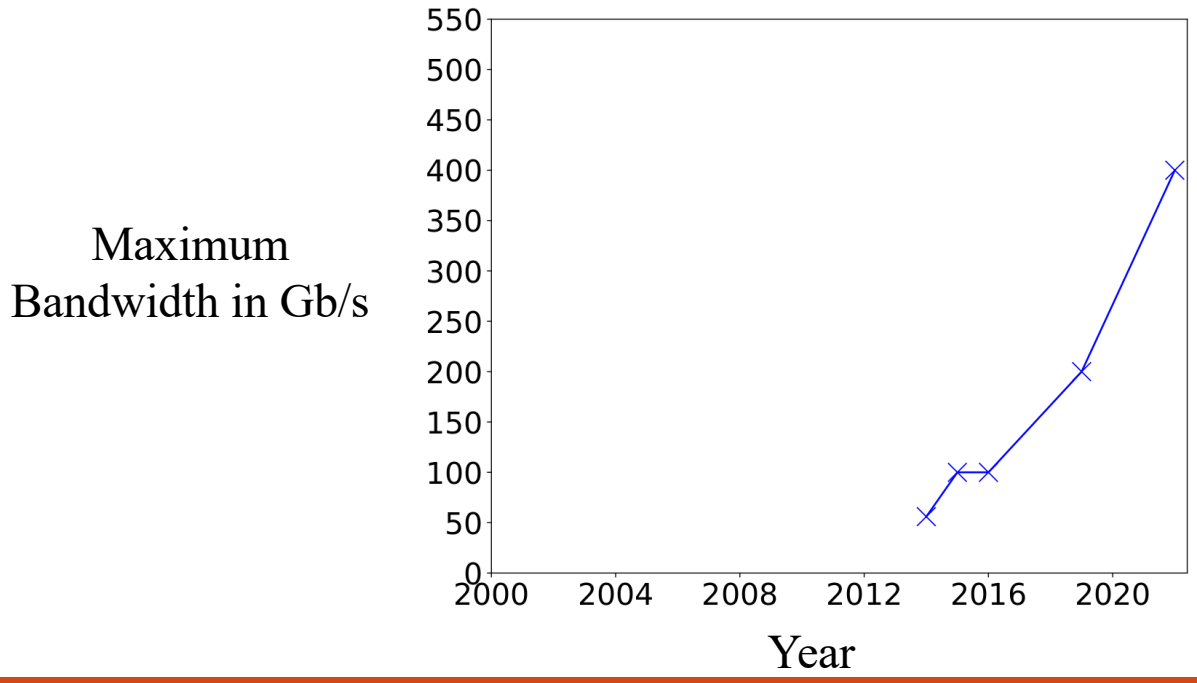
Stop Taking the Scenic Route: the Shortest Distance Between the CPU and NIC is MMIO

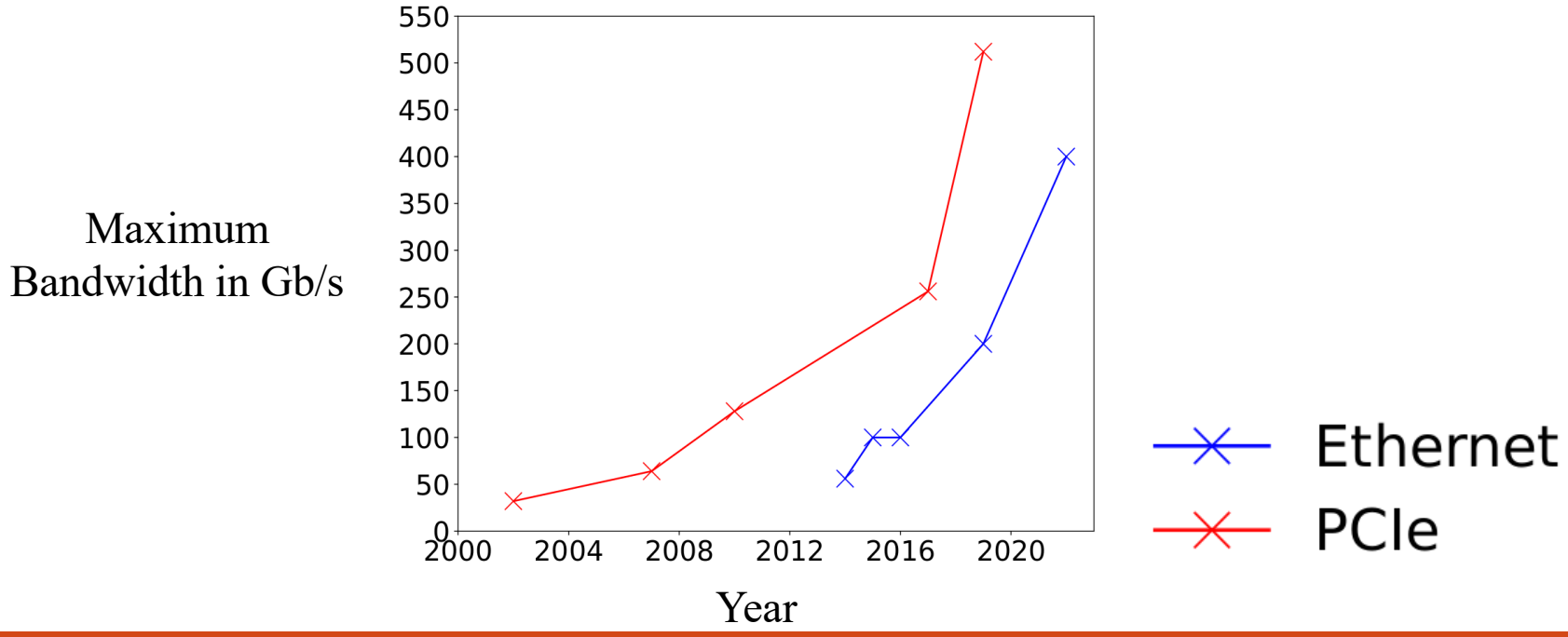
WEI SIEW LIEW

UNIVERSITY OF UTAH

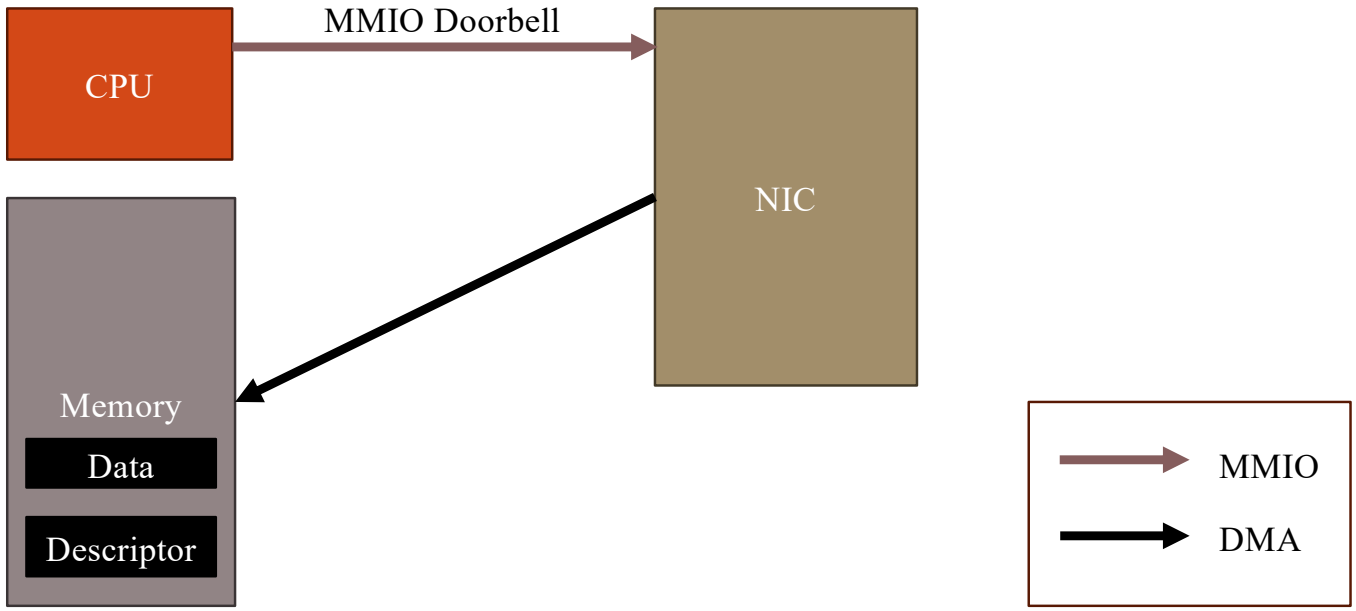
Ethernet Bandwidth is Increasing



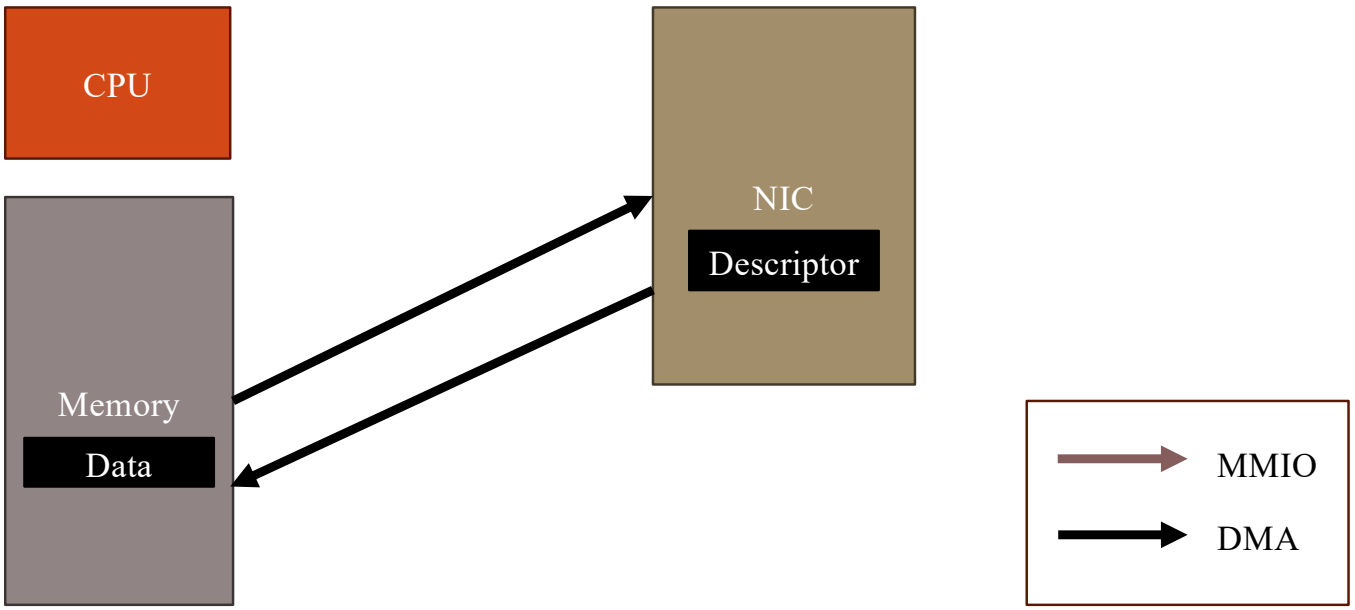
Optimize communication between CPU and NIC to benefit from increase in bandwidth



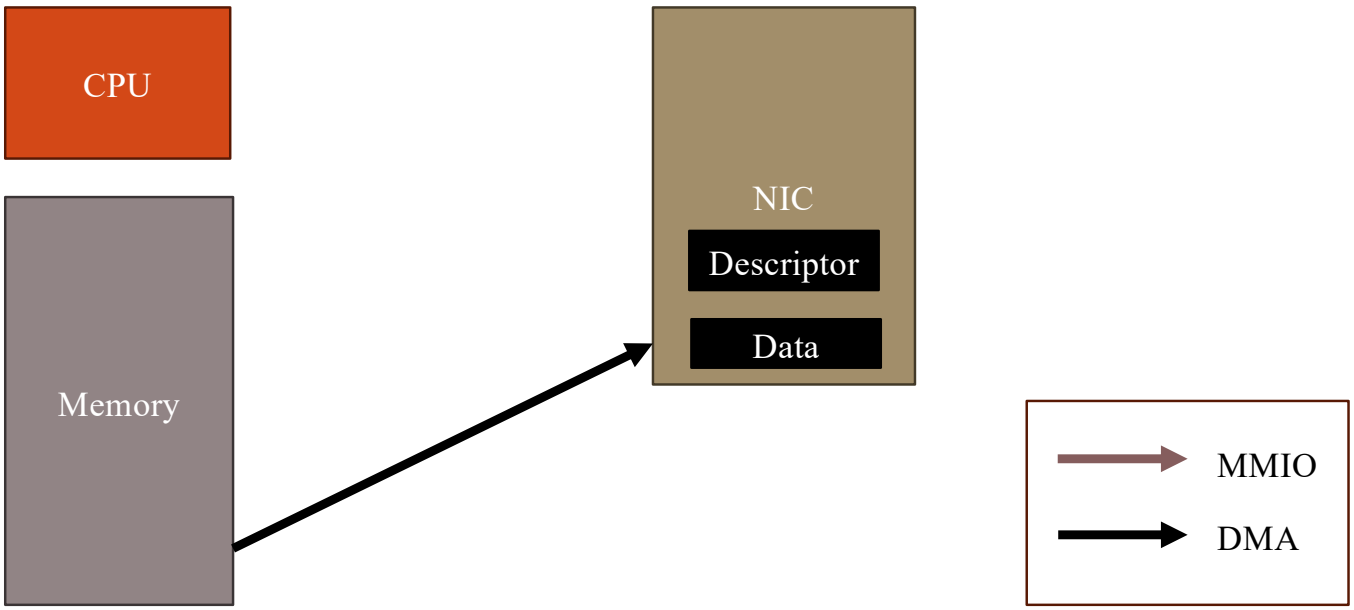
The Scenic Route: How Data is Sent from CPU to NIC Today



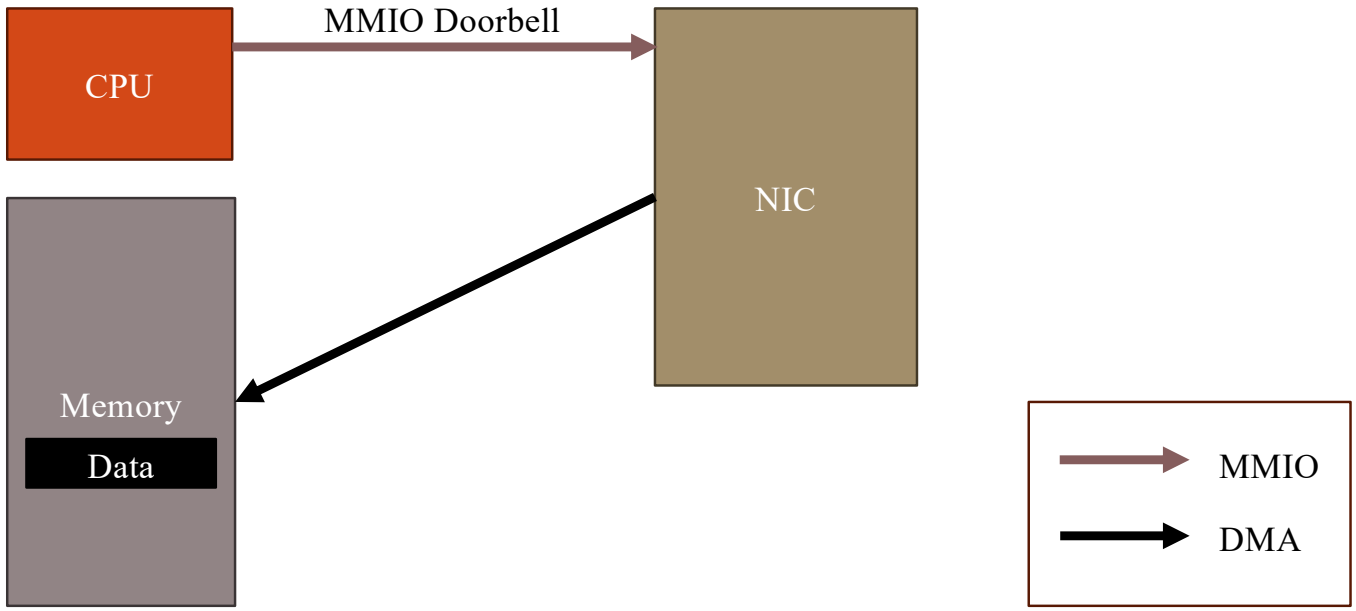
The Scenic Route: How Data is Sent from CPU to NIC Today



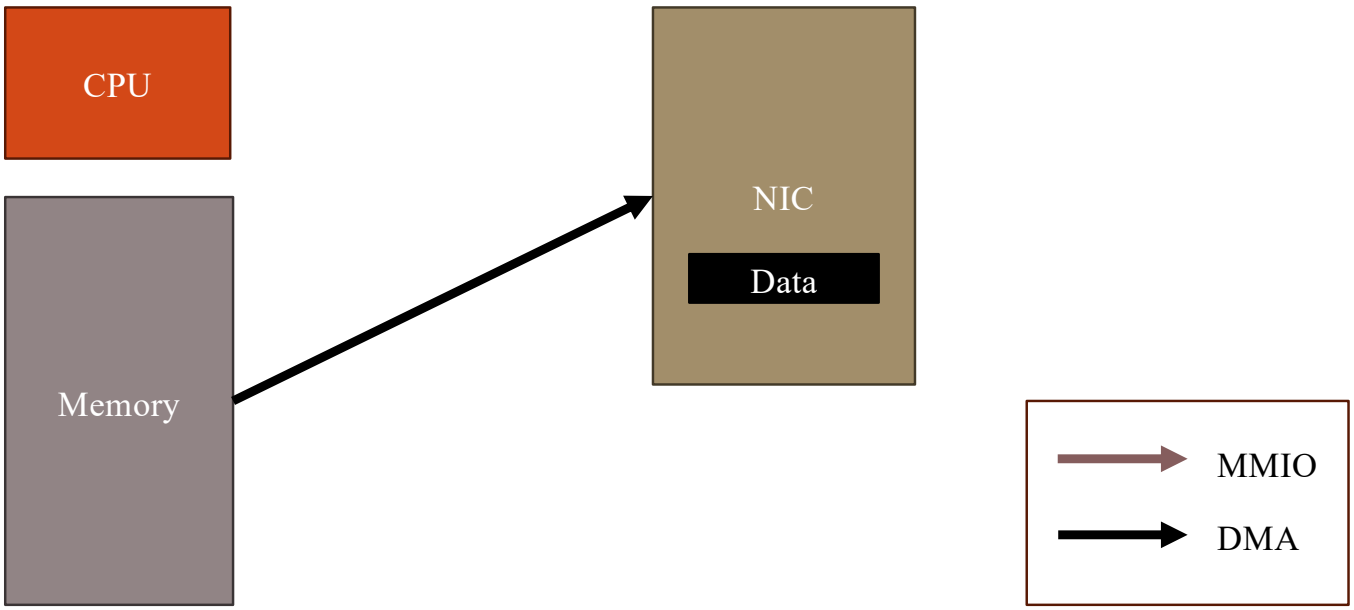
The Scenic Route: How Data is Sent from CPU to NIC Today



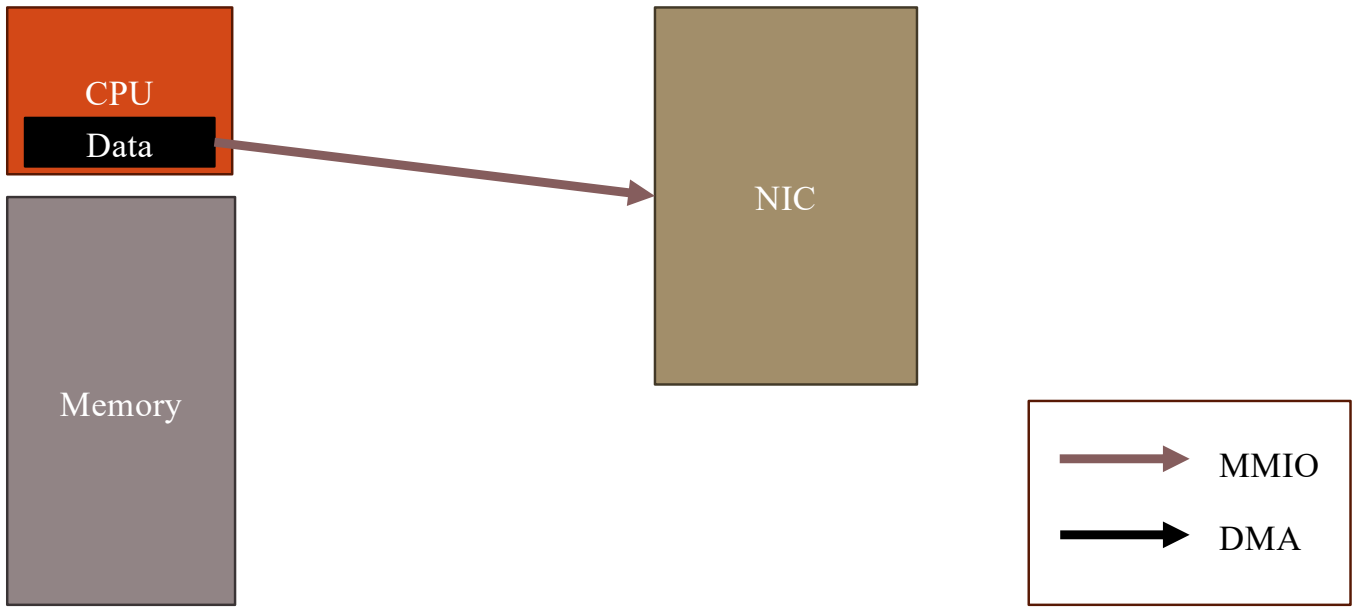
One Optimization: Ensō



One Optimization: Ensō

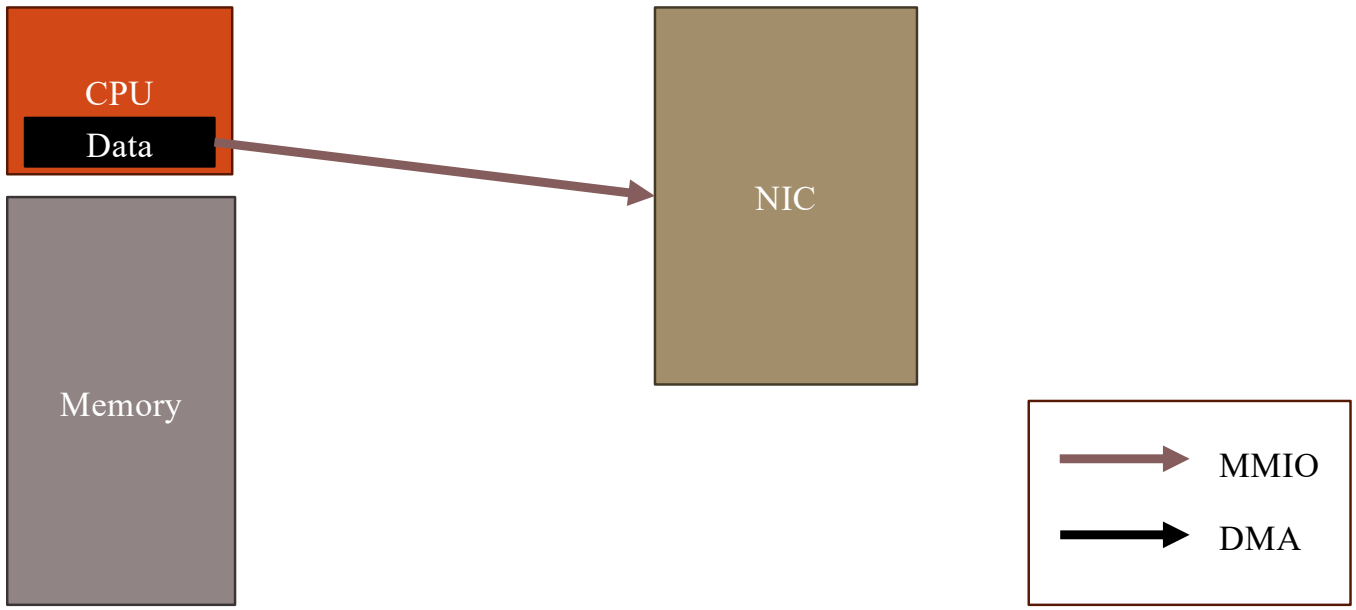


Stop Taking the Scenic Route?

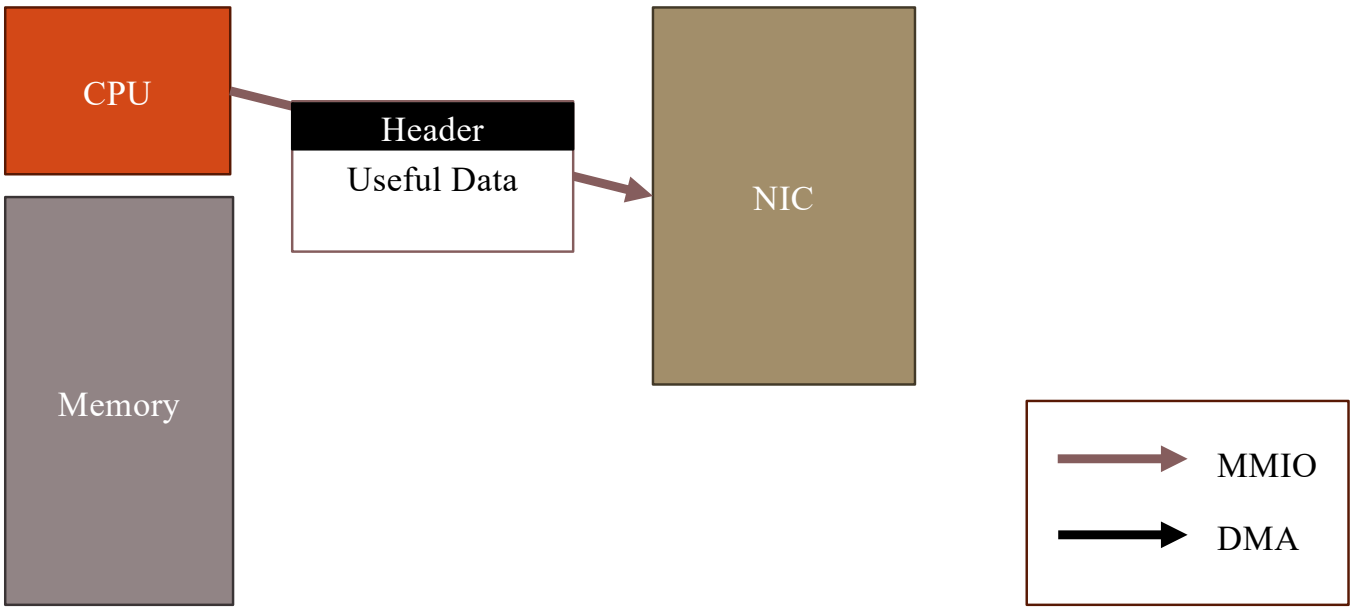


Stop Taking the Scenic Route?

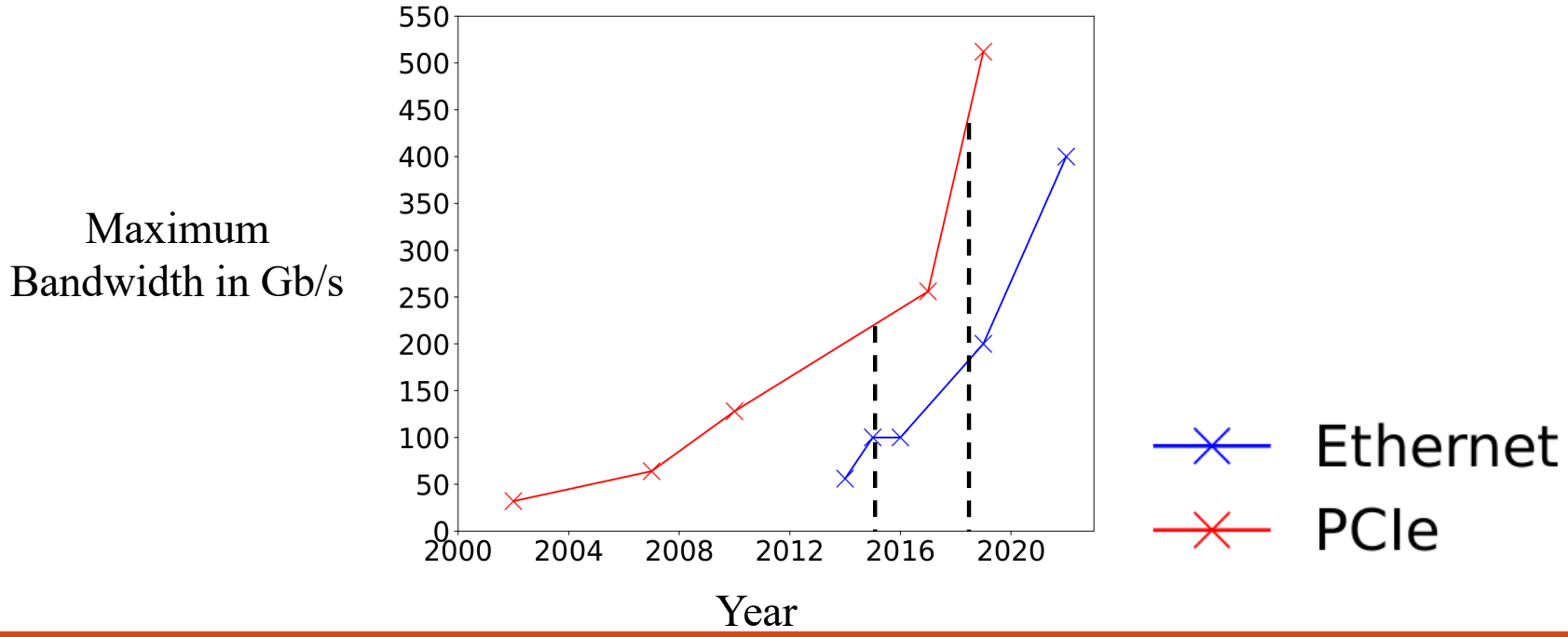
Objection 1: Is CPU Fast Enough?



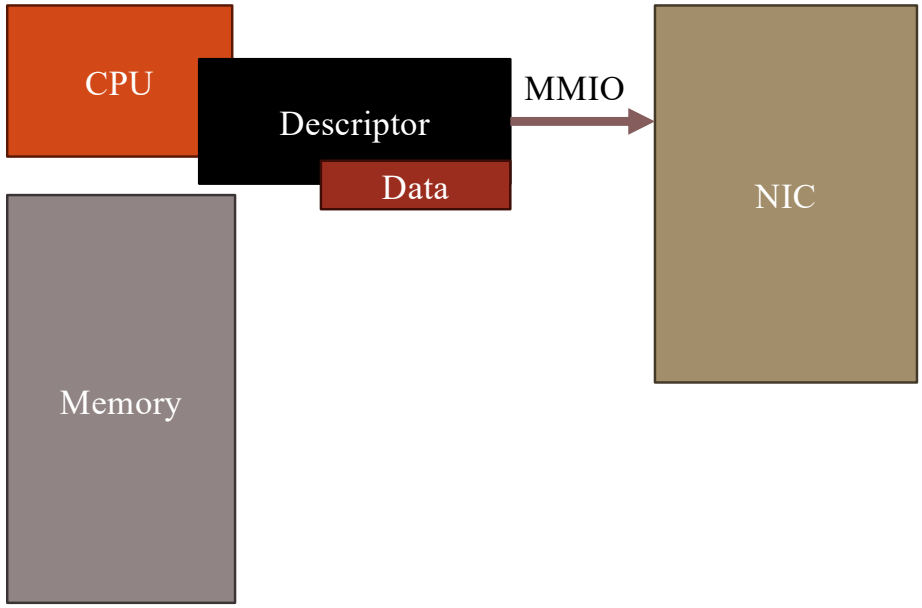
Objection 2: PCIe Bandwidth Overheads



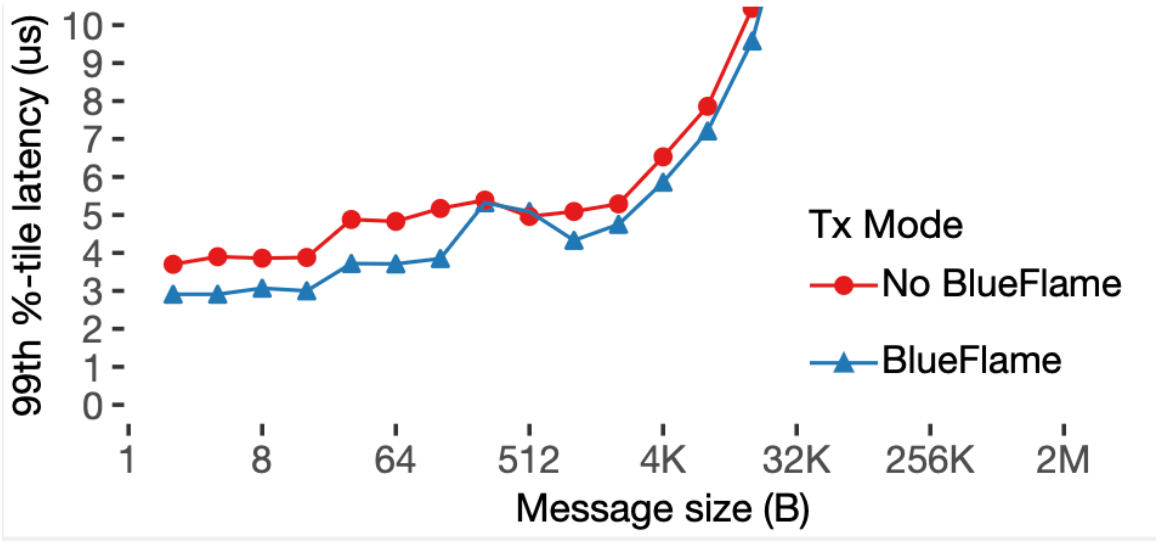
Objection 2: PCIe Bandwidth Overheads



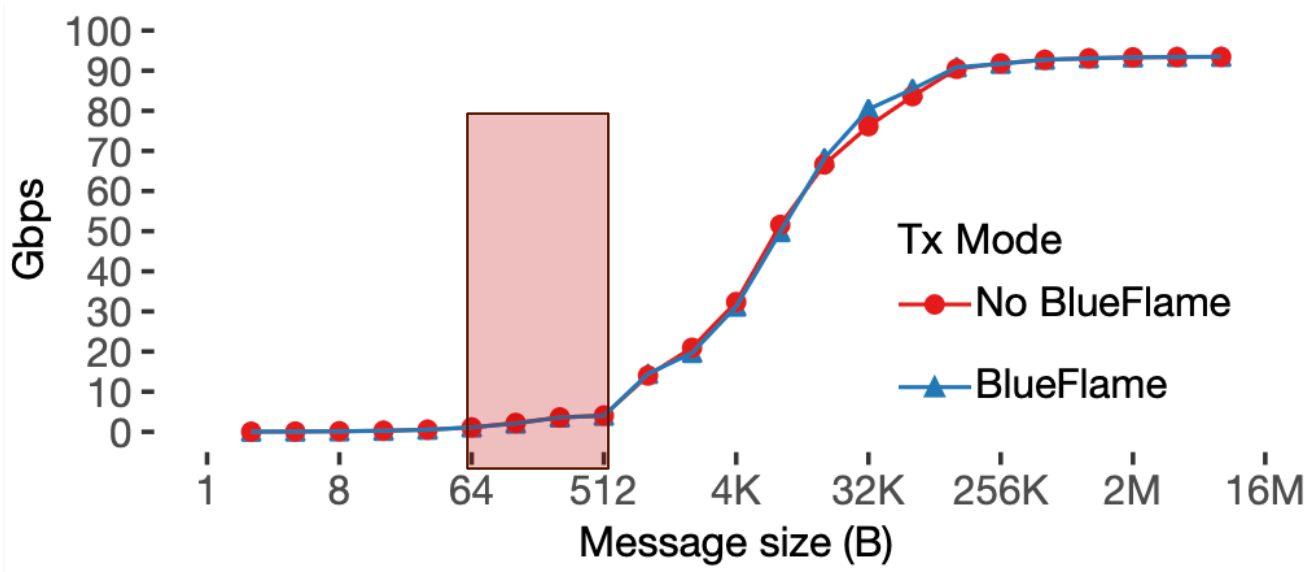
Objection 3: Don't existing NICs already do MMIO?



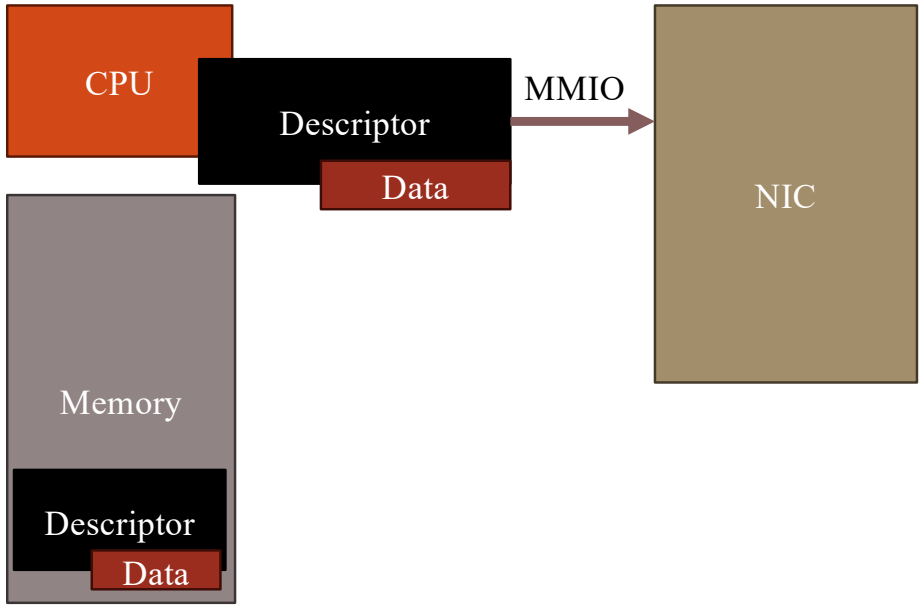
BlueFlame Optimization Improves Latency



Throughput is low with BlueFlame

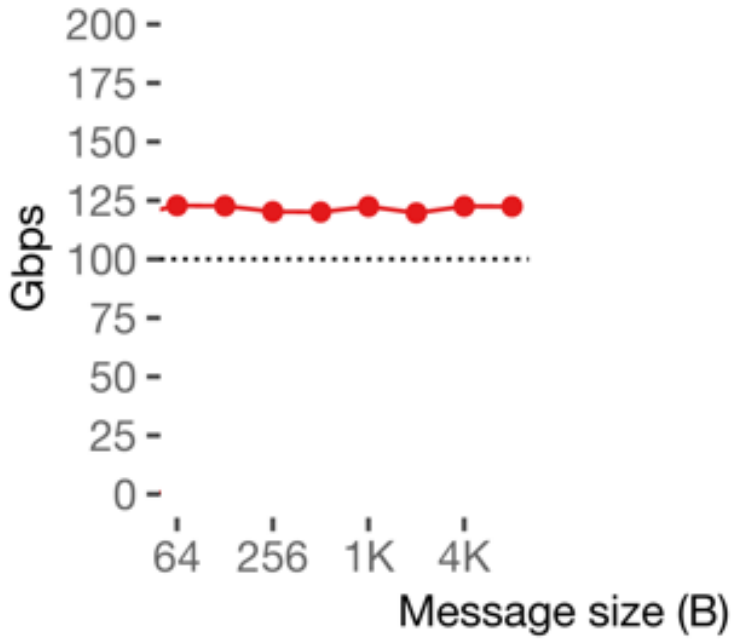


Why? MMIO is an Afterthought!

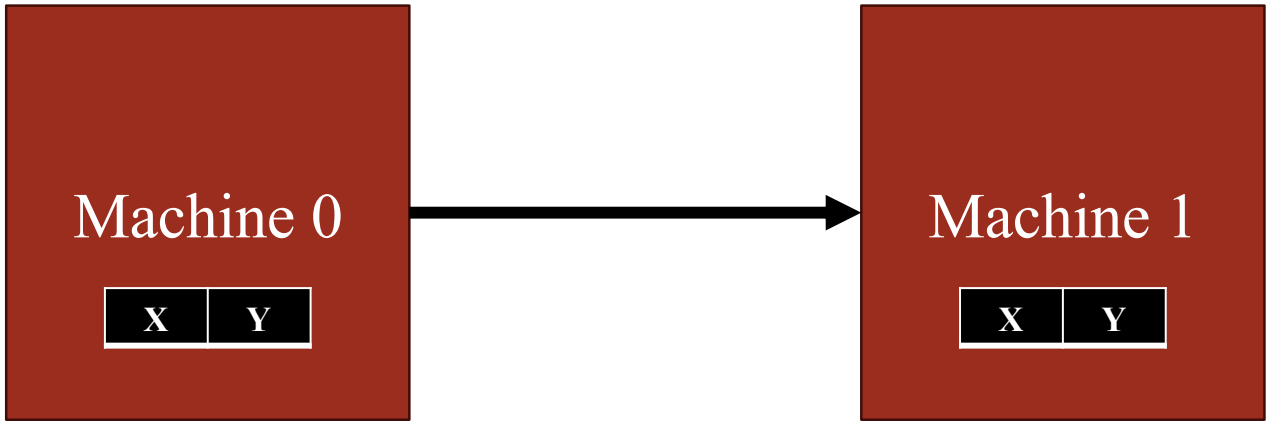


How fast can we go with MMIO?

Single Core MMIO Throughput Exceeds 100 Gbps

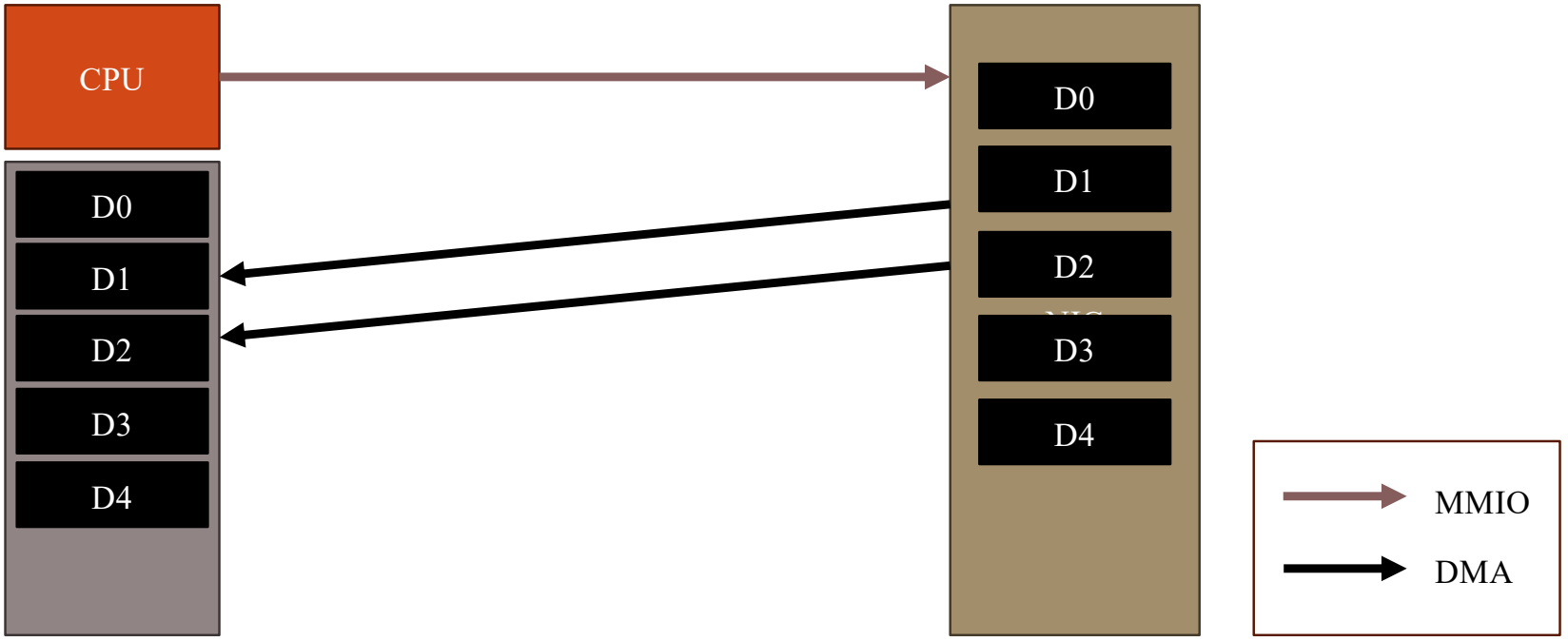


What about Ordering?

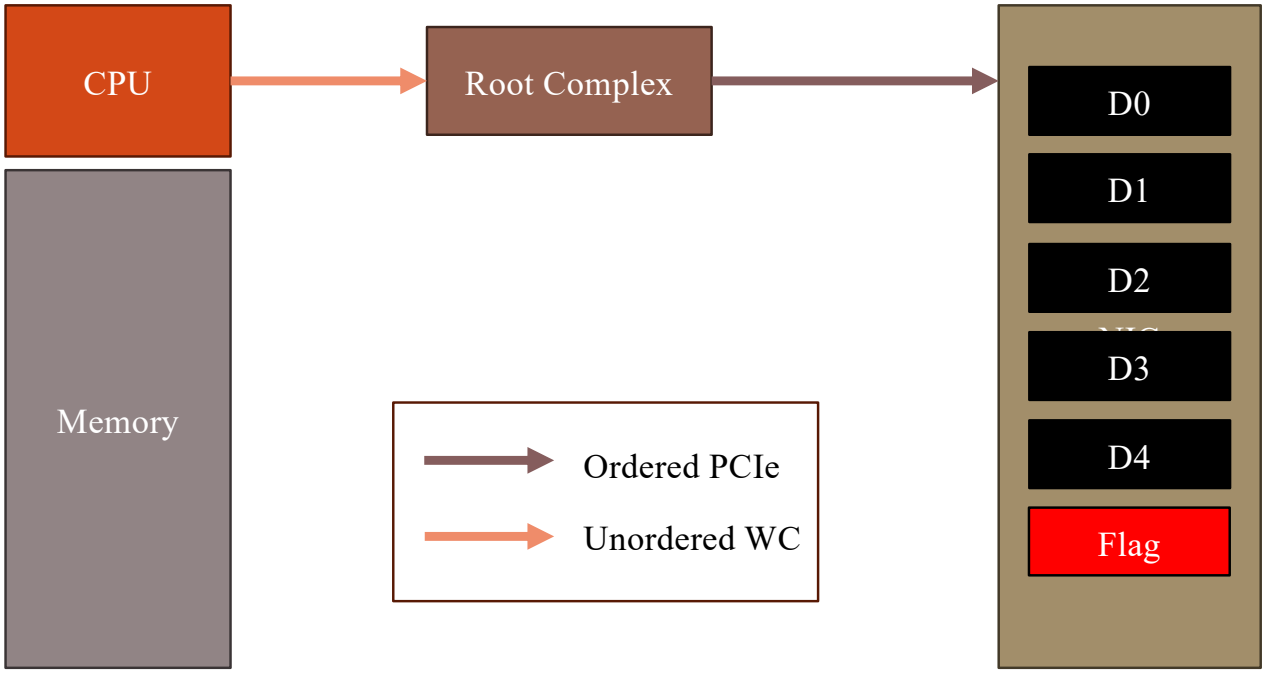


What ordering is required for transmission?

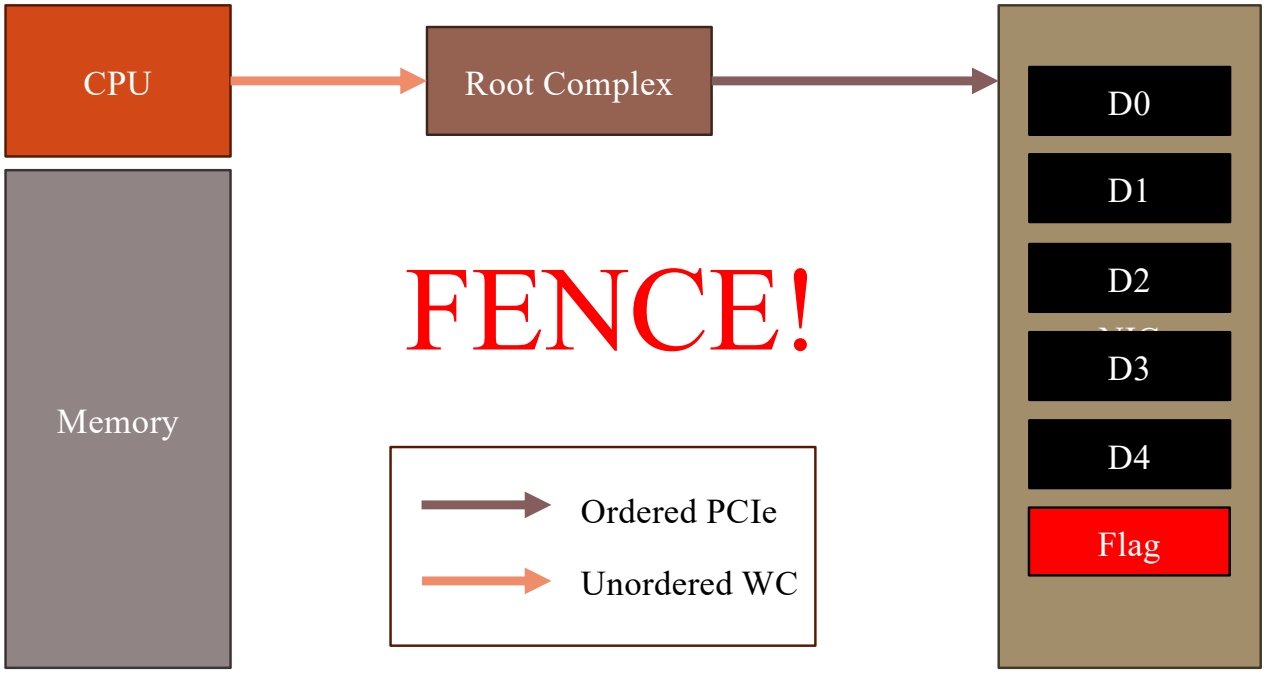
How DMA provides Ordering



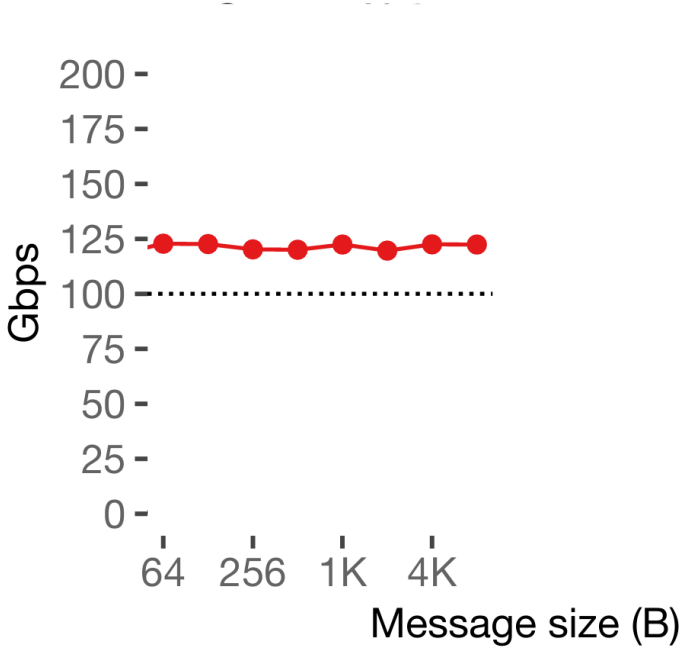
The Problem with MMIO Write Ordering



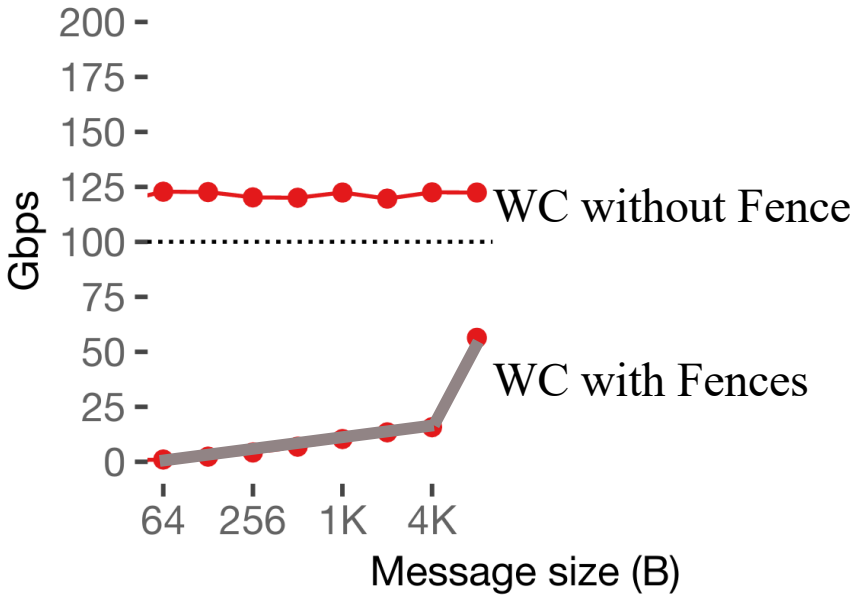
Write Ordering: The Fast Path is not Toll-Free



Is MMIO Really the Fast Path?



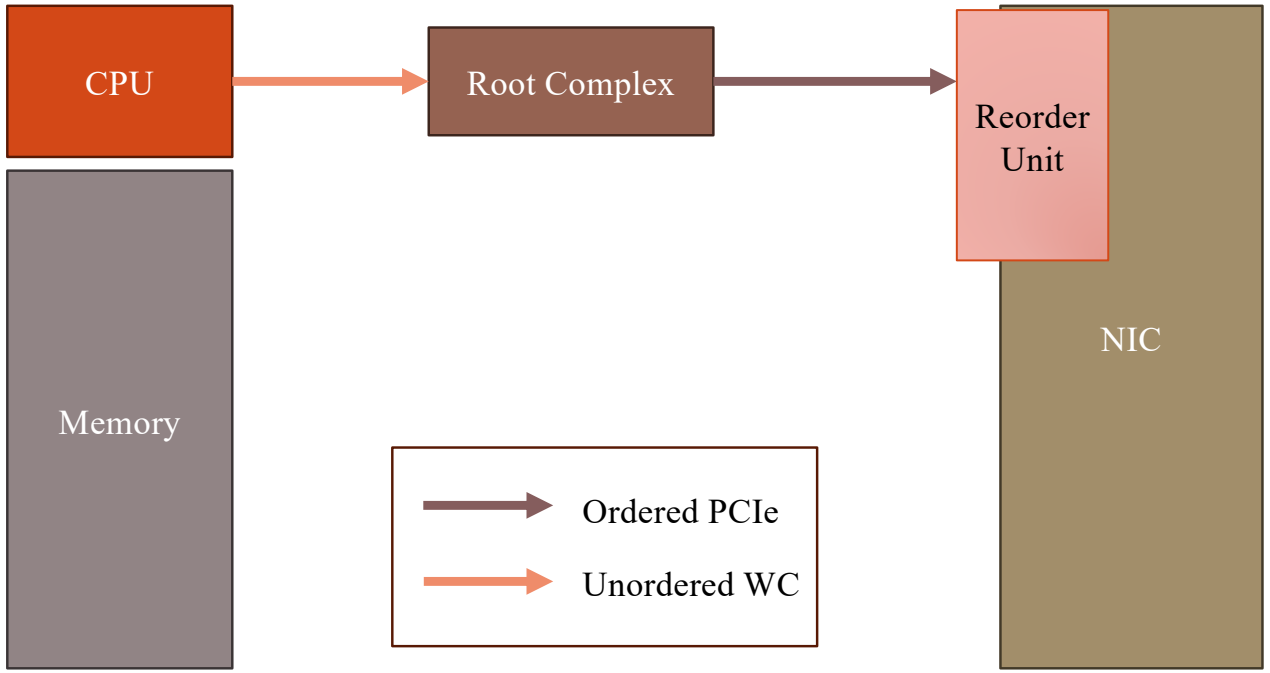
While high write throughput is possible with MMIO, it comes at the cost of ordering!



Can we get fast, ordered MMIO?

End-to-end ordering: Allow writes to be re-ordered as long as hardware at the NIC can recover order

Efficient end-to-end Ordering with the Hardware Reorder Unit



Efficient end-to-end Ordering with the Hardware Reorder Unit

Idea: Instead of ordering writes with fences, keep track of whether the complete message is received at the NIC.

Byte Address	0	1	2	3	4	5
Valid bit	0	0	0	0	0	0
Data						

Efficient end-to-end Ordering with the Hardware Reorder Unit

Idea: Instead of ordering all writes, keep track of whether the complete message is received at the NIC.

Byte Address	0	1	2	3	4	5
Valid bit	0	1	0	1	0	0
Data		B1		B3		

Efficient end-to-end Ordering with the Hardware Reorder Unit

Idea: Instead of ordering all writes, keep track of whether the complete message is received at the NIC.

Byte Address	0	1	2	3	4	5
Valid bit	1	1	0	1	0	0
Data	B0	B1		B3		

Message Size

Efficient end-to-end Ordering with the Hardware Reorder Unit

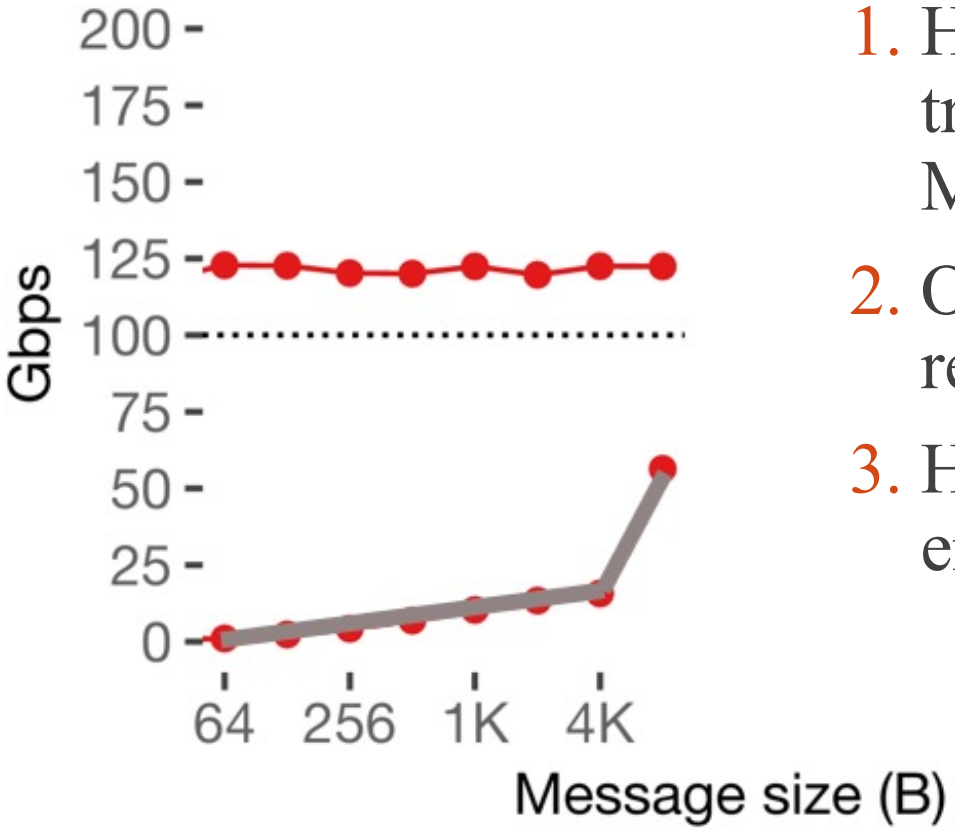
Idea: Instead of ordering all writes, keep track of whether the complete message is received at the NIC.

Byte Address	0	1	2	3	4	5
Valid bit	1	1	1	1	0	0
Data	B0	B1	B2	B3		

Message Size

Discussion

1. How much does this matter for real applications?
2. Do we still need DMA on the transmit path?
3. What about the receive path?
4. What about coherent interconnects?



1. High throughput transmission with WC MMIO.
2. Ordering using fences reduces throughput
3. Hardware support for end-to-end ordering.