

# To Copy or Not to Copy: Making In-Memory Databases Fast on Modern NICs



THE  
UNIVERSITY  
OF UTAH<sup>®</sup>

Aniraj Kesavan  
Robert Ricci  
Ryan Stutsman

# Introduction

- Didn't we solve DB I/O when we got rid of disks ?
- Today: Copy records to transmit buffer to send
- Zero copy: Transmit Data directly from records
- Do you copy or not?
- Is there a way to get best of both worlds?

# The Setup

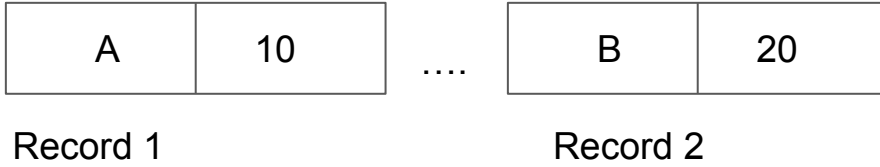
- Mellanox Infiniband Cx-3 and Connect-ib
- Peak B/W - 5.8 GB/s, latency  $\sim 1 \mu\text{s}$ , kernel bypass
- 15 clients:
  - Copy Out
  - Zero Copy

# Copy Out

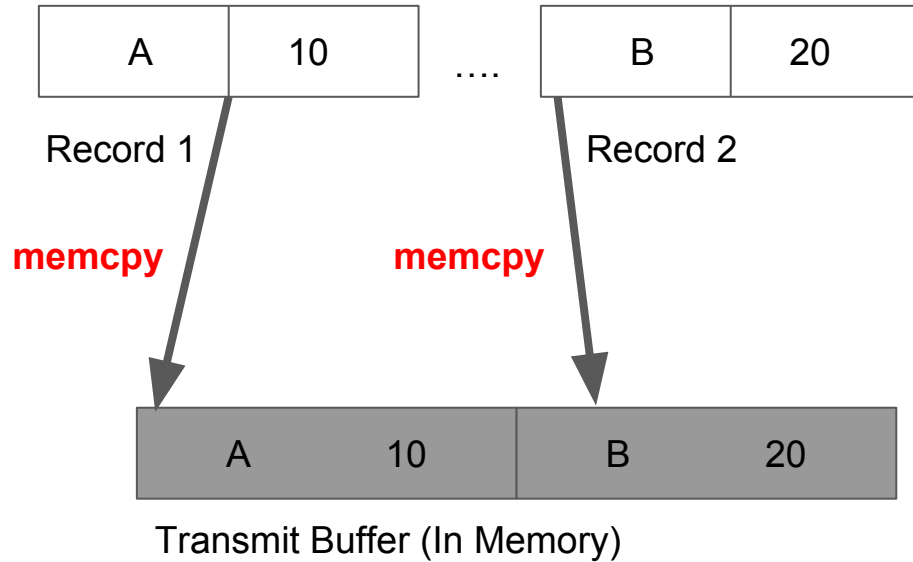
A	10
---	----

Record 1

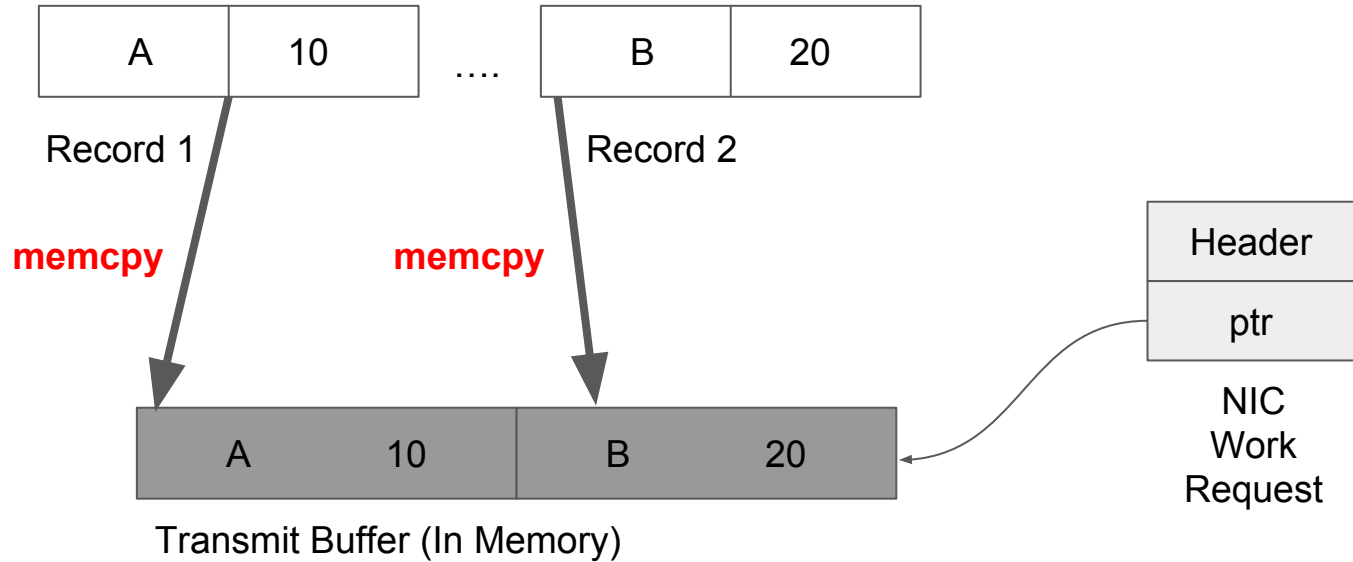
# Copy Out



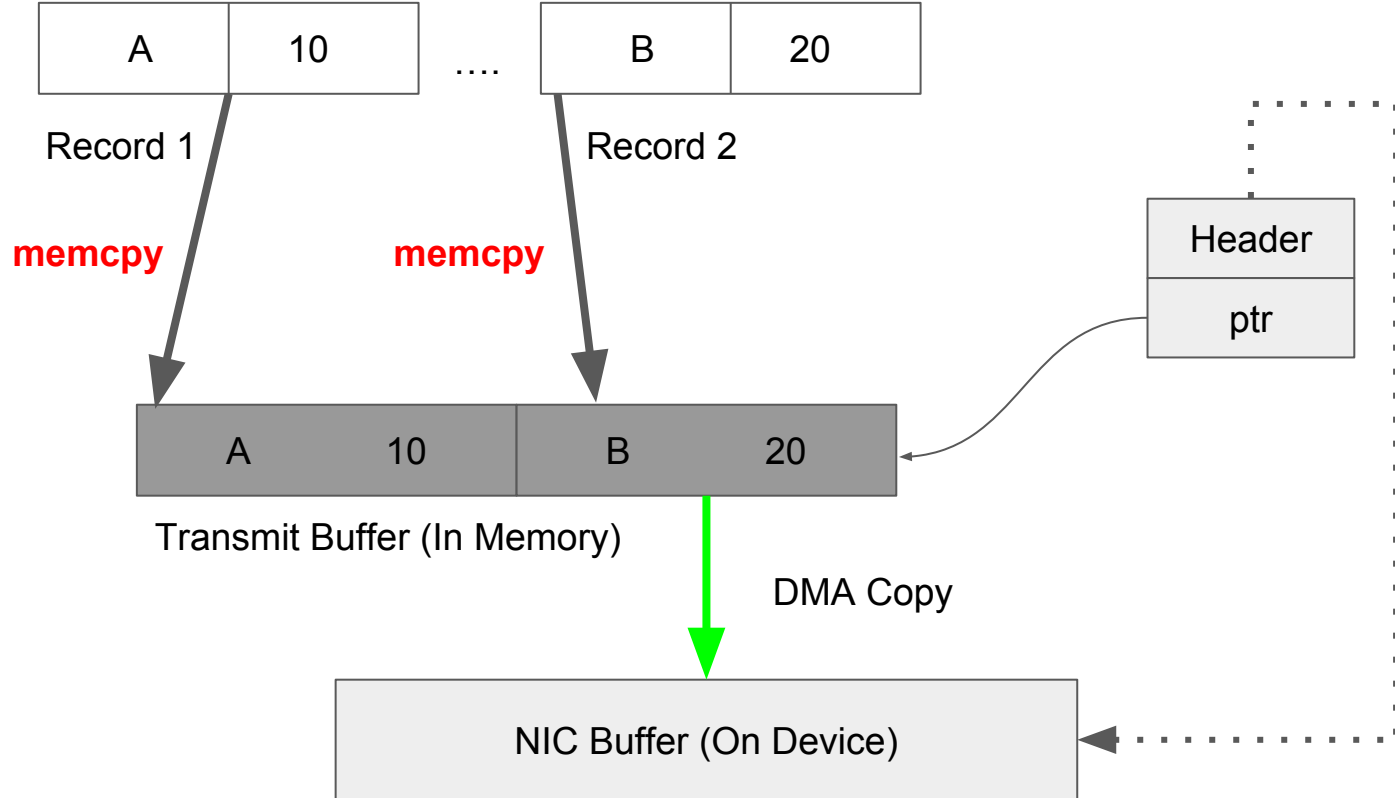
# Copy Out



# Copy Out



# Copy Out





# Takeaways - Copy Out

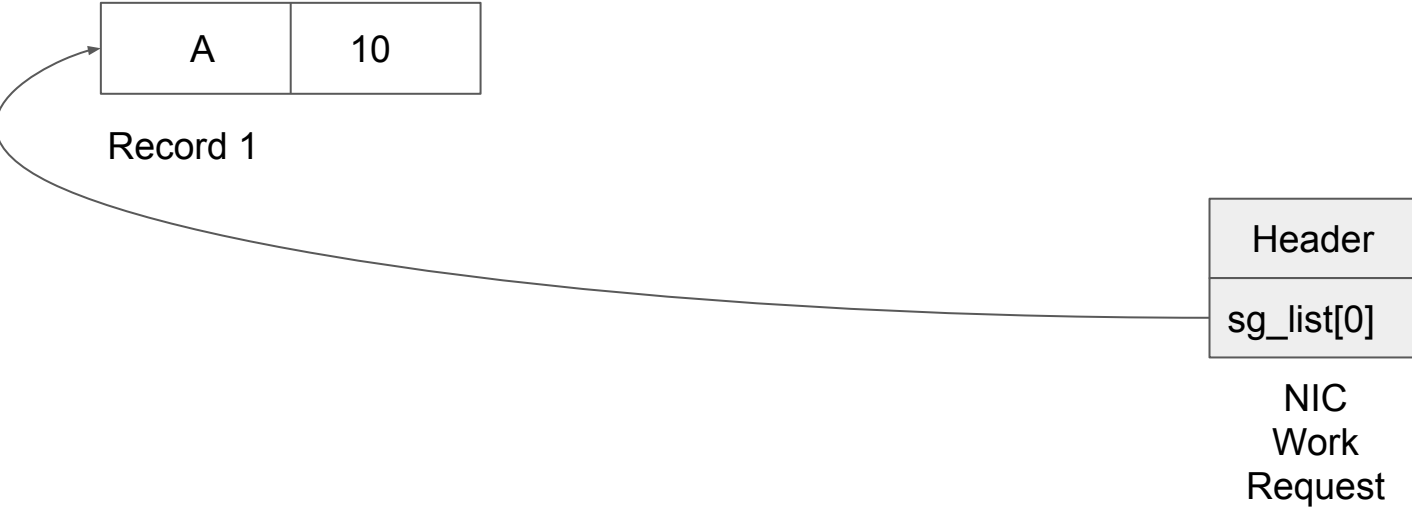
- Involves CPU cycles to copy
- Involves a memcpy - 2X more memory bandwidth

# Zero Copy

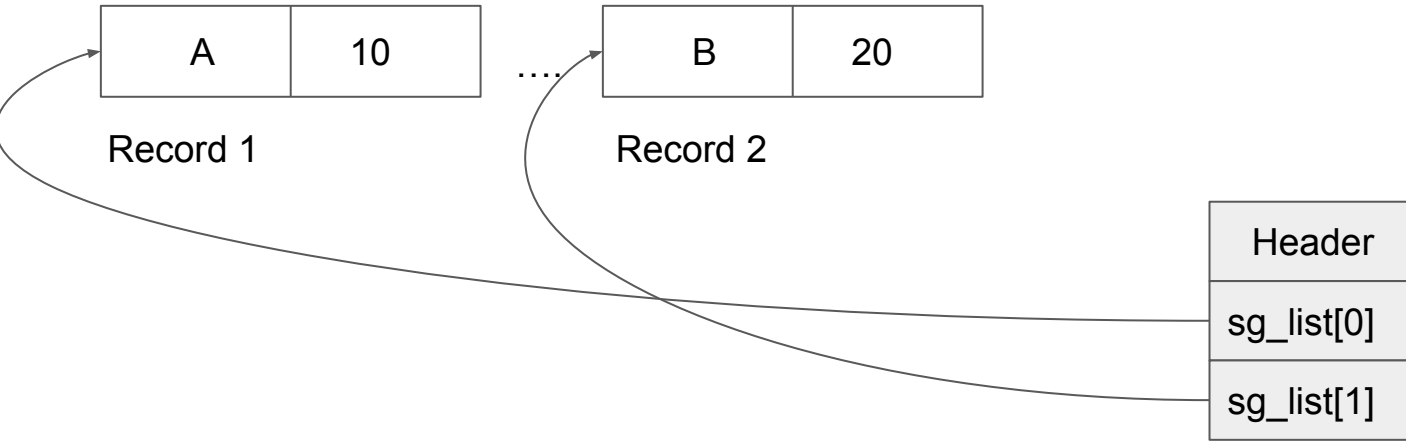
A	10
---	----

Record 1

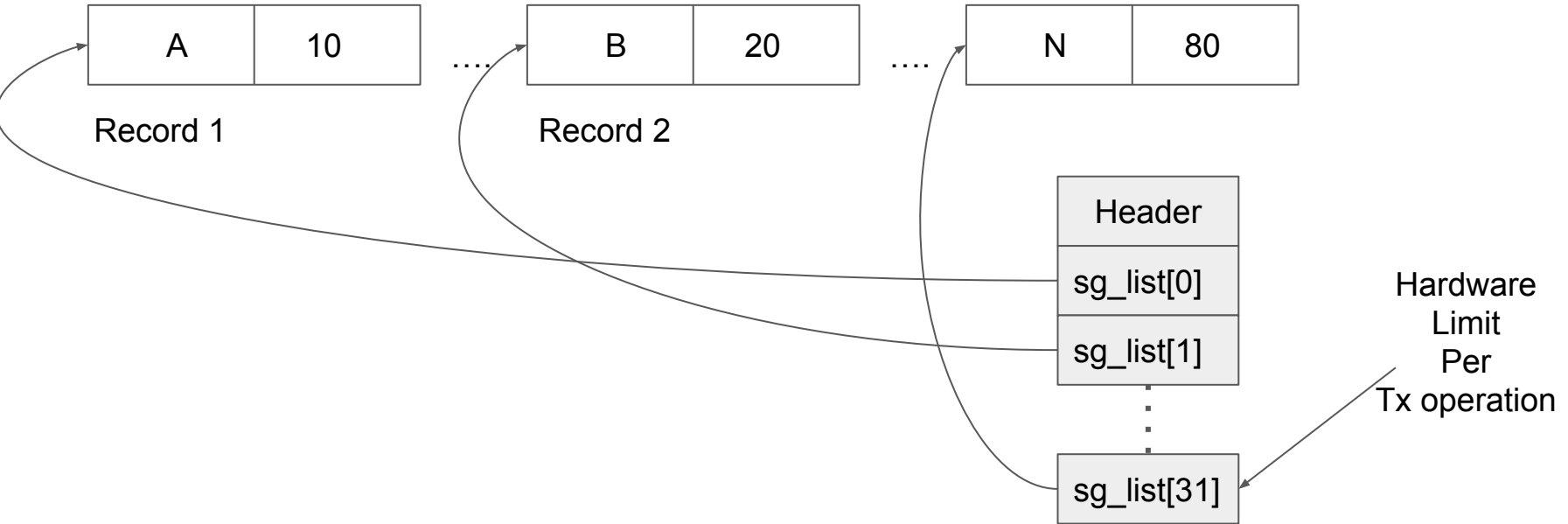
# Zero Copy



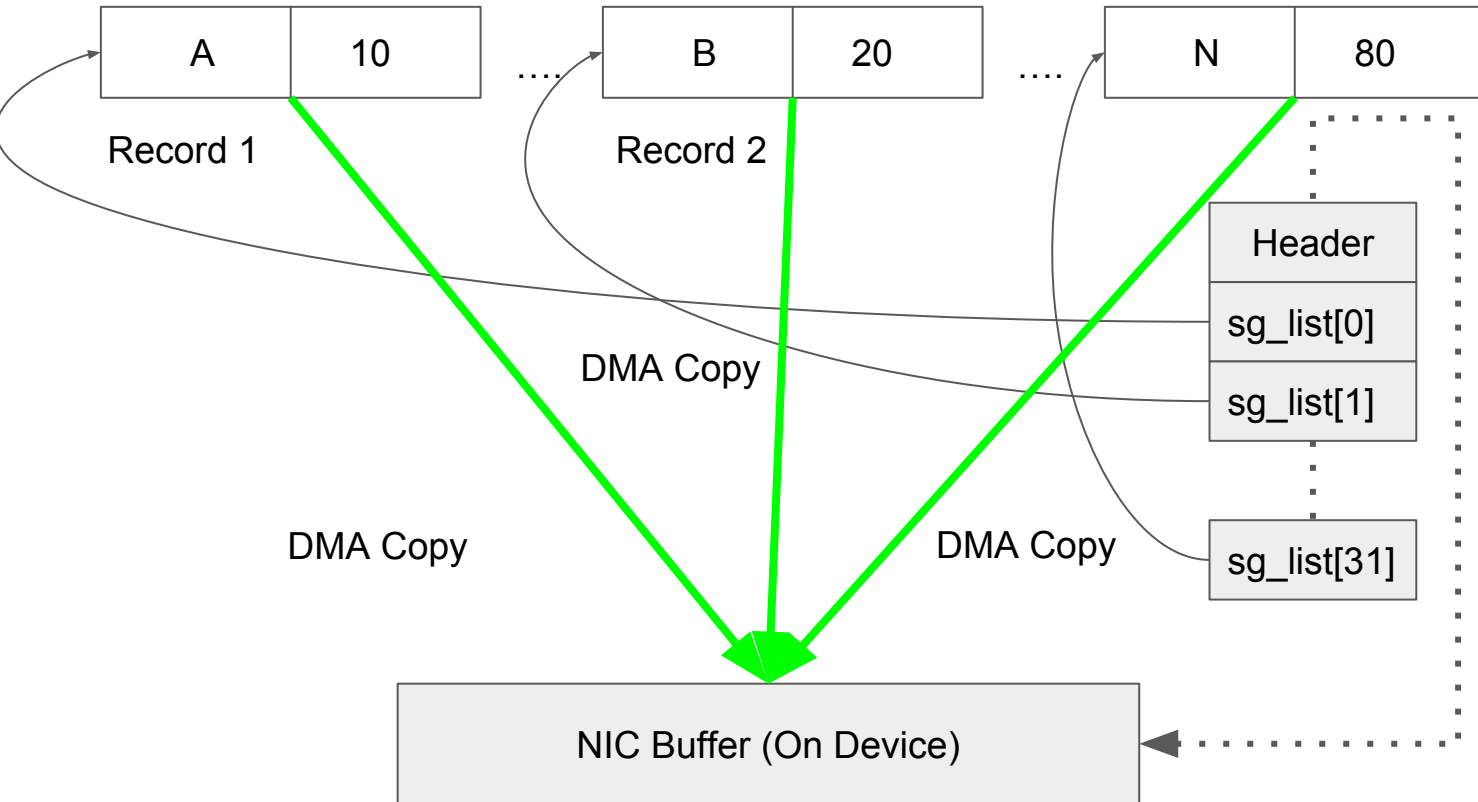
# Zero Copy



# Zero Copy



# Zero Copy



# Takeaways - Zero Copy

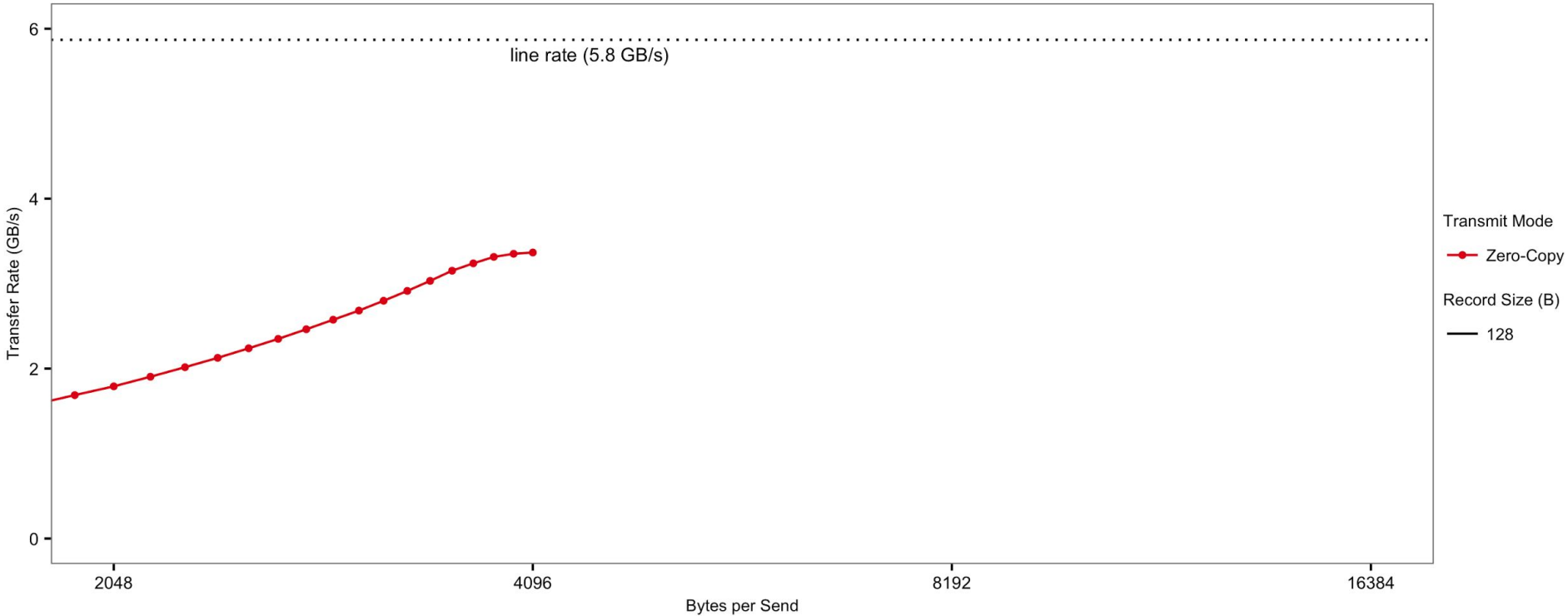
- No memcpy
- Limited number of records per transmission

# Experiments

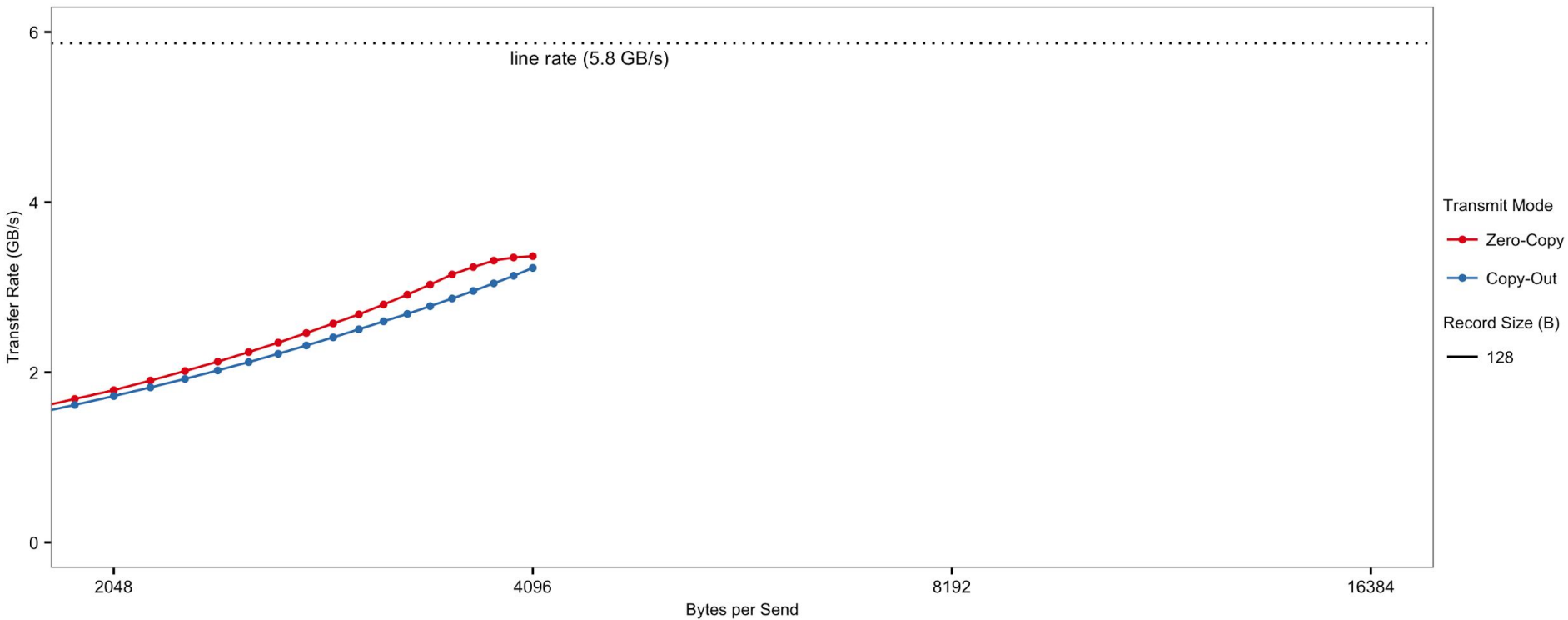
- Measuring effects of layout
  
- When to use which?



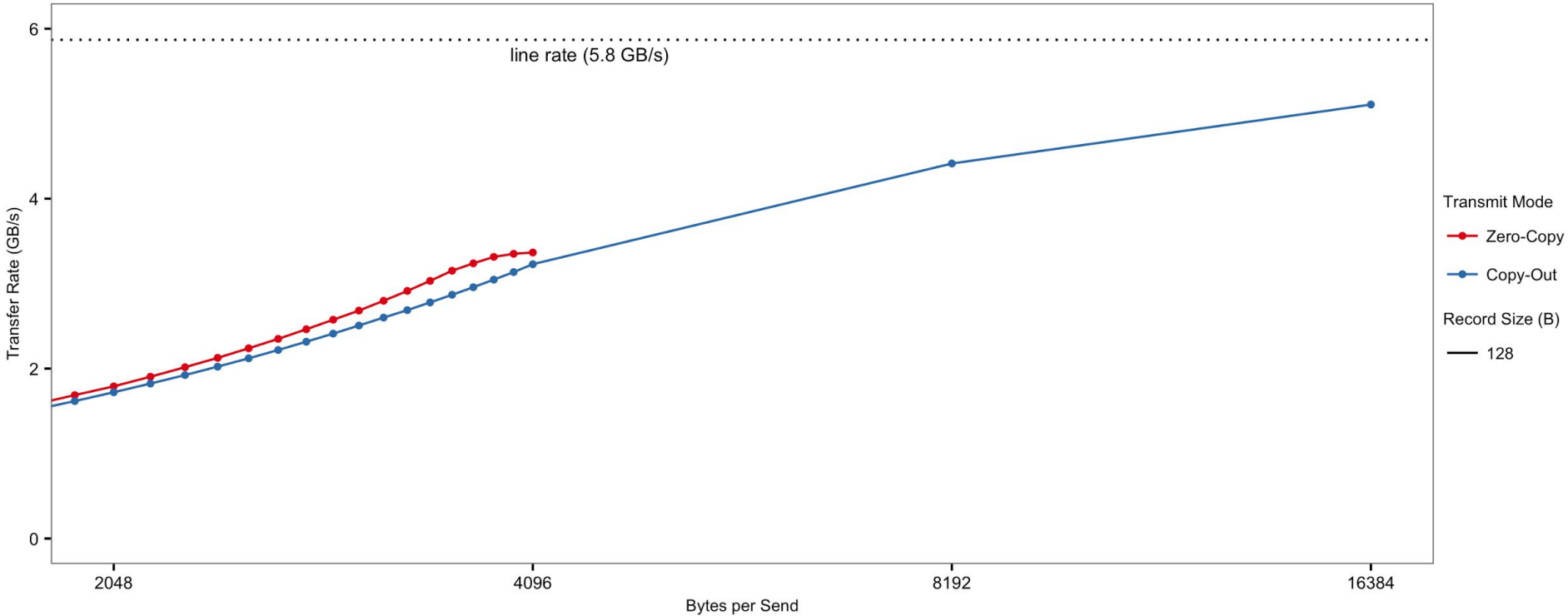
# Transmission Throughput (Zero Copy)



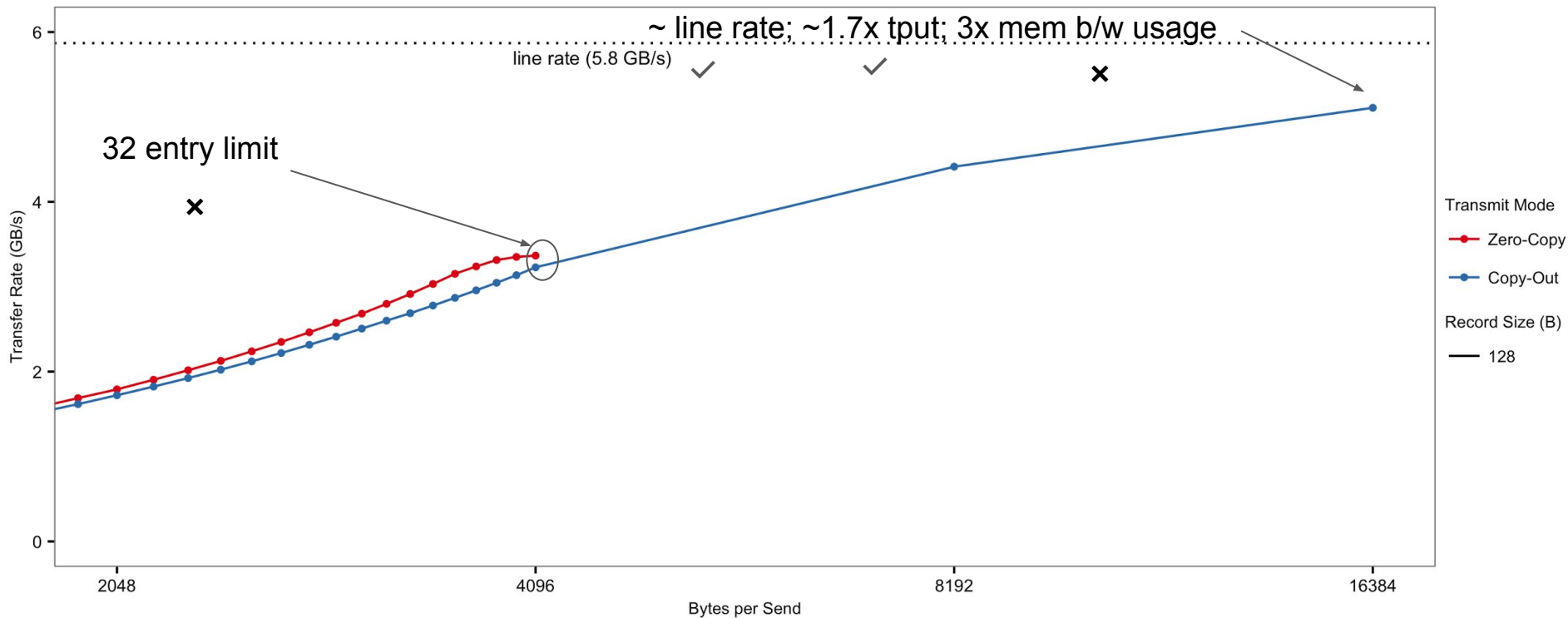
# Transmission Throughput (Comparison)



# Transmission Throughput (Comparison)



# Transmission Throughput (Comparison)



# What makes the NIC happy?

- Large Chunks of data - better throughput
- A few chunks of data that it can gather
- Stable data
  - Zero Copy requires records to be locked over transmission

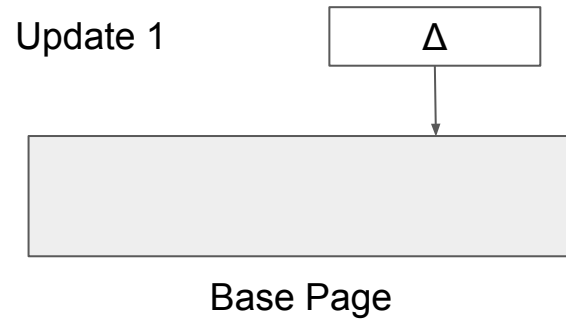
# Bw-Tree

[Bw-Tree - Levandoski et al., 2013]

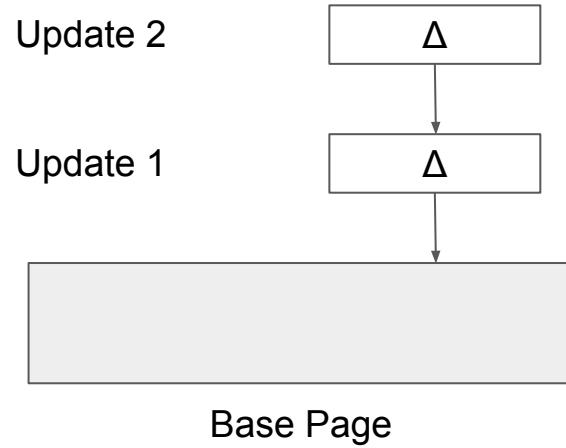


Base Page

# Bw-Tree

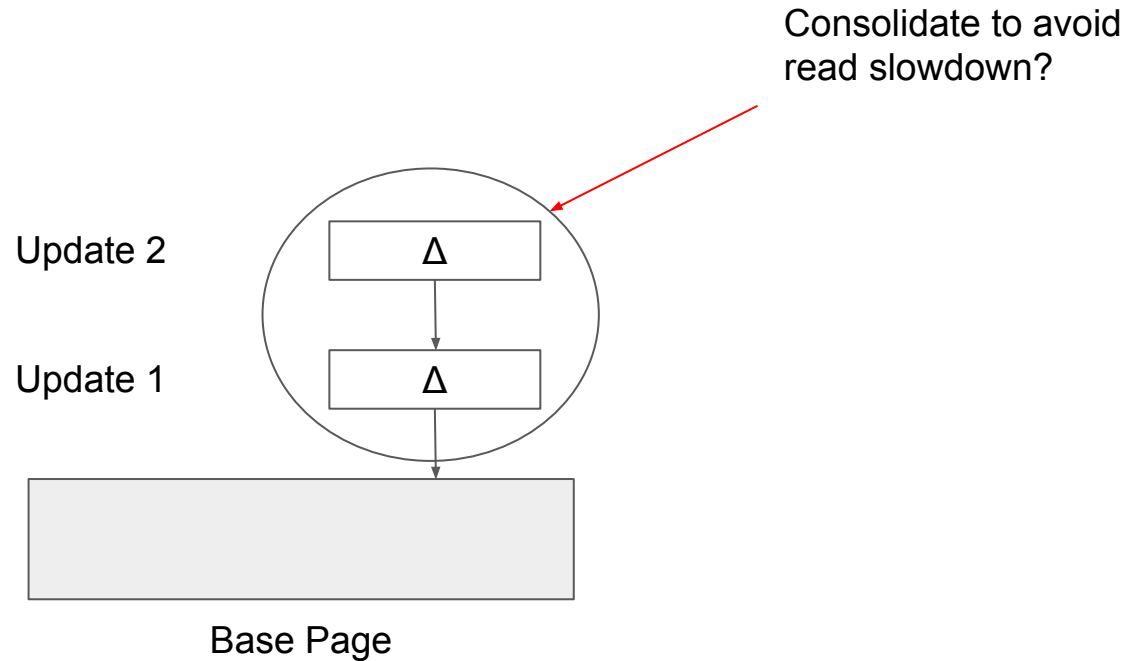


# Bw-Tree

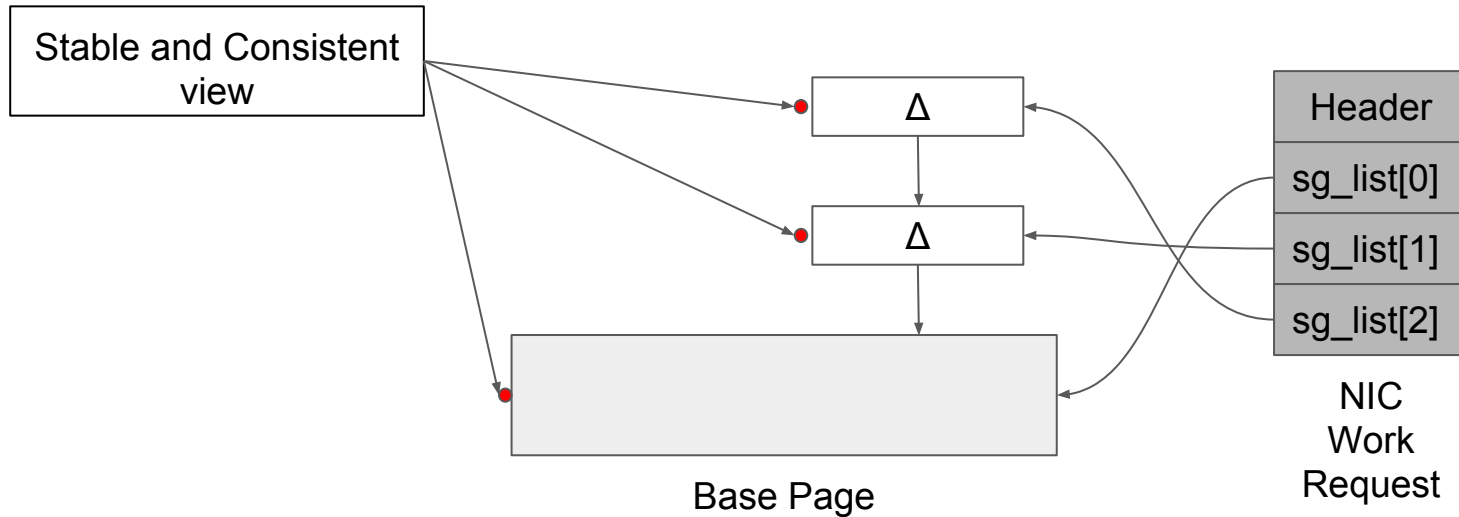




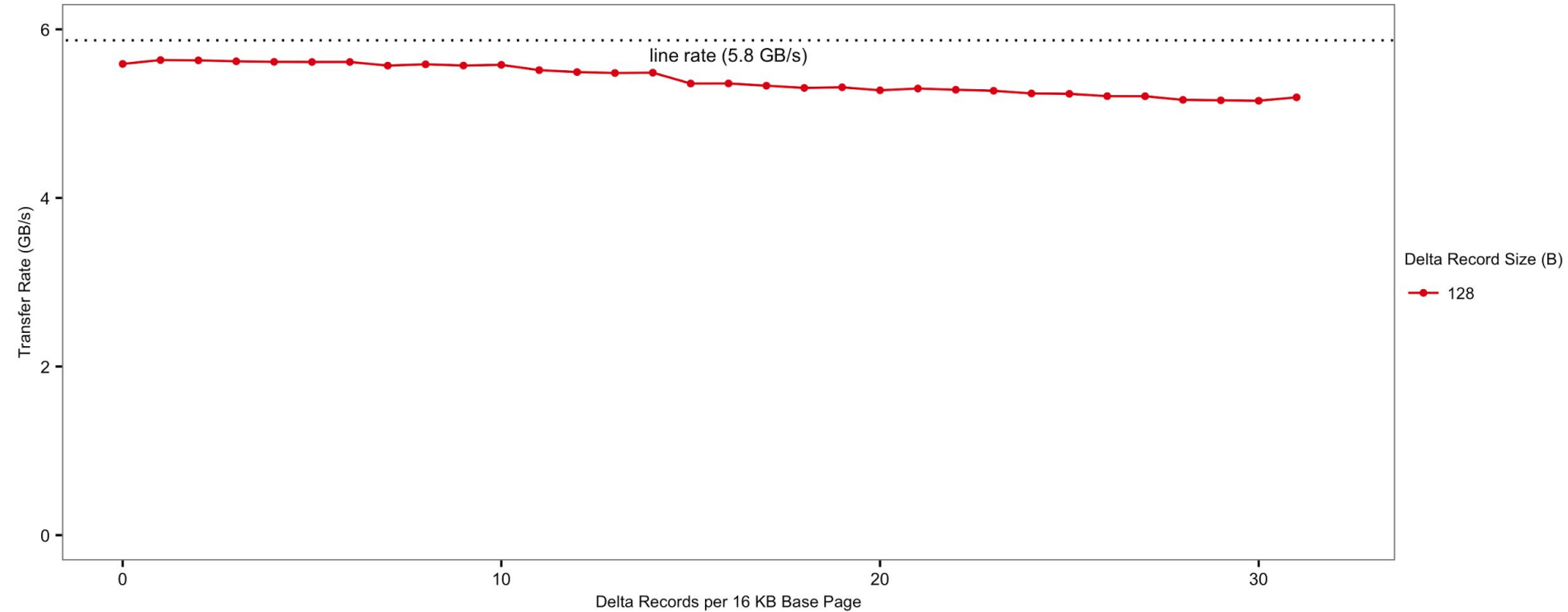
# Bw-Tree



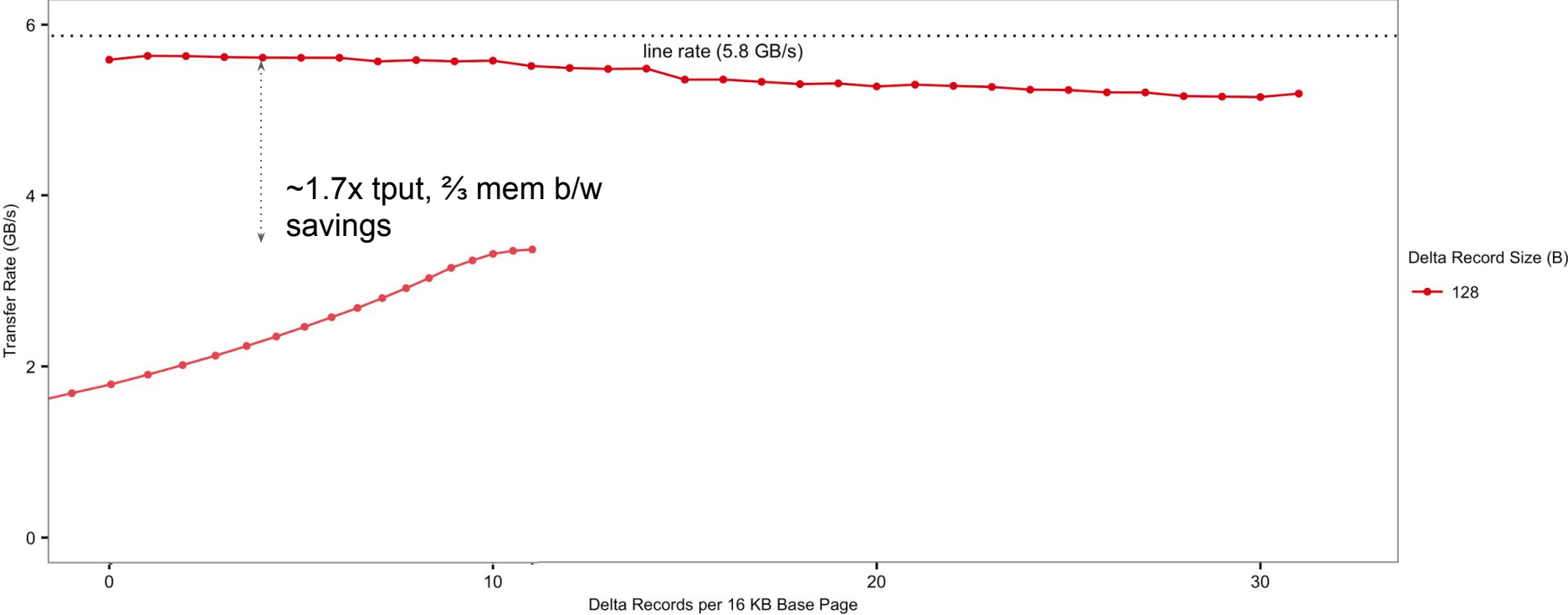
# Bw-Tree



# Deltas make NIC happy



# Deltas make NIC happy



# Conclusion

- Does DB layout matter for NIC performance?
  - Yes. If you care about mem b/w and CPU cycles.
- No updates in place structures like Bw-Tree gives us the best of both worlds by:
  - Transmitting bigger chunks directly aiding throughput
  - Transmitting smaller chunks directly saving memory b/w and CPU consolidation costs

# QnA

Source:

To Copy or Not to Copy: Making In-Memory Databases Fast on Modern NICs:



- A.Kesavan, R. Ricci, R. Stutsman

# CPU Overheads

