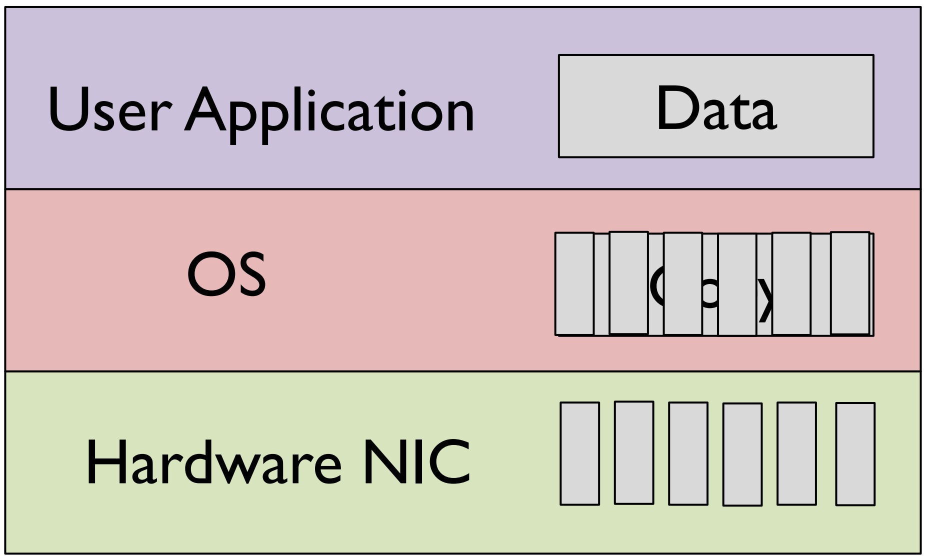
RDMA

ECE/CS598HPN

Radhika Mittal

Traditional Network Stack

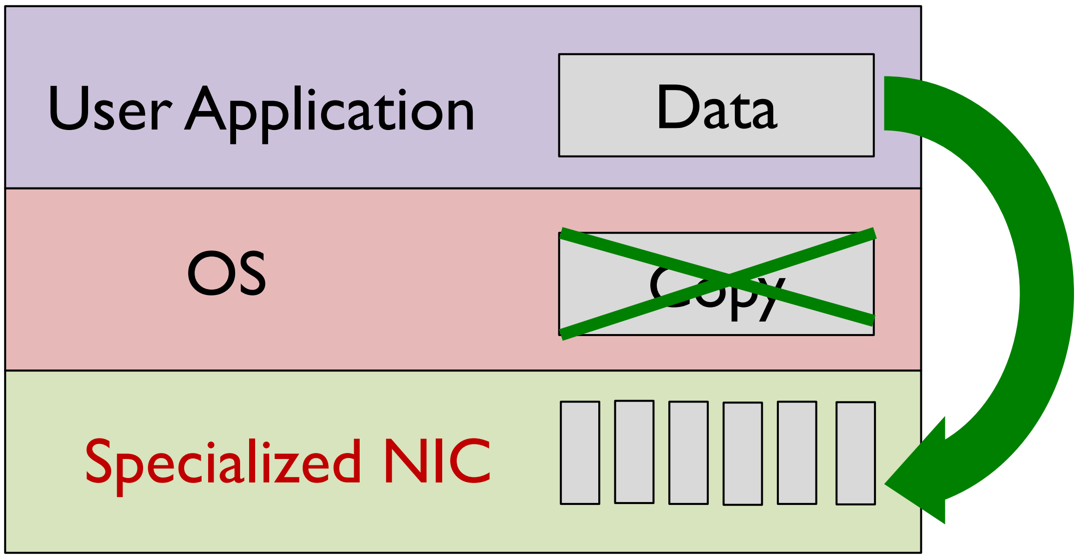


Packet processing in OS incurs high latency, cannot support high throughput, and leads to high CPU utilization.

Not acceptable in today's datacenters:

- few microseconds of latency
- tens to hundred Gbps bandwidth
- cpu = $$$

Remote Direct Memory Access

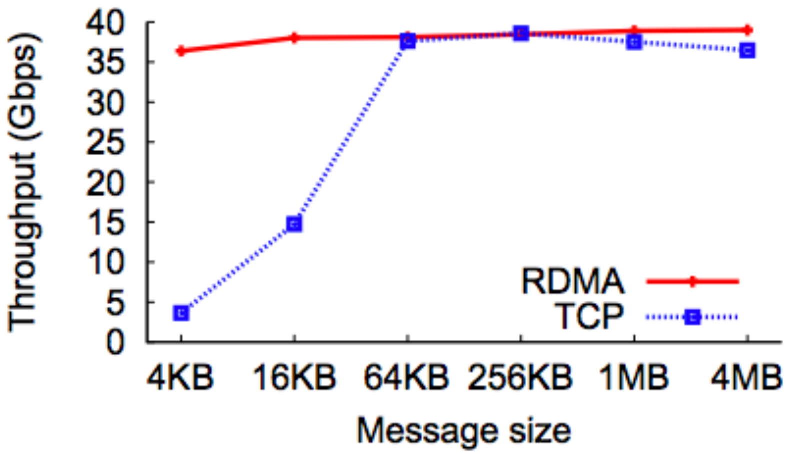


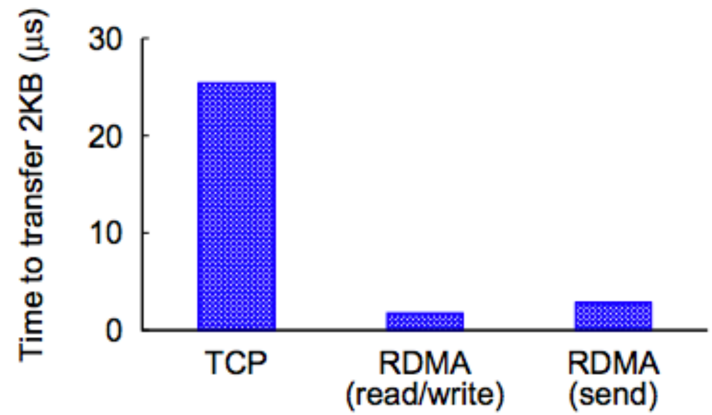
- Protocol offload
- Zero-copy
- OS by-pass

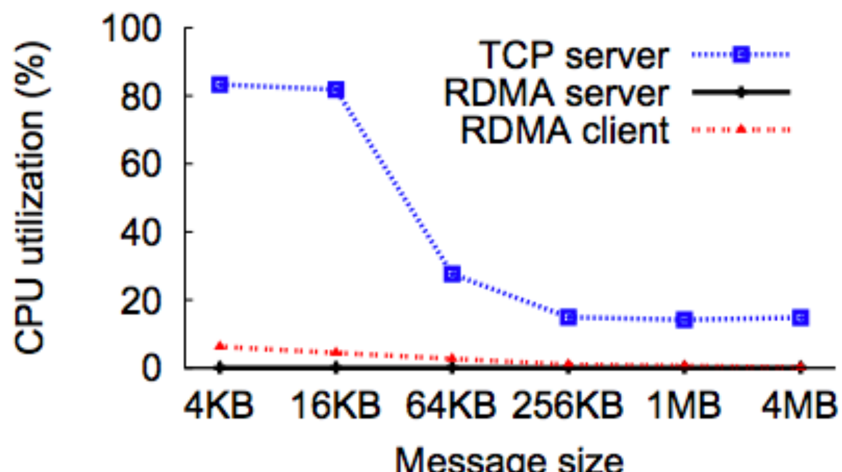
Traditionally used in Infiniband clusters for HPC.

Achieves low latency, high throughput and negligible CPU utilization.

Performance Benefits of RDMA







From *"Congestion Control for Large-Scale RDMA Deployments"*, Zhu et. al., SIGCOMM 2015

RDMA usecases in datacenters

- Distributed storage:
  - Distributed key-value stores
    - *Pilaf (ATC'13), FaRM (NSDI'14, SOSP'15), HERD (SIGCOMM'14), FASST(OSDI'16),…*
  - Distributed file systems
  - NVMe over Fabric
- Applications requiring low latency (e.g. search queries)
- GPU Direct communication (by-pass CPU): ML training
- Other proposals
  - Resource disaggregation *(OSDI'16)*, Remote swapping *(NSDI'17)*,…
  - use RDMA interface to support arbitrary computation without involving server CPU (NSDI'22)

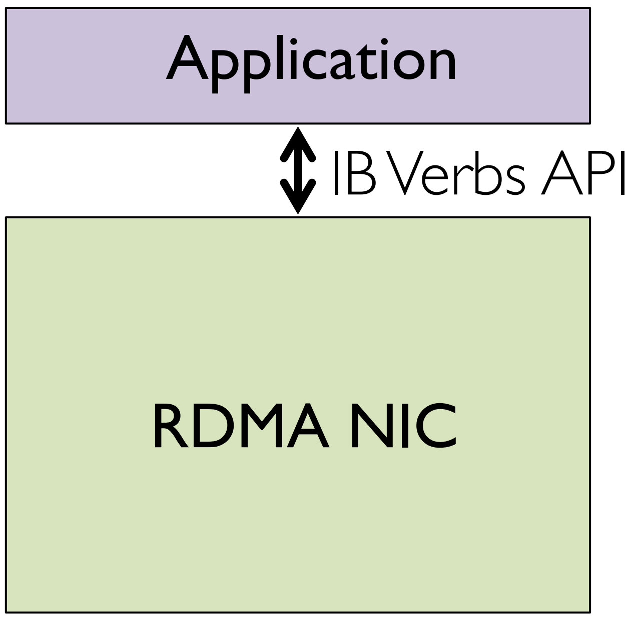
Focus of today's lecture

- Overview of RDMA
- RDMA deployment in today's datacenters

Focus of today's lecture

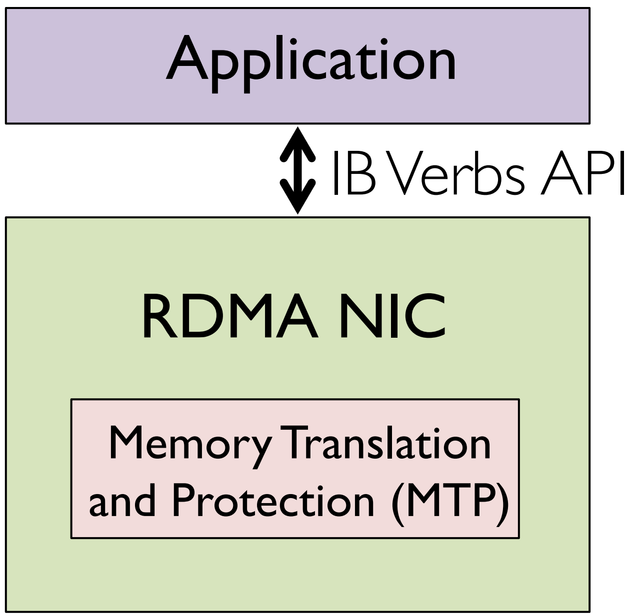
- Overview of RDMA
- RDMA deployment in today's datacenters

RDMA Overview and Components



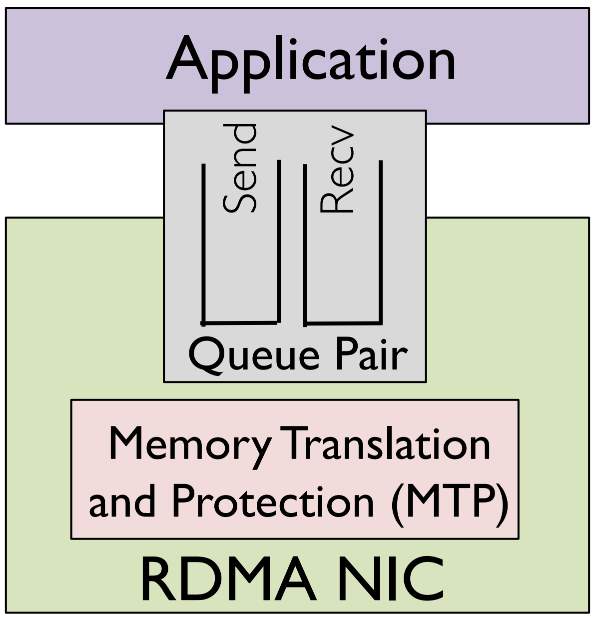
Applications bypass the kernel and interact directly with the RDMA NIC using the **IB verbs** API provided by the NIC driver.

Memory Translation and Protection



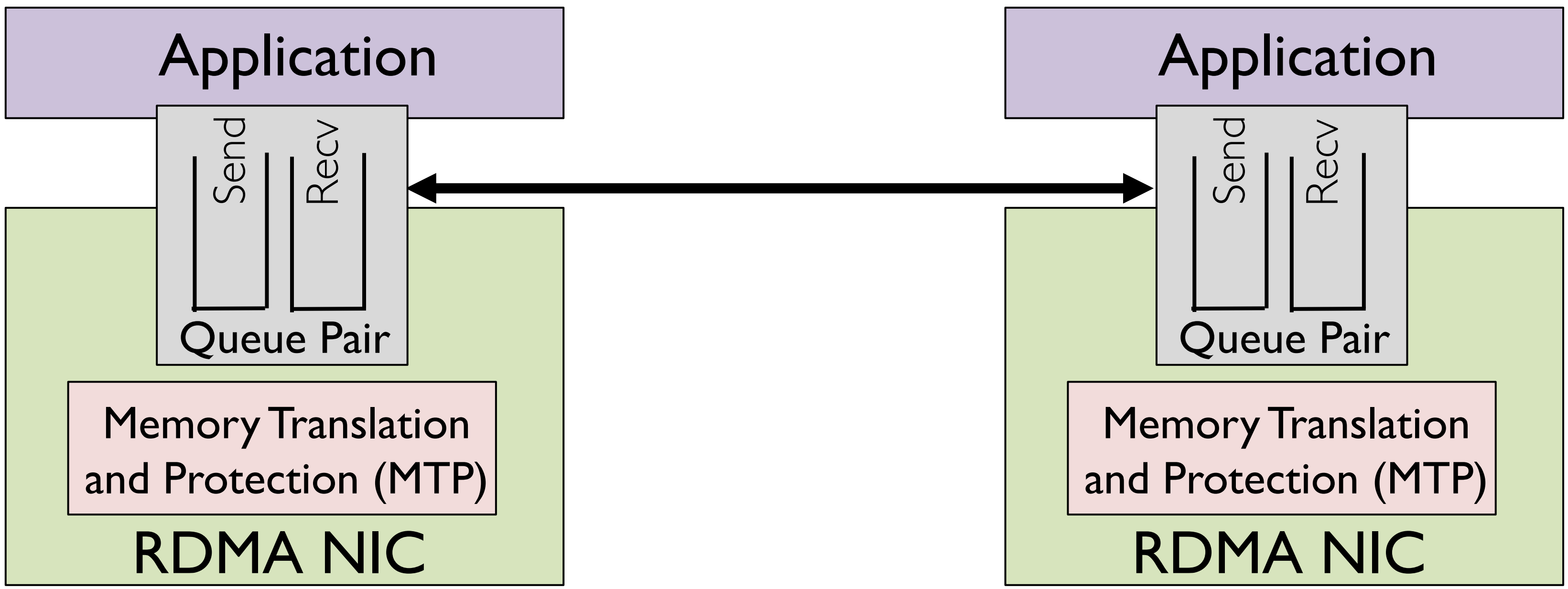
- Applications register *memory regions* with the NIC.
- **Translation:** MTP maintains *virtual address* to *physical address* mapping.
- **Protection:** MTP assigns local and remote access keys to memory region.

Queue Pairs (QP)



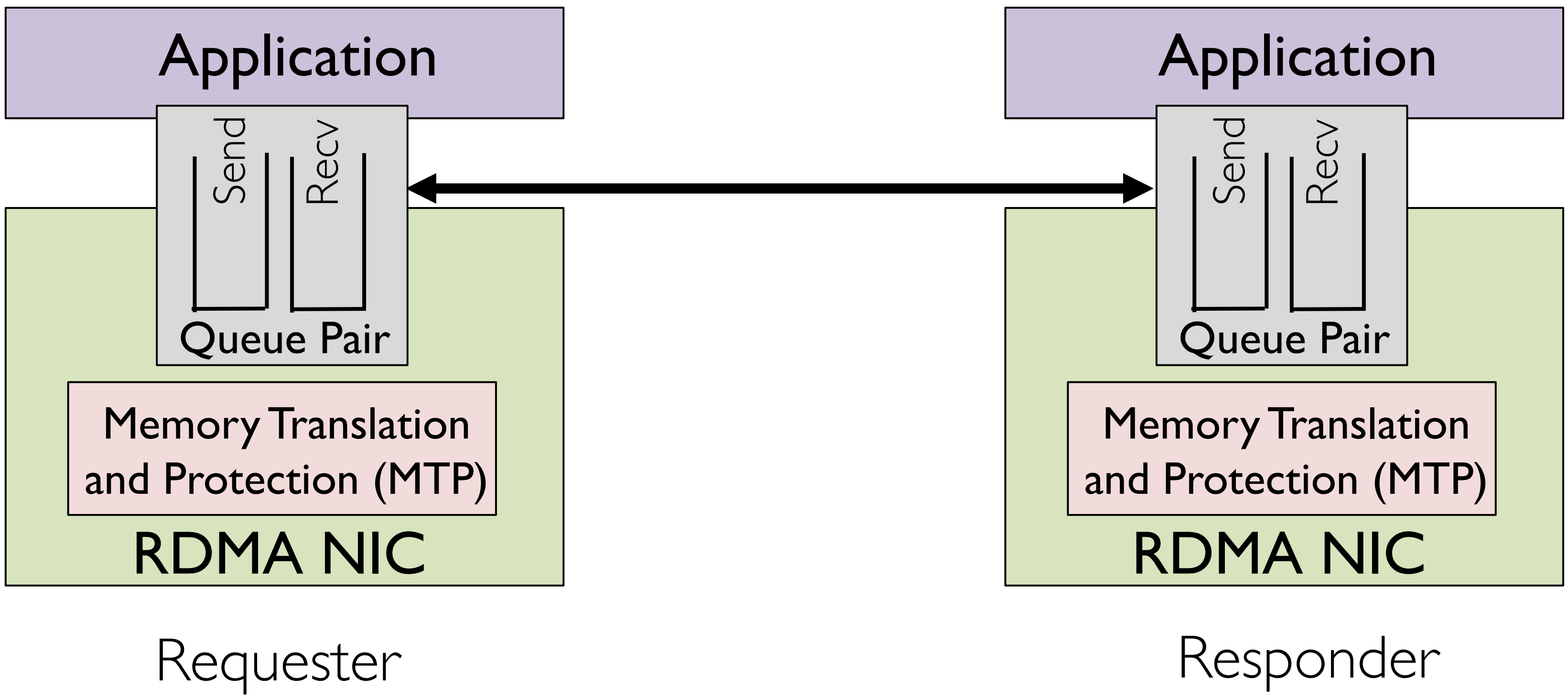
- QPs are interfaces between the application and the NIC.
- Different types:
  - Connection-oriented vs Datagram
  - Reliable vs unreliable.
- Reliable Connected (RC) QPs
  - Analogous to a TCP connection.
  - Support all types of operations.

Connection Establishment

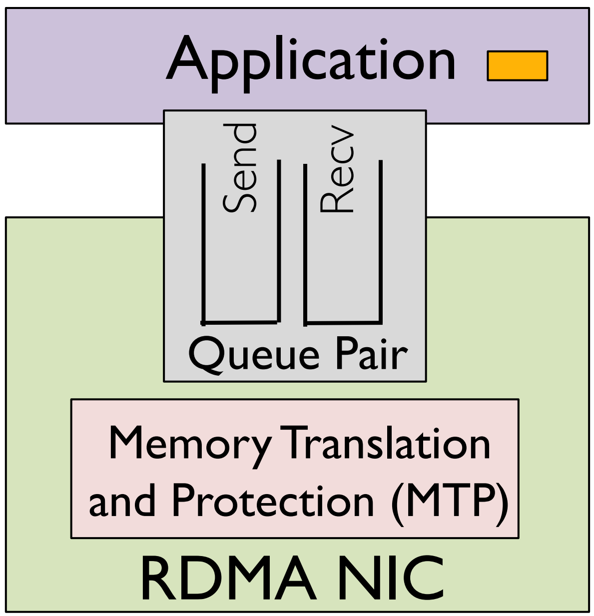


Connection establishment requires out-of-band exchange of node identifiers, QP id, and remote keys.

Work Requests

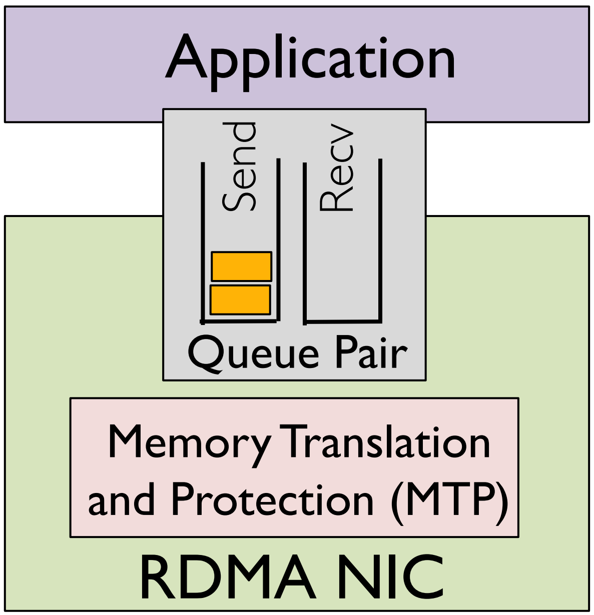


Work Requests



- Application issues a *work request* (WR) for a QP.
- WR contains all the metadata associated with a message transfer.

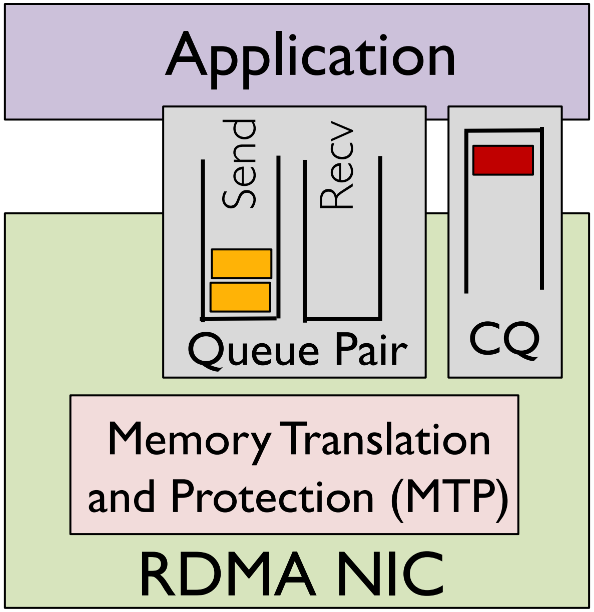
Work Queue Element (WQE)



Requester

- This WR gets stored as a *Work Queue Element (WQE)* at the QP's send queue.
- Multiple WQEs can get queued up in the send queue.
- RDMA NIC processes these WQEs one after another.

Completion Queue Element (CQE)



Requester

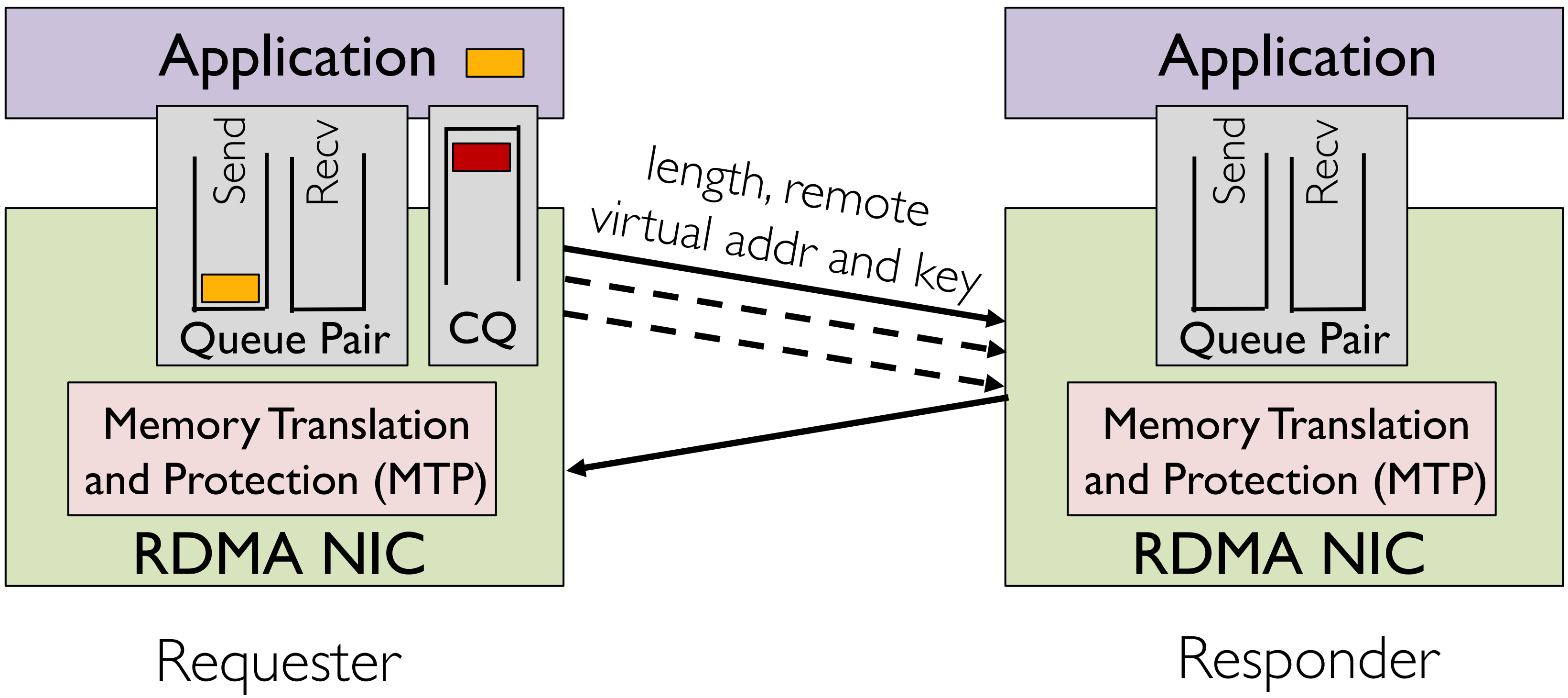
- Each QP is associated with a completion queue (CQ).
- Upon request completion,
  - The WQE expires.
  - CQE is created.
- CQE notifies request completion to application.

Four Types of RDMA Operations

- **RDMA Write:** Write data from local node to specified address at remote node.
- **RDMA Read:** Read data from specified address at remote node to local node.
- **RDMA Atomic:** Atomic fetch-add and compare-swap operations at specified location at remote node.
- **Send/Receive:** Send data to a remote node.

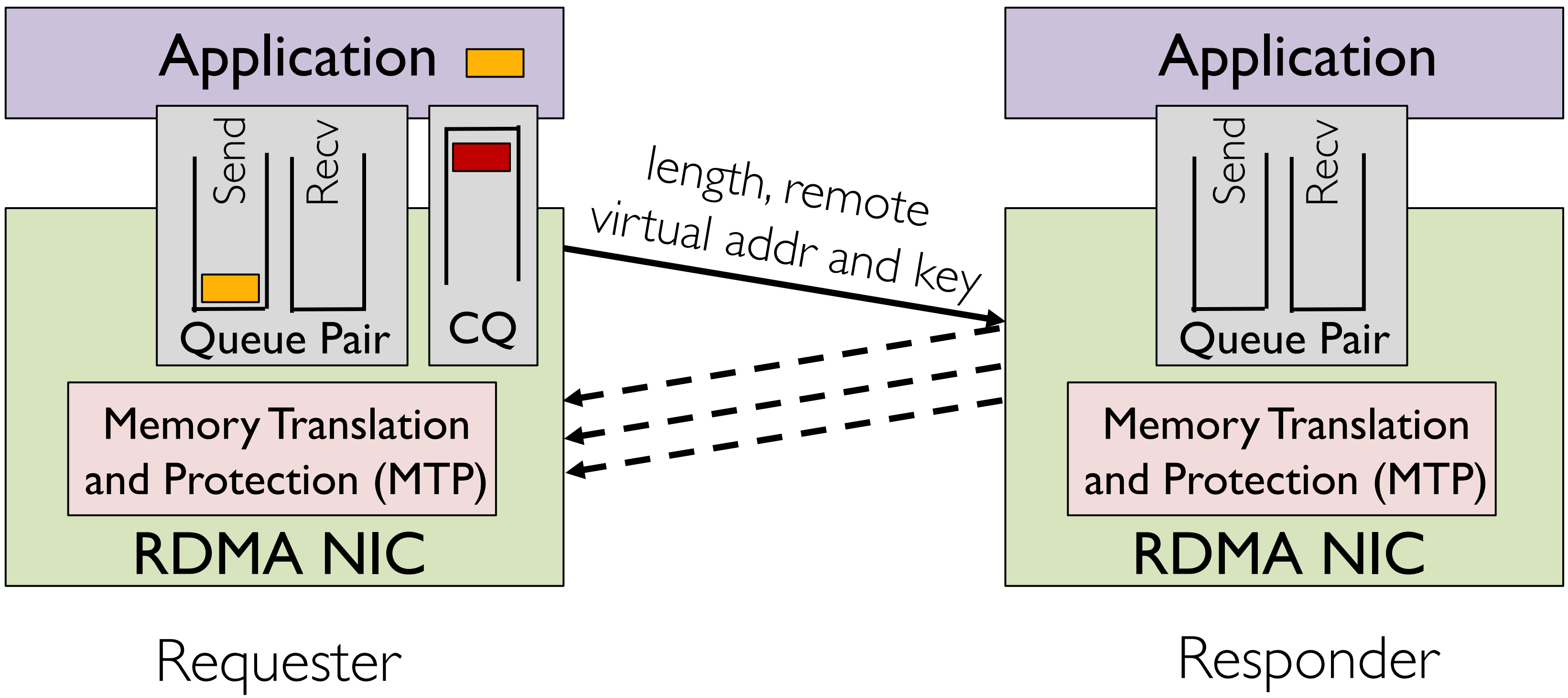
RDMA Write

WR/WQE metadata: local source virtual addr, local key, data length, remote sink virtual addr, remote key



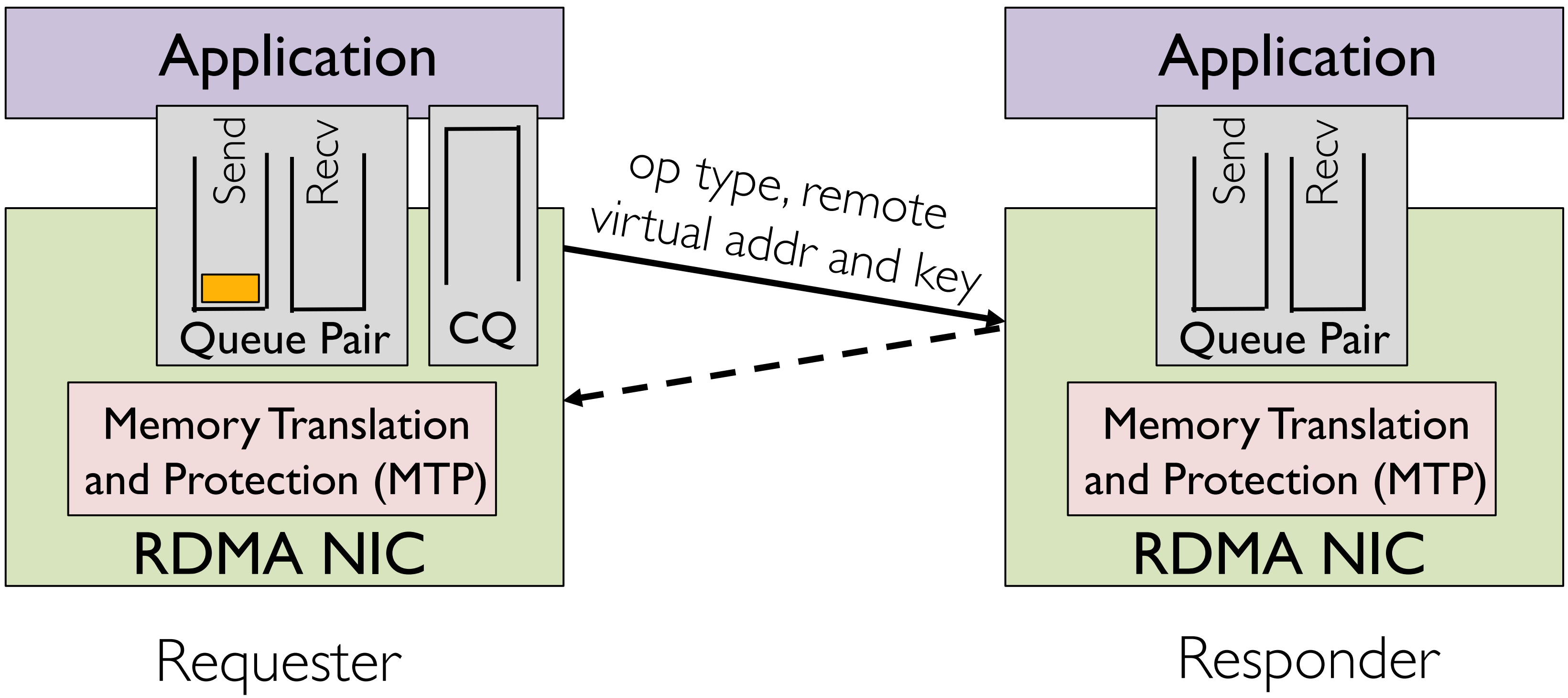
RDMA Read

WR/WQE metadata: *local sink* virtual addr, local key, data length, *remote source* virtual addr, remote key



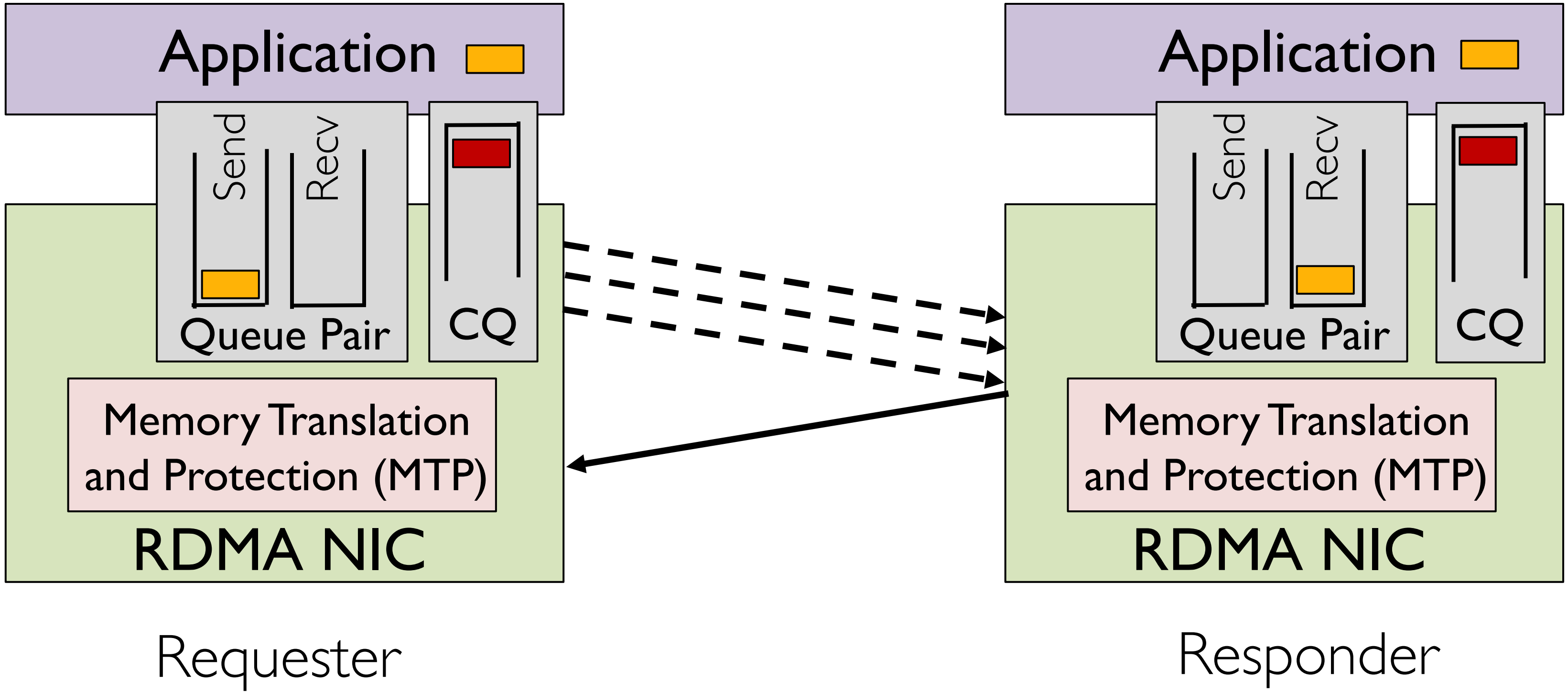
RDMA Atomic

WR/WQE metadata: local sink virtual addr, local key, *atomic operation*, remote source virtual addr, remote key



RDMA Send and Receive

Send WQE metadata: local source virtual addr, local key, data length.
Receive WQE metadata: local sink virtual addr



Four Types of RDMA Operations

- **RDMA Write:** Write data from local node to specified address at remote node.
- **RDMA Read:** Read data from specified address at remote node to local node.
- **RDMA Atomic:** Atomic fetch-add and compare-swap operations at specified location at remote node.
- **Send/Receive:** Send data to a remote node.

Lower layers for RDMA

- Traditionally designed for Infiniband.
  - Own set of networks protocols and addressing.
- RDMA over Converged Ethernet (RoCE)
  - Allows running RDMA over Ethernet.
- RoCEv2
  - Allows running RDMA over IP.

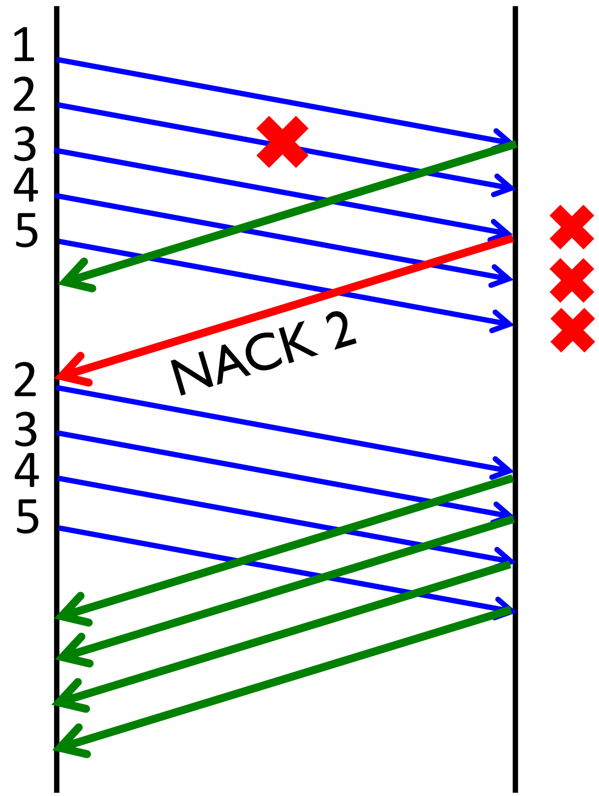
Focus of today's lecture

- Overview of RDMA
- RDMA deployment in today's datacenters

Conventional RDMA

- RDMA traditionally used in Infiniband clusters.
  - A different network protocol supporting high bandwidth.
- Infiniband links use credit-based flow control.
  - Losses are rare.
- Transport layer in RDMA NICs not designed to deal with losses efficiently.
  - Receiver discards out-of-order packets.
  - Sender does *go-back-N* on detecting packet loss.

Go-back-N Loss Recovery



Receiver discards all out-of-order packets.

Sender retransmits all packets sent after the last acked packet.

Conventional RDMA

- RDMA traditionally used in Infiniband clusters.
  - A different network protocol supporting high bandwidth.
- Infiniband links use credit-based flow control.
  - Losses are rare.
- Transport layer in RDMA NICs not designed to deal with losses efficiently.
  - Receiver discards out-of-order packets.
  - Sender does *go-back-N* on detecting packet loss.

RDMA in datacenters

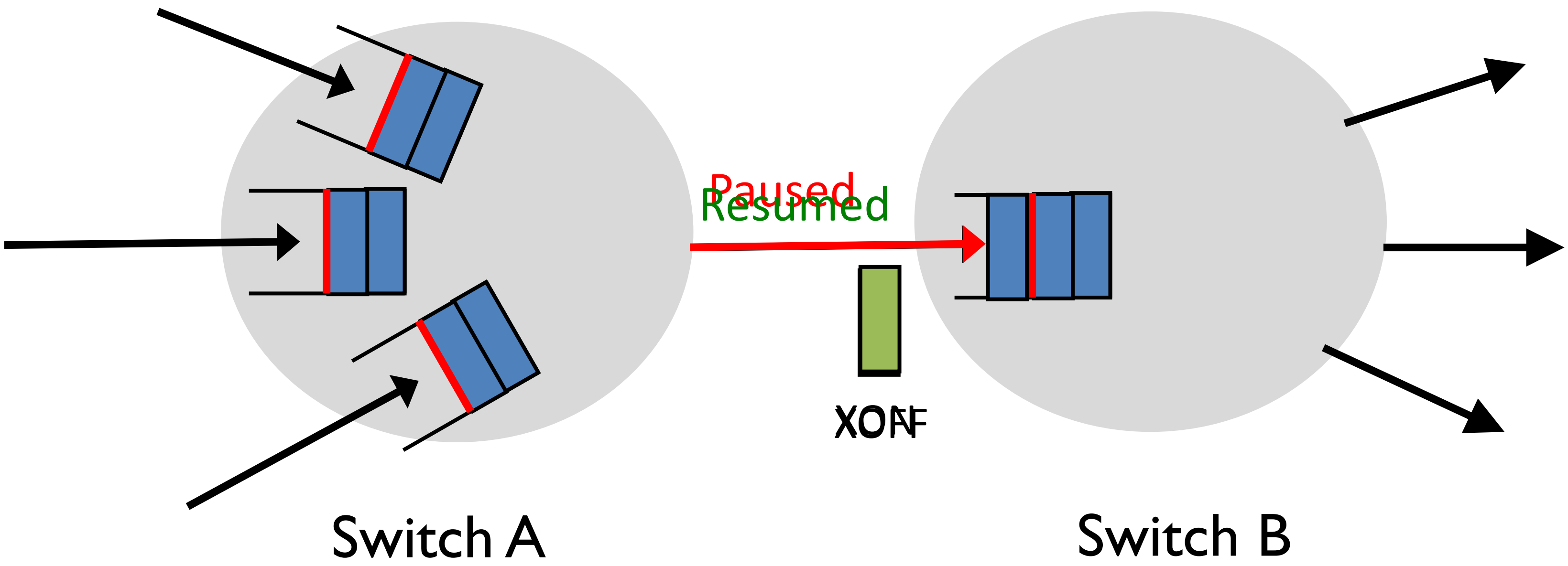
- Desire to run RDMA over commodity Ethernet.

- RoCE: RDMA over Ethernet fabric.
  - RoCEv2: RDMA over IP-routed networks.

- Infiniband transport was adopted as it is.
  - Go-back-N loss recovery.
  - Needs a lossless network for good performance.

Network made lossless by enabling PFC

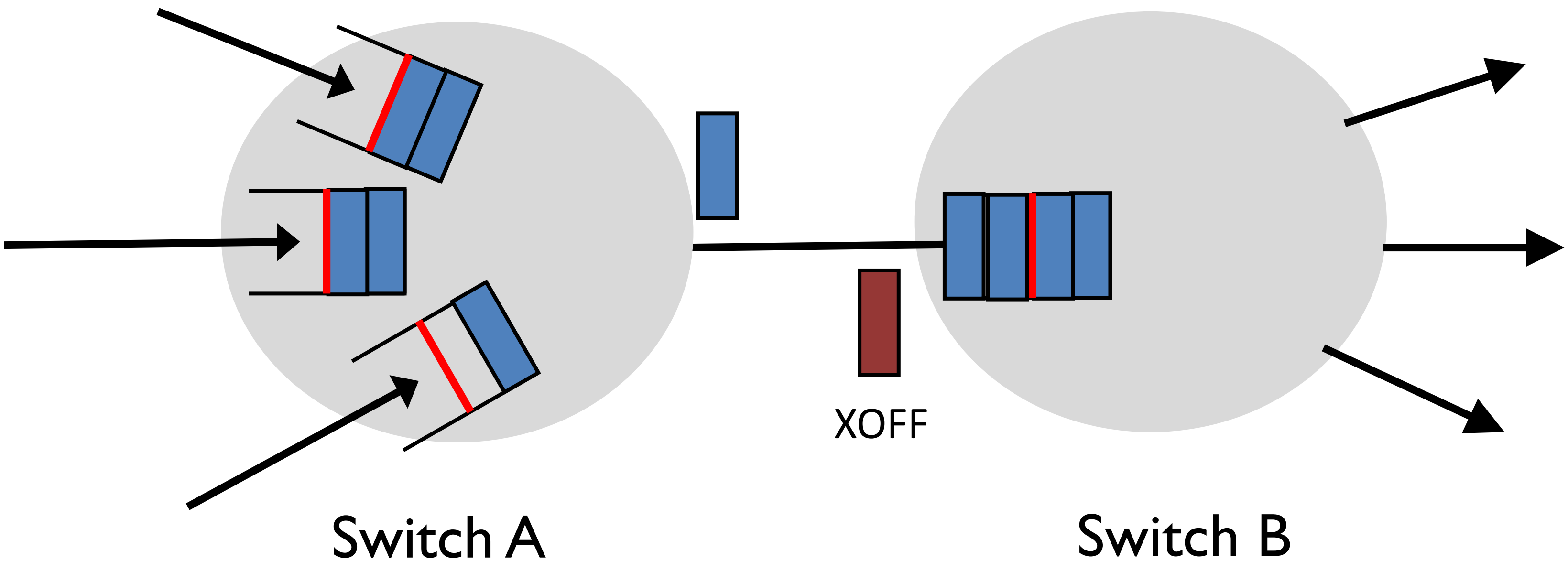
- Priority Flow Control (PFC)
  - Pause transmission when queuing exceeds a certain threshold.



Drawbacks of PFC

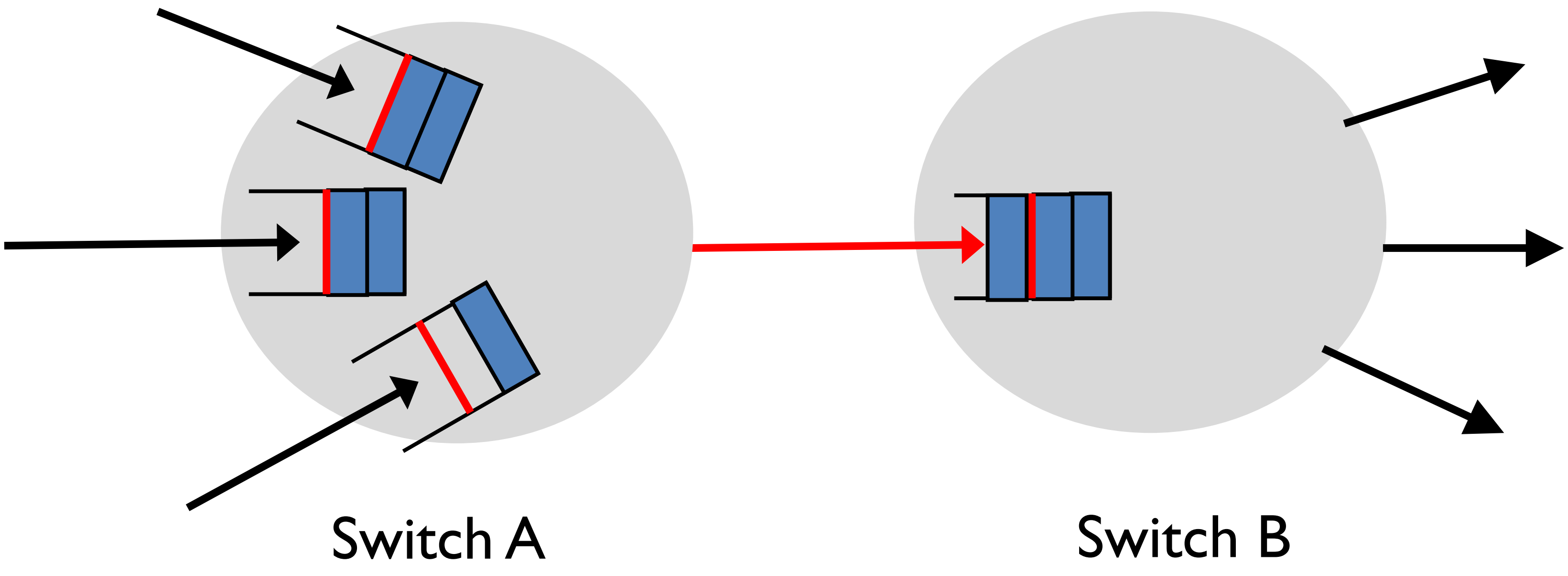
Complicates network management.

PFC threshold requires careful configuration.



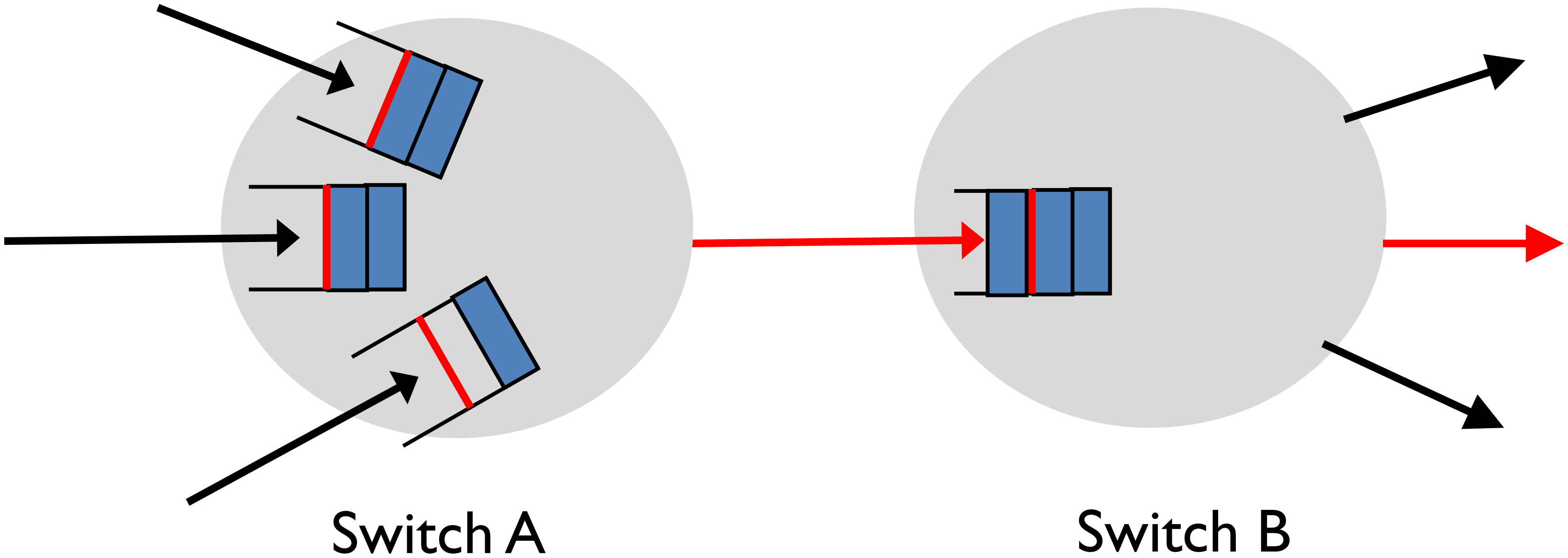
Drawbacks of PFC

Unfairness and Head-of-Line blocking



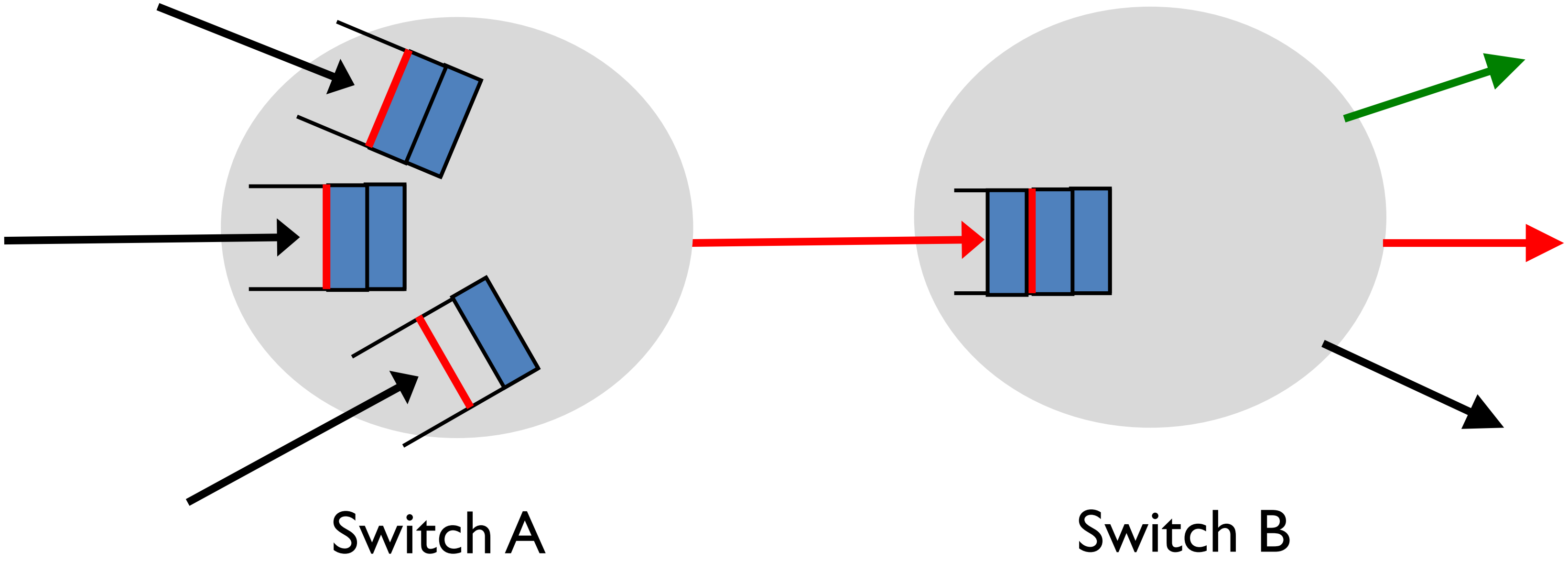
Drawbacks of PFC

Unfairness and Head-of-Line blocking



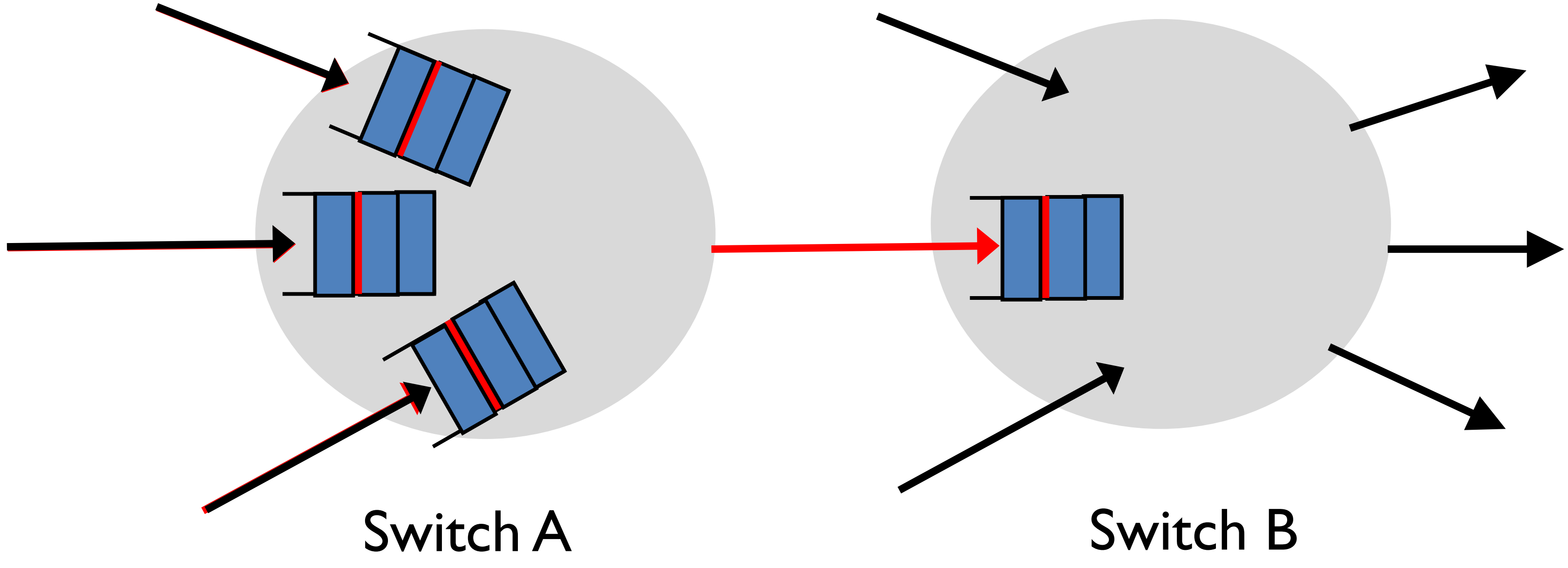
Drawbacks of PFC

Unfairness and Head-of-Line blocking



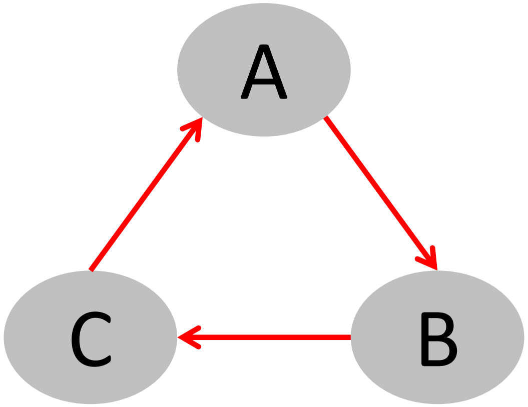
Drawbacks of PFC

Congestion Spreading



Drawbacks of PFC

Deadlocks caused by cyclic buffer dependency

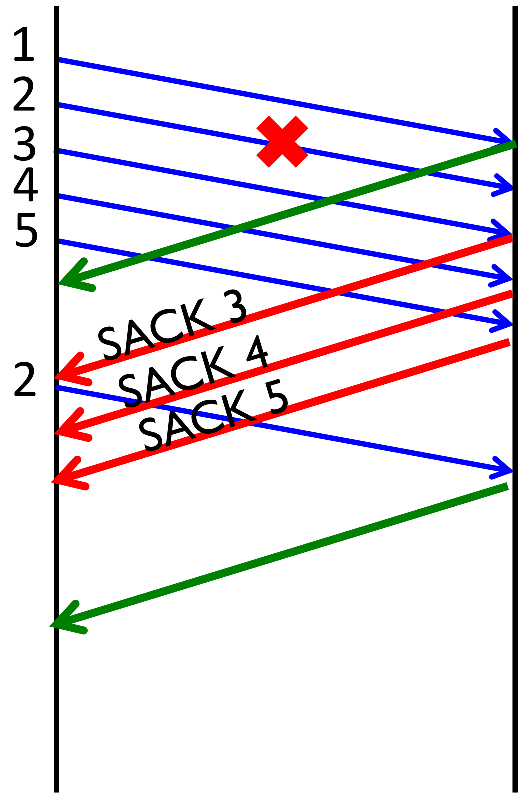


Lots of work highlighting PFC issues

- **Congestion control to mitigate PFC issues**
  - Delay-based TIMELY, *Mittal et al, SIGCOMM'15*
  - ECN-based DCQCN, *Zhu et al, SIGCOMM'15*
  - Recent update: Zero-touch RoCE based on delay-based congestion control (NVIDIA/Microsoft)
- **Deployment experience**
  - RDMA over commodity Ethernet at scale, *Guo et al, SIGCOMM'16*
- **Deadlock avoidance**
  - Deadlocks in datacenter: why do they form and how to avoid them, *Hu et al, HotNets 2016*
  - Unlocking credit loop deadlock, *Shpiner et al, HotNets 2016*
  - Tagger: Practical PFC deadlock prevention in datacenter networks, *Hu et al, CoNext 2017*

A potential fix

Update the RoCE NIC design to better handle losses *(IRN, SIGCOMM'18)*

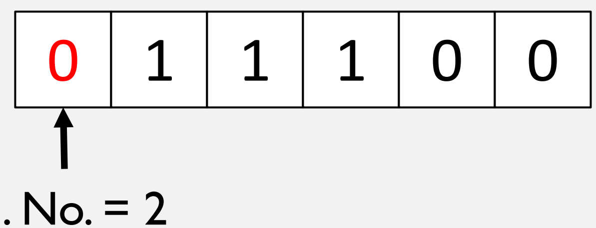


Might not scale to cross-datacenter RDMA transfers.

Receiver does not discard out-of-order packets and *selectively acknowledges* them.

Sender retransmits only the lost packets.

Use bitmaps to track lost packets.



Limit no. of in-flight packets to bandwidth-delay product.

System-level Challenges

- Limited NIC cache (FaRM, NSDI'14)
  - Performance decreased with amount of memory registered for remote access
    - More page table entries required.
    - Couldn't fit all entries in NIC cache.
  - Solution: use large pages (2GB).
    - Implemented a kernel driver.
    - Unit of address mapping, of recovery, and of registration with memory.

System-level Challenges

- Limited NIC cache (FaRM, NSDI'14)
  - Performance decreased as cluster size increased.
    - Larger number of QPs required.
      - Ideally, $2 \times m \times t^2$
      - Reduced to $2 \times m \times t$
      - m = no. of machines, t = no. of threads/machine
    - Couldn't fit all QP context in NIC cache.
  - Solution: use fewer QPs (2mt / q)
    - Larger 'q', higher sharing overhead.

Challenges of deploying RDMA in DCs

- Need for a lossless network
  - Congestion control to mitigate PFC issues (DCQCN, Timely, ZTR).
  - Better loss recovery in the NIC (IRN, SIGCOMM'18)
  - Large enough buffers + congestion control (eRPC, NSDI'19)
- Limited NIC cache:
  - Use bigger pages for memory translation (FaRM, NSDI'14).
  - Optimizing number of QPs (FaRM, NSDI'14; FASST, OSDI'16).
- Limited resource sharing and isolation
  - Kernel re-direction (LITE, SOSP'17)
- Supporting RDMA for VMs (para-virtual RDMA)
  - Commercial solution from VMWare requiring NIC support.
- Limited flexibility (tied to increased heterogeneity)
  - FPGA-based implementation / firmware patches.

Today's reading

- Empowering Azure Storage with RDMA (NSDI'23)

- What was the primary usecase?

- What are the additional challenges that arise?

Is RDMA the right choice for datacenters?

What will a clean slate approach look like?

Thank you for your feedback!

- Many of you want harder assignments ☺
- Student presentations
- Broader variety of topics – more papers per class?
- Sometimes discussions tend to drag…
- More background before diving into details.