDMA ODP is an example of "SVA mirroring" using HMM that needs to dynamically map/unmap large numbers of single pages.

2. VFIO PCI live migration code is building a very large "page list“ for the device. Instead of allocating a scatter list entry per allocated page it can just allocate an array of 'struct page *', saving a large amount of memory.

3. NVMe PCI demonstrates how a BIO can be converted to a HW scatter list without having to allocate then populate an intermediate SG table.

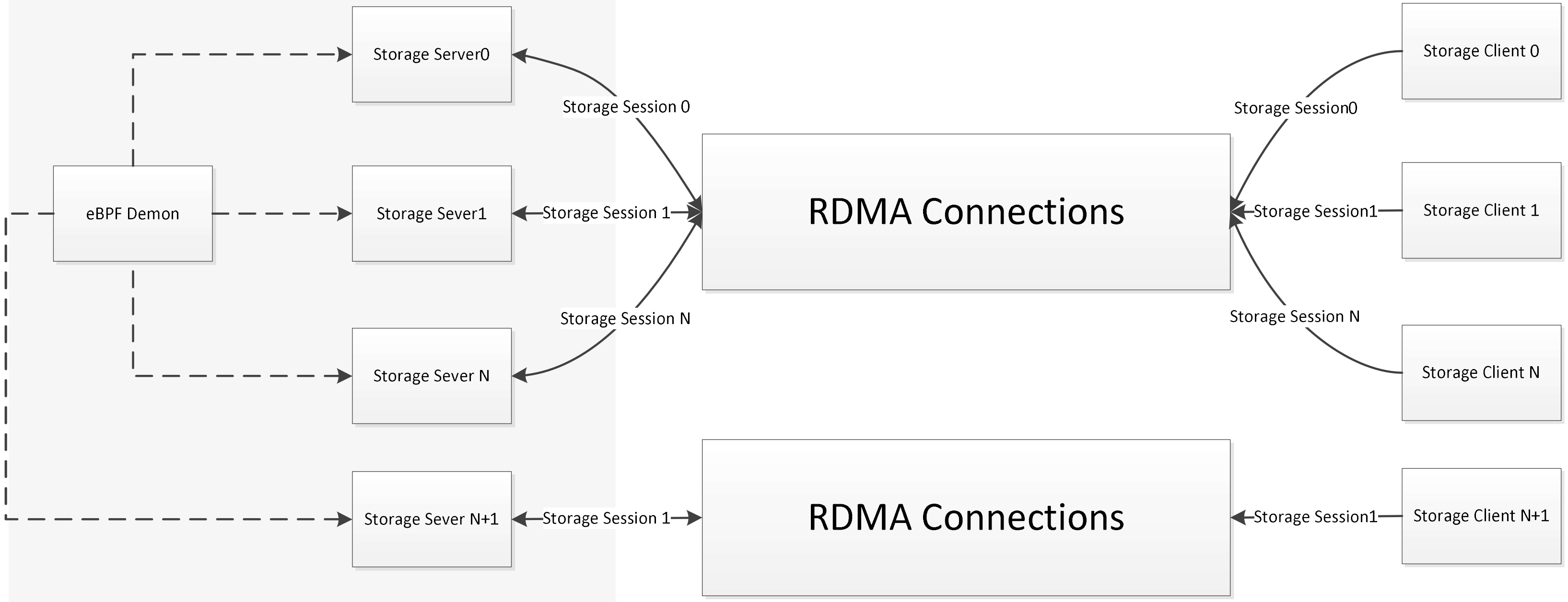
The description of the problem



In this scenario, multiple storage sessions share a single RDMA connection, and the number of storage sessions increases steadily over time.When a new RDMA connection is initially created, it is used by only a few storage sessions, so performance is typically very good. However, as more and more storage sessions begin to use the same RDMA connection, they gradually consume its entire capacity. Once the number of storage sessions reaches the connection's upper limit, adding another storage session degrades overall performance — for example, by increasing latency and reducing bandwidth.

Solution 1:

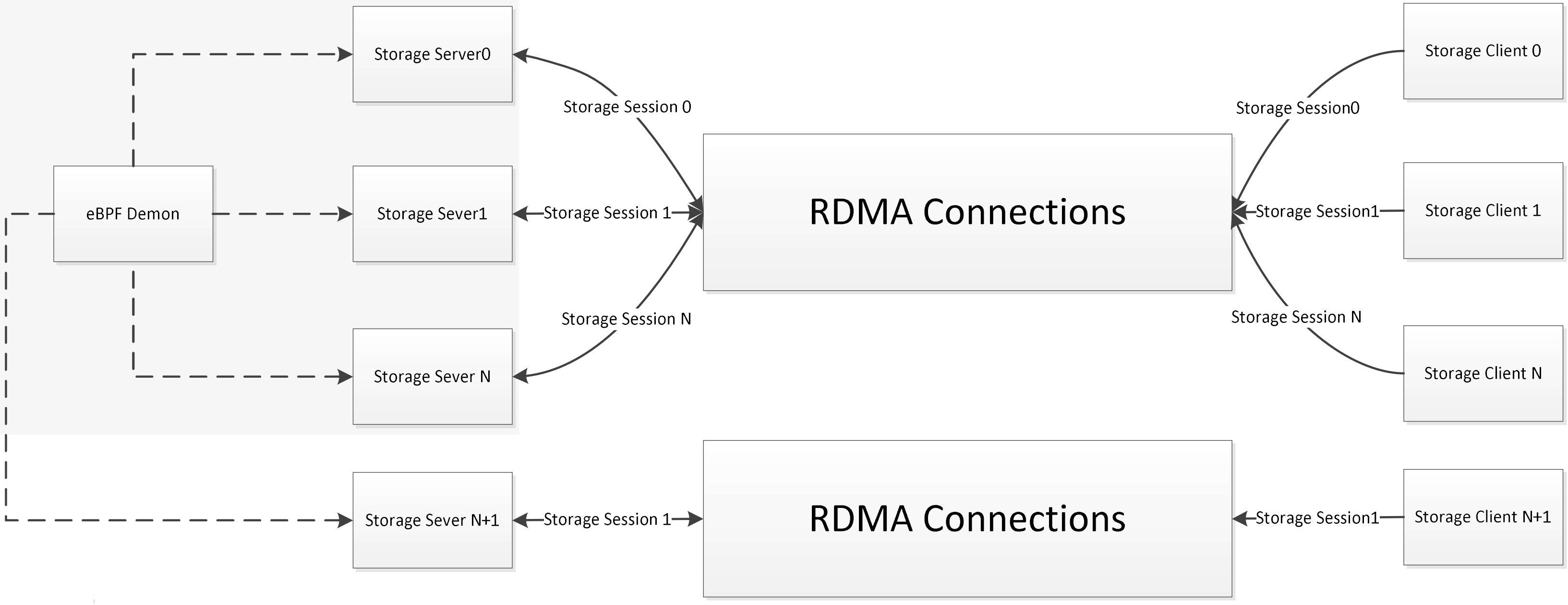
The root cause is insufficient DMA memory to support so many storage sessions. An eBPF application can be used to monitor the performance (latency and bandwidth) of the storage sessions. If performance degrades, it indicates that the number of storage sessions has reached the upper limit of the RDMA connection. In this case, a new RDMA connection should be created, and the new storage session should use it.



Solution 1 analysis

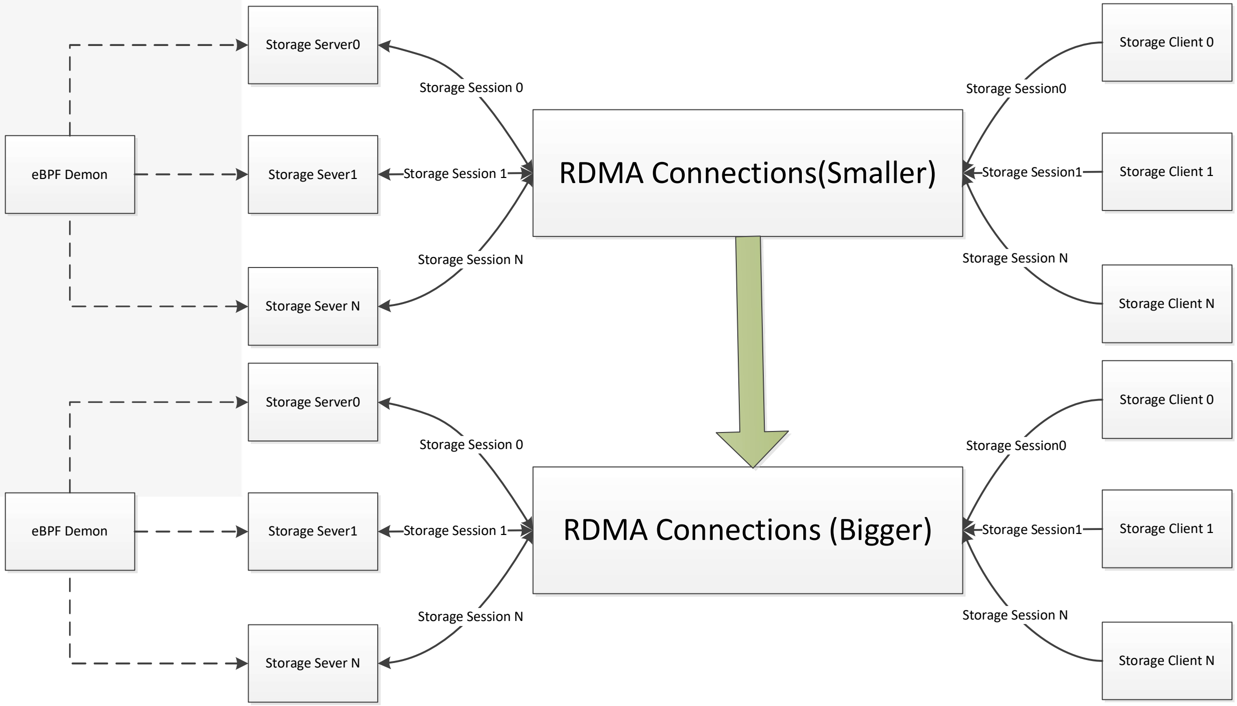
1. Pros: Easy to implement, create a new RDMA connection and let the new Storage sessions use this new RDMA connection.

2. Cons: this solution requires additional resources to manage multiple RDMA connections. Having too many RDMA connections can consume excessive resources, leading to inefficiencies on the host.



Solution 2:

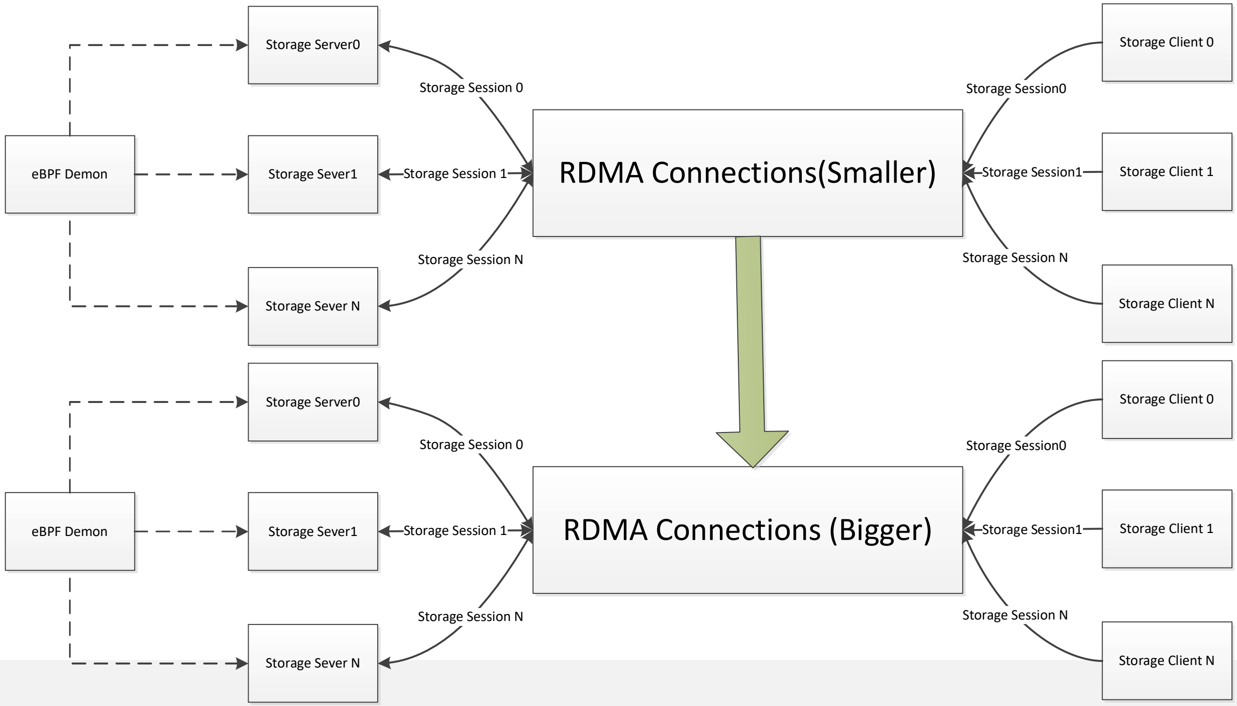
To avoid excessive resource usage from managing too many RDMA connections, eBPF will continue monitoring the performance of storage sessions. As more storage sessions are added, and the number reaches the RDMA connection's upper limit, eBPF will create a new RDMA connection with a larger DMA memory allocation. Then, it will gradually migrate all storage sessions from the old, smaller-DMA RDMA connection to the new one.



Solution 2 analysis

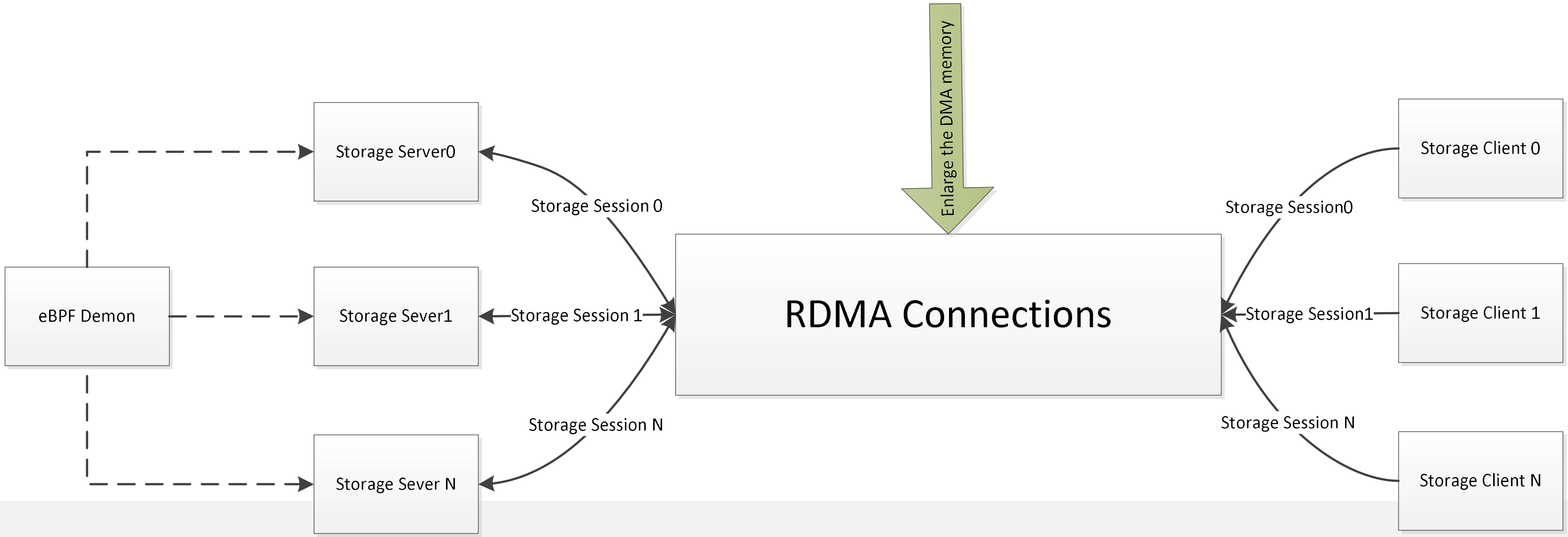
Pros: avoid excessive resource usage from managing too many RDMA connections

Cons: This solution requires sufficient system memory to maintain both RDMA connections during the transition. There is a risk of memory allocation failure when creating the new RDMA connection. Additionally, migrating storage sessions involves restarting them, which can impact the overall system behavior.



Solution 3

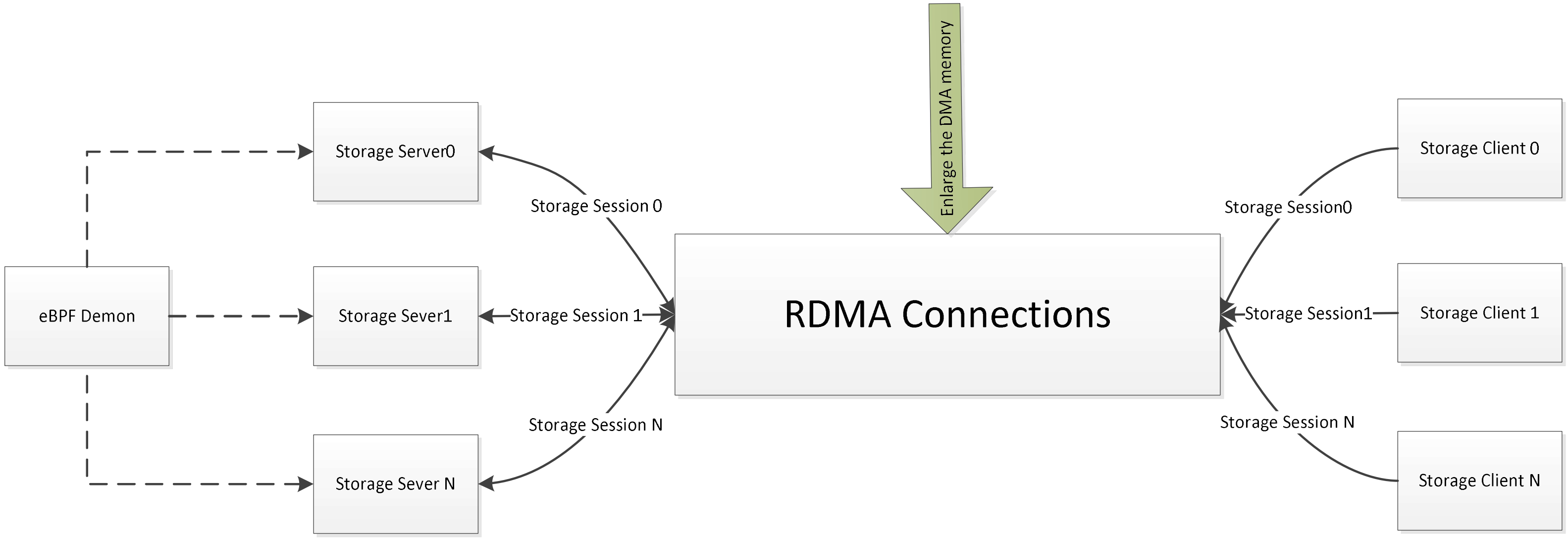
This solution is based on Leon's new patch series, 'Provide a new two-step DMA mapping API.' At the beginning, a large IOVA window is allocated from the IOMMU, but only a portion of it—such as one-quarter—is initially mapped to physical DMA memory. When the number of storage sessions approaches the upper limit of the RDMA connection, eBPF can allocate additional physical DMA memory and map it to another quarter of the pre-allocated IOVA window. This effectively extends the available DMA memory, improving the performance of the storage sessions. As more storage sessions are added and the new limit is reached, more DMA memory can be mapped, further extending the capacity. Ultimately, the maximum number of storage sessions is limited by the total DMA memory available on the host. Once all DMA memory is exhausted, no further storage sessions can be accepted by the RDMA connection.



Solution 3 Analysis

Procs: Avoid creating new RDMA sessions.

Cons: Operate on the same RDMA connection to enlarge the DMA memory. Need to check the performance of Storage sessions.



Thanks for your attention!

- Questions