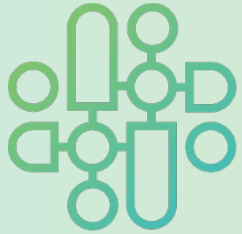


SC23
Denver, CO | i am hpc.

Enhancing Metadata Transfer Efficiency: Unlocking the Potential of DAOS in the ADIOS context

Ranjan Sarpangala Venkatesh, **Greg Eisenhauer***, Scott Klasky, Ada Gavrilovska

Georgia Institute of Technology

U.S. DEPARTMENT OF ENERGY | Office of Science

Introduction

Exponential increase in amount of data

Focus on increasing I/O bandwidth and concurrency

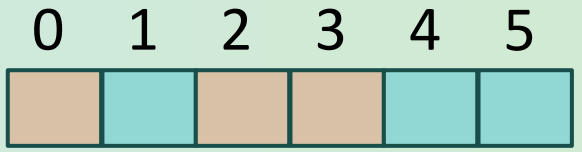
Metadata costs have NOT received same attention

Background - ADIOS

- **I/O library**
- Scalable
- **Global Multidimensional Array**
- **Timestep based**

Timestep 1

Var A



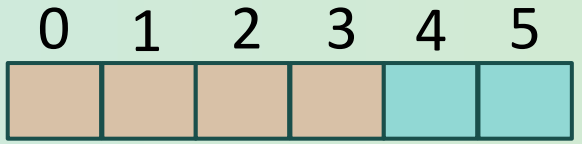
Metadata

Rank 1 : VarA - [0,1] and [4,5]

Rank 2 : VarA - [2,3]

Timestep 2

Var A



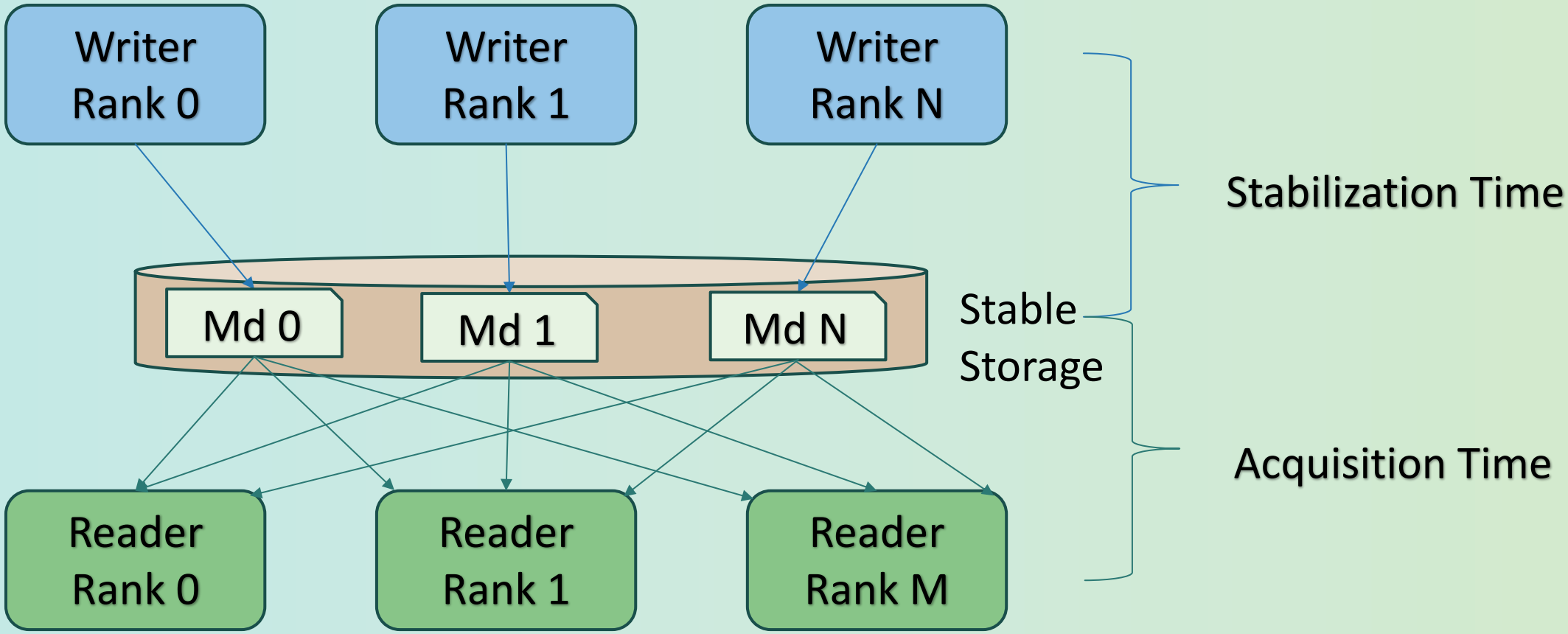
Metadata

Rank 1 : VarA - [0,3]

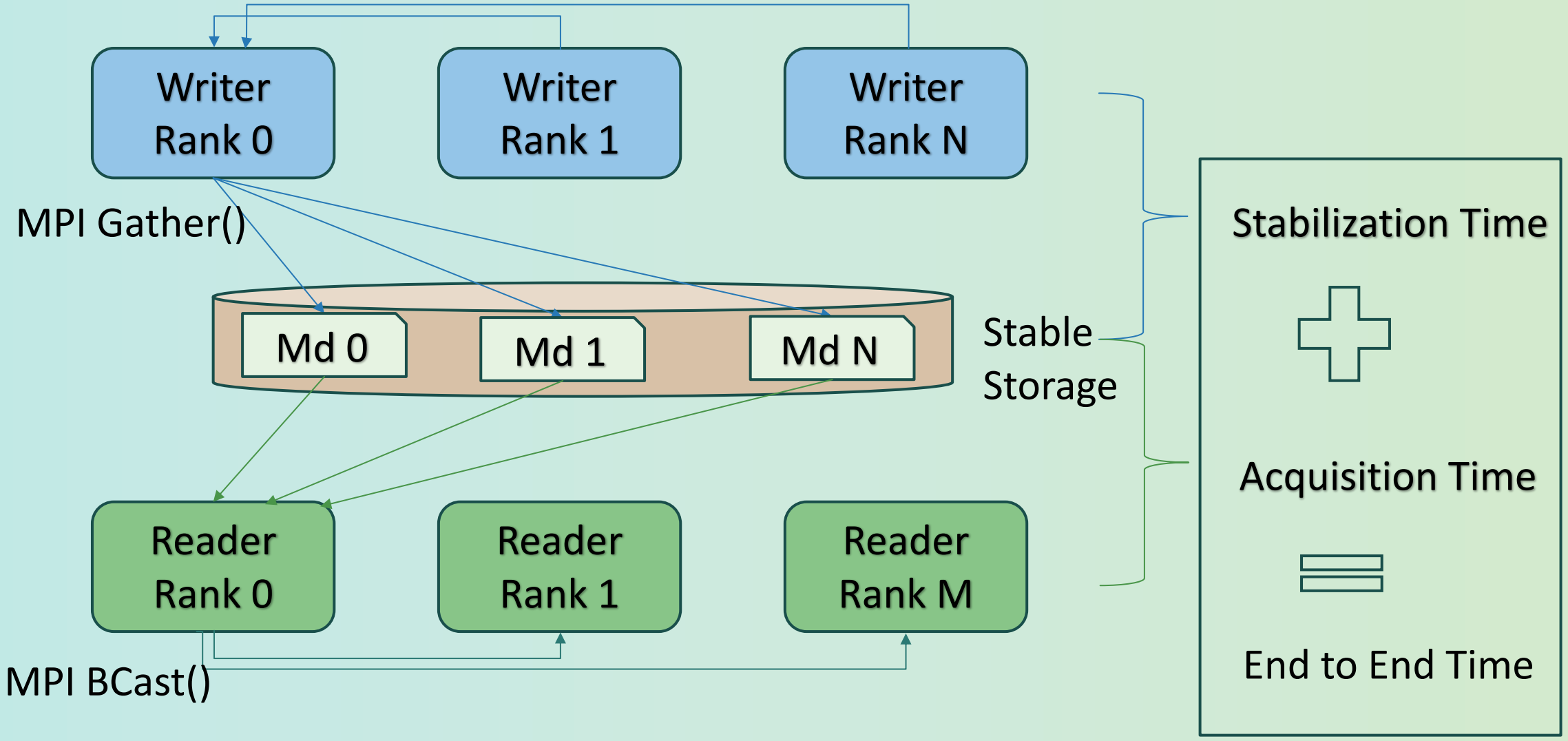
Rank 2 : VarA - [4,5]

Rank 1 Rank 2

ADIOS metadata transfer in a timestep

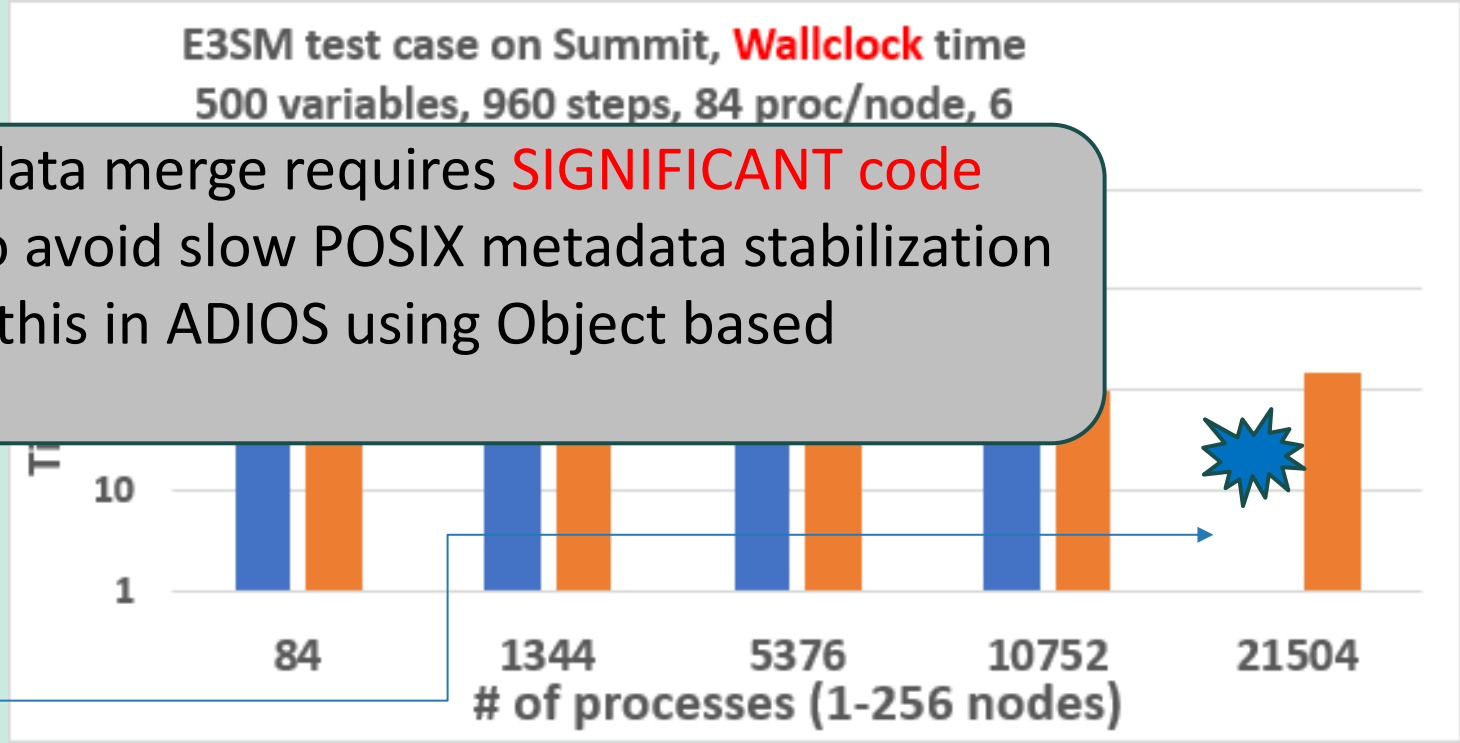


ADIOS metadata transfer in a timestep



E3SM Metadata Stabilization – Expensive?

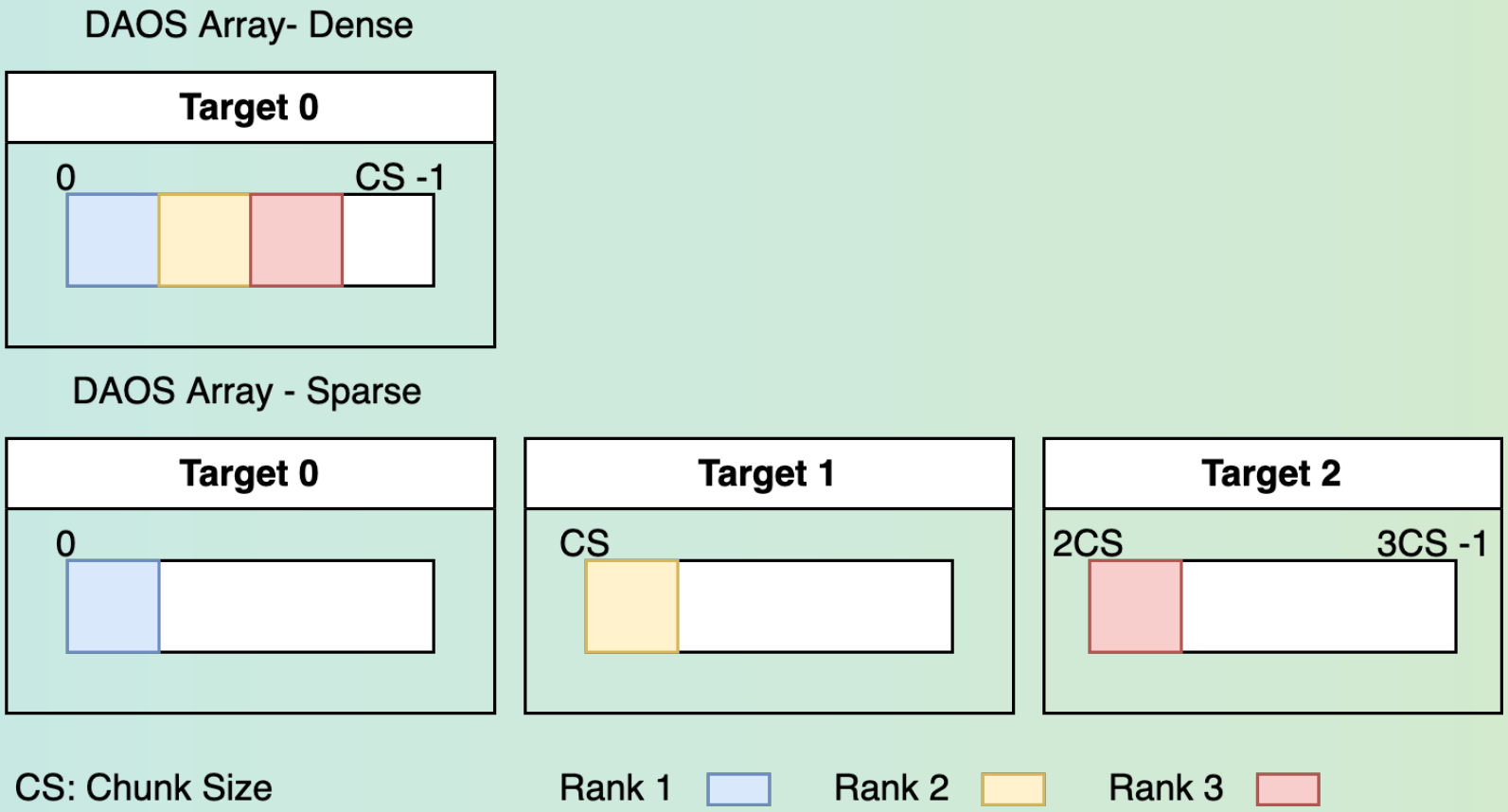
- Both original and app-level data merge write SAME out of E3SM data
- App-level data mer reduces ADIOS me aggregating data before passing to ADIOS
- Original is prohibitively expensive at scale



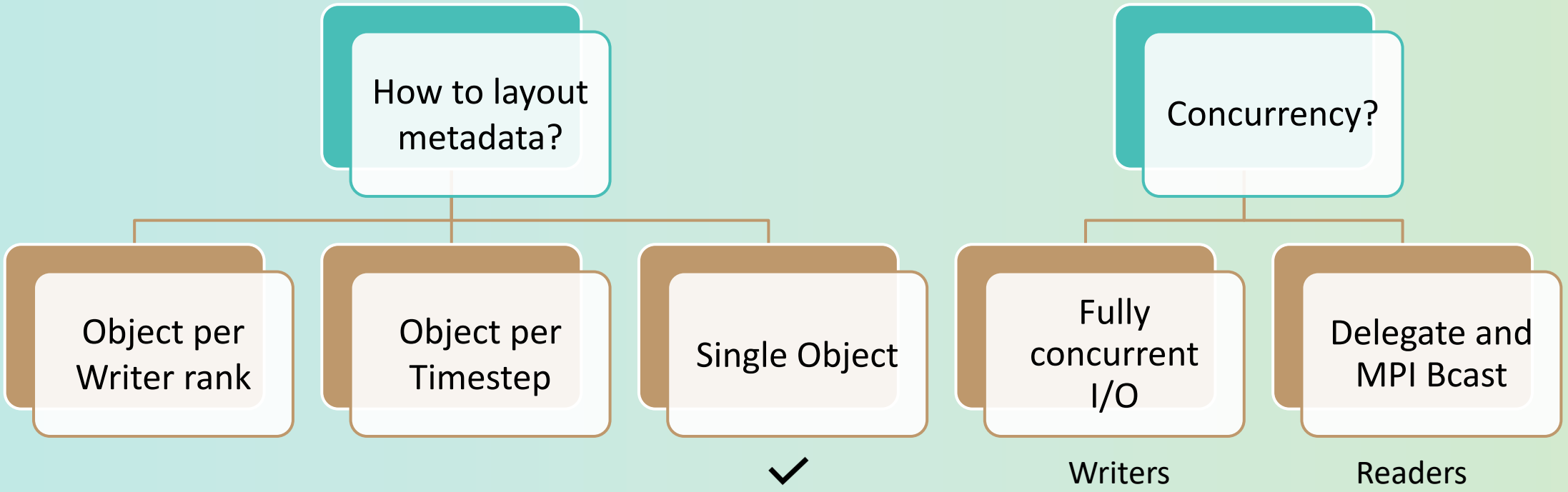
- App-level data merge requires SIGNIFICANT code changes, to avoid slow POSIX metadata stabilization
- Can we fix this in ADIOS using Object based storage?

DAOS APIs

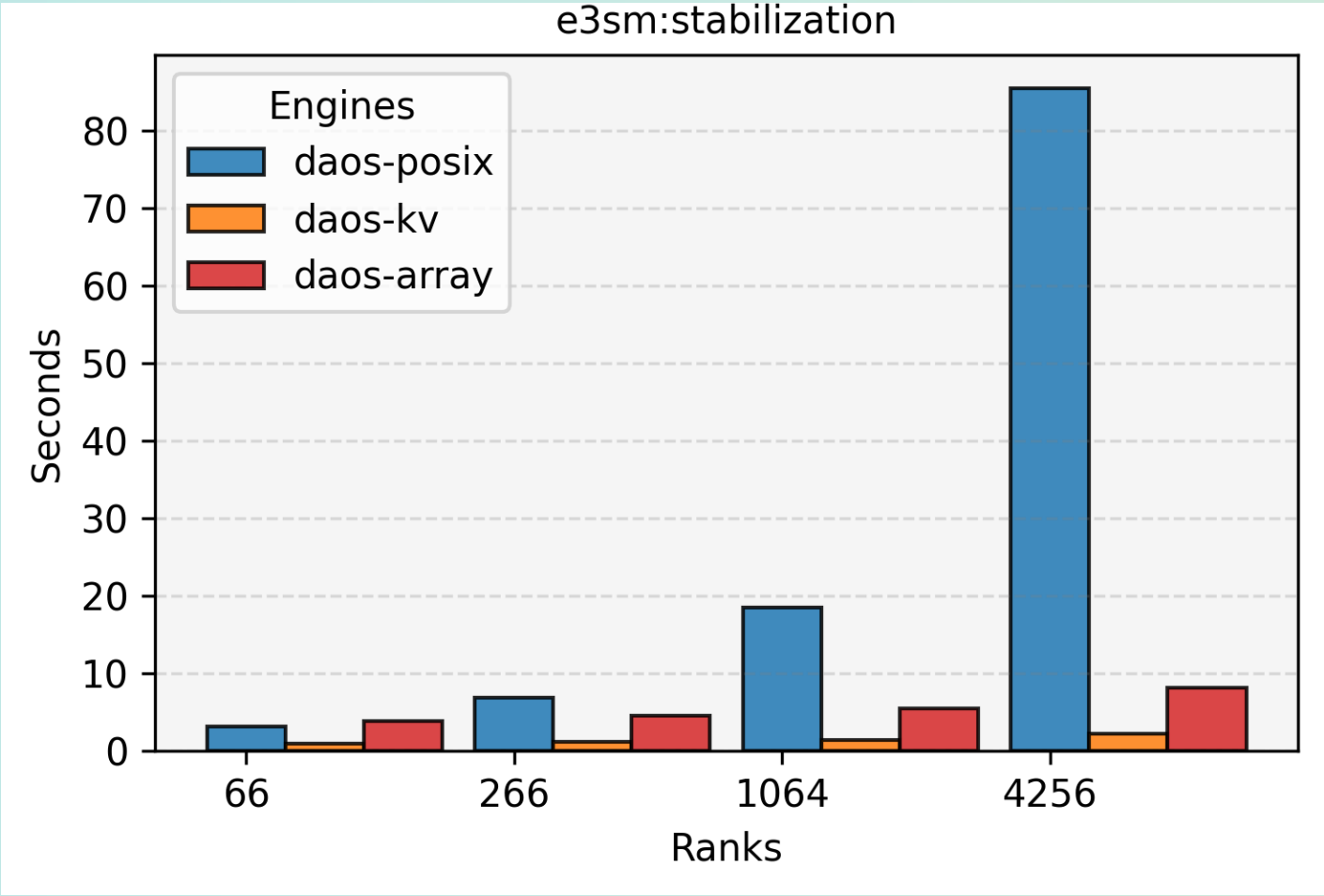
- DAOS POSIX
- DAOS Array
- DAOS Key Value



Design Options

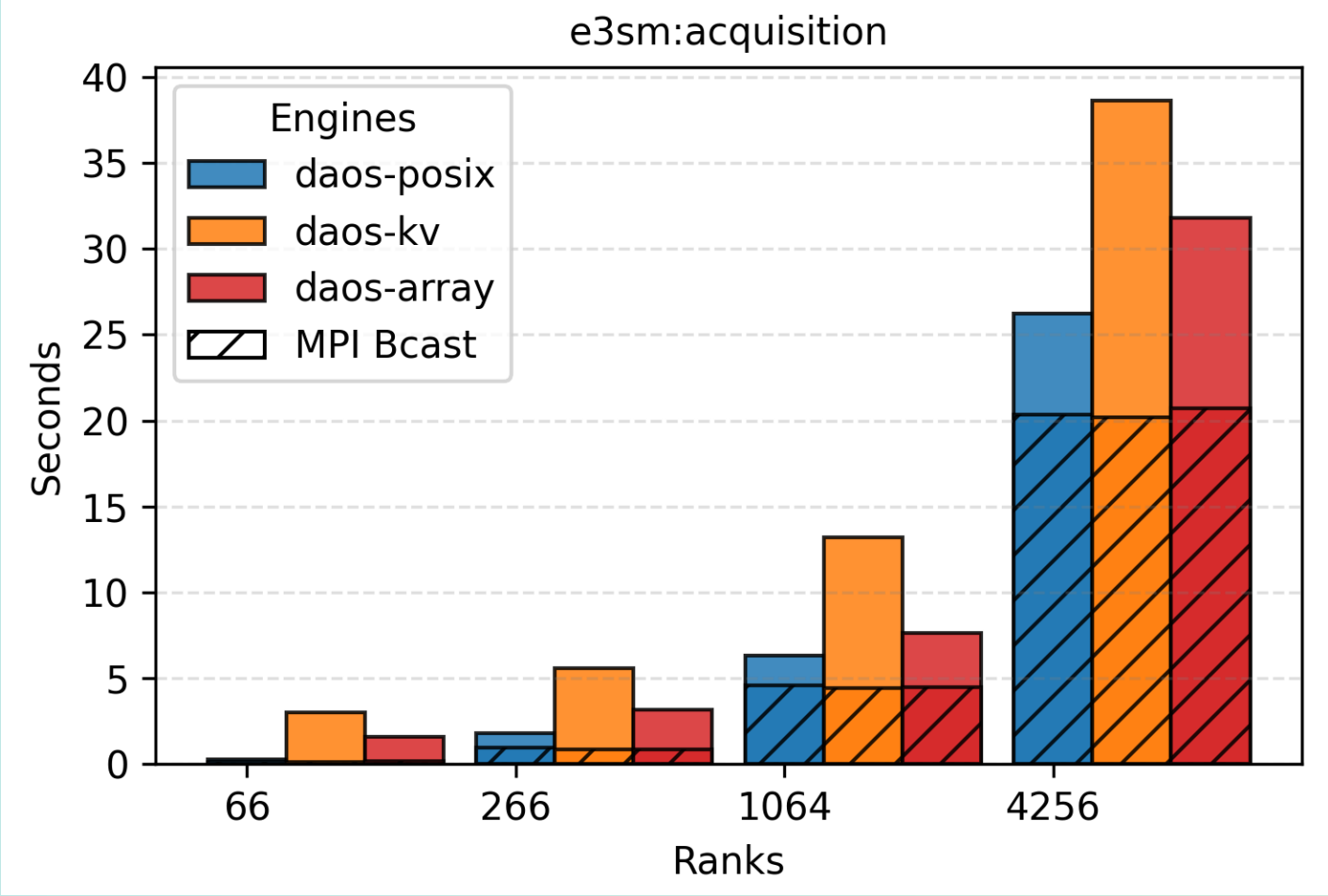


(E3SM – 56KB) – Stabilization Time



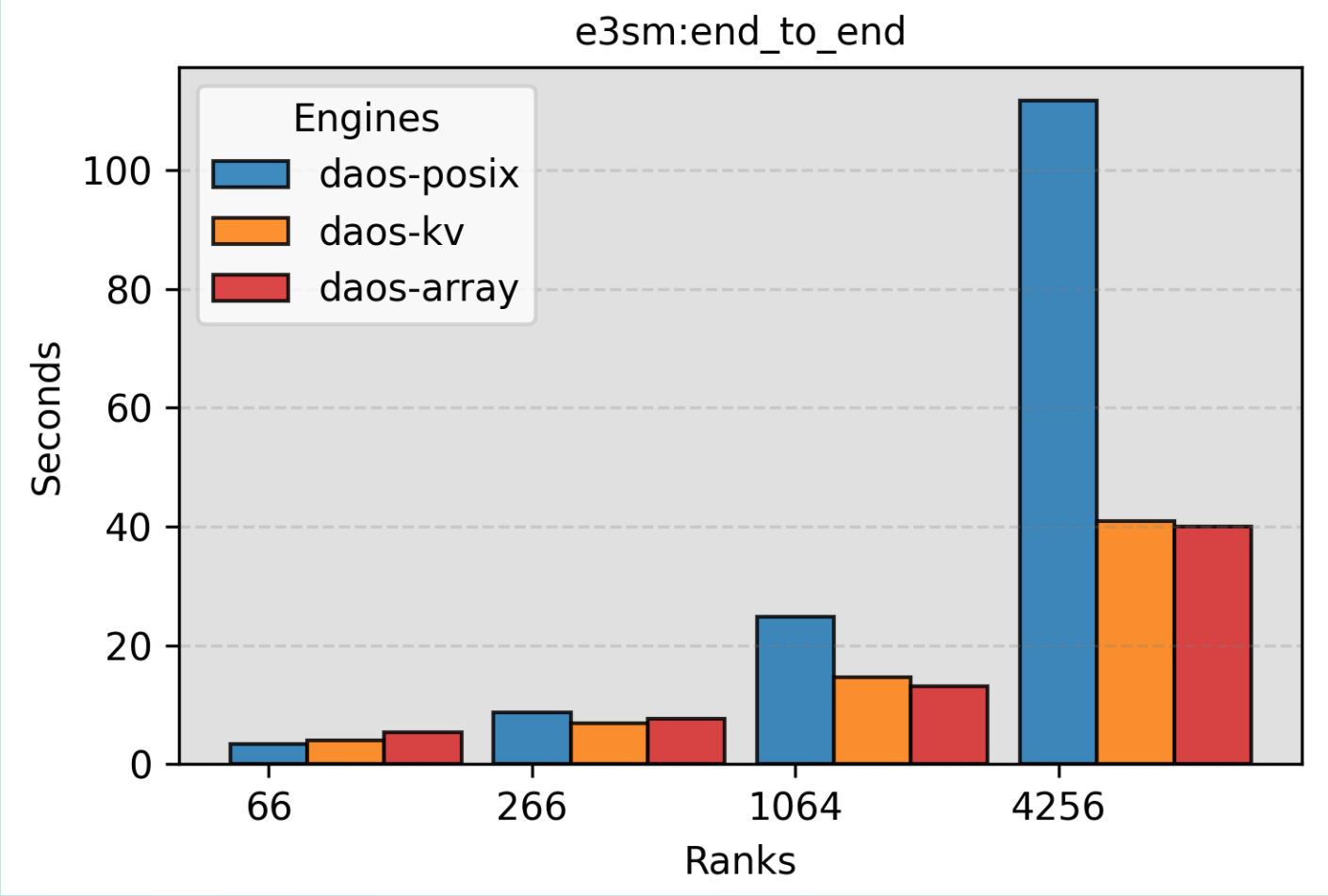
- Stabilization time of E3SM is known to dominate overall time
- POSIX is unsustainable at scale
- KV more than 10X faster than POSIX

(E3SM – 56KB) – Acquisition Time



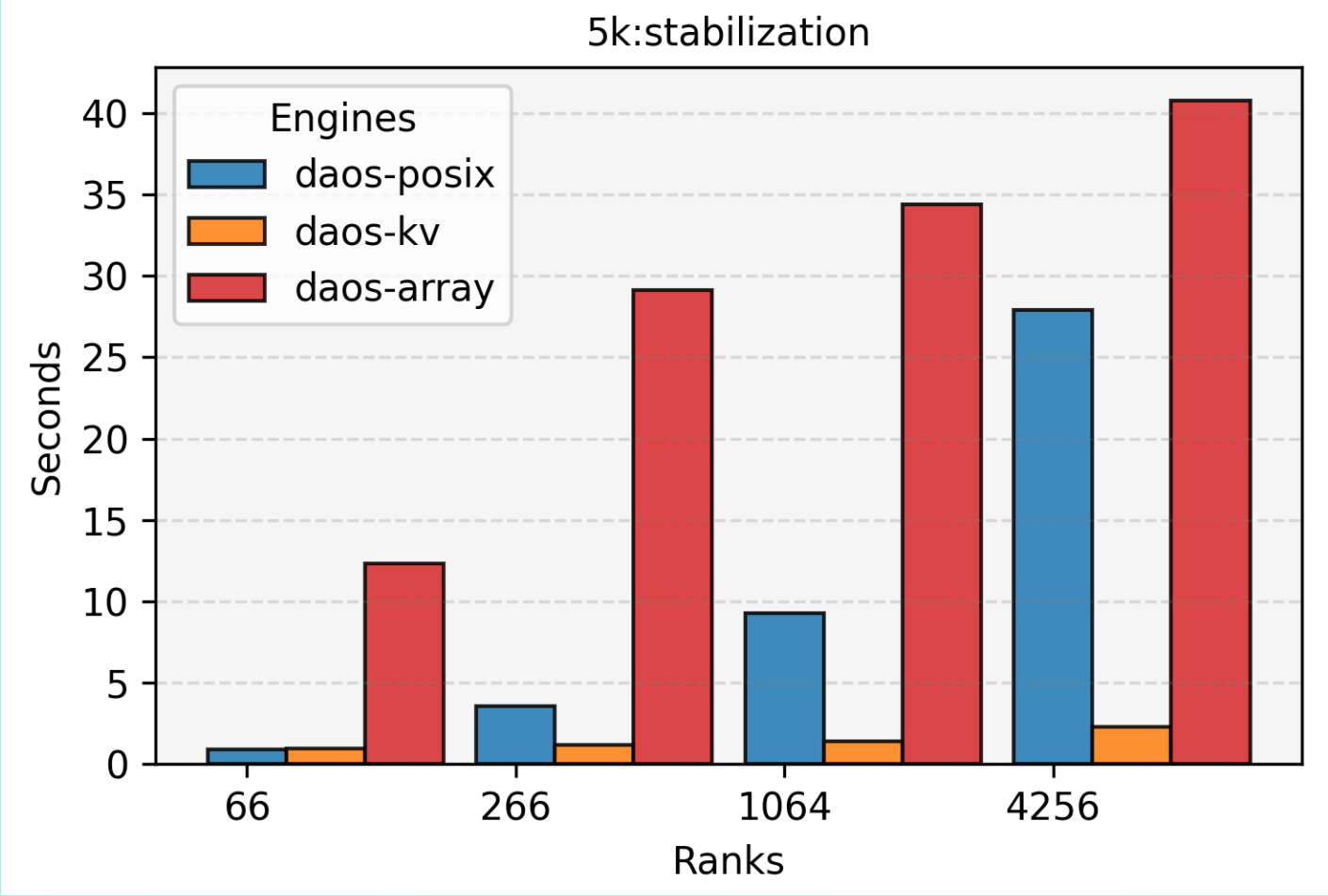
- Here KV is slower than Array and POSIX
- No single winner across stabilization and acquisition

(E3SM – 56KB) – End to End Time



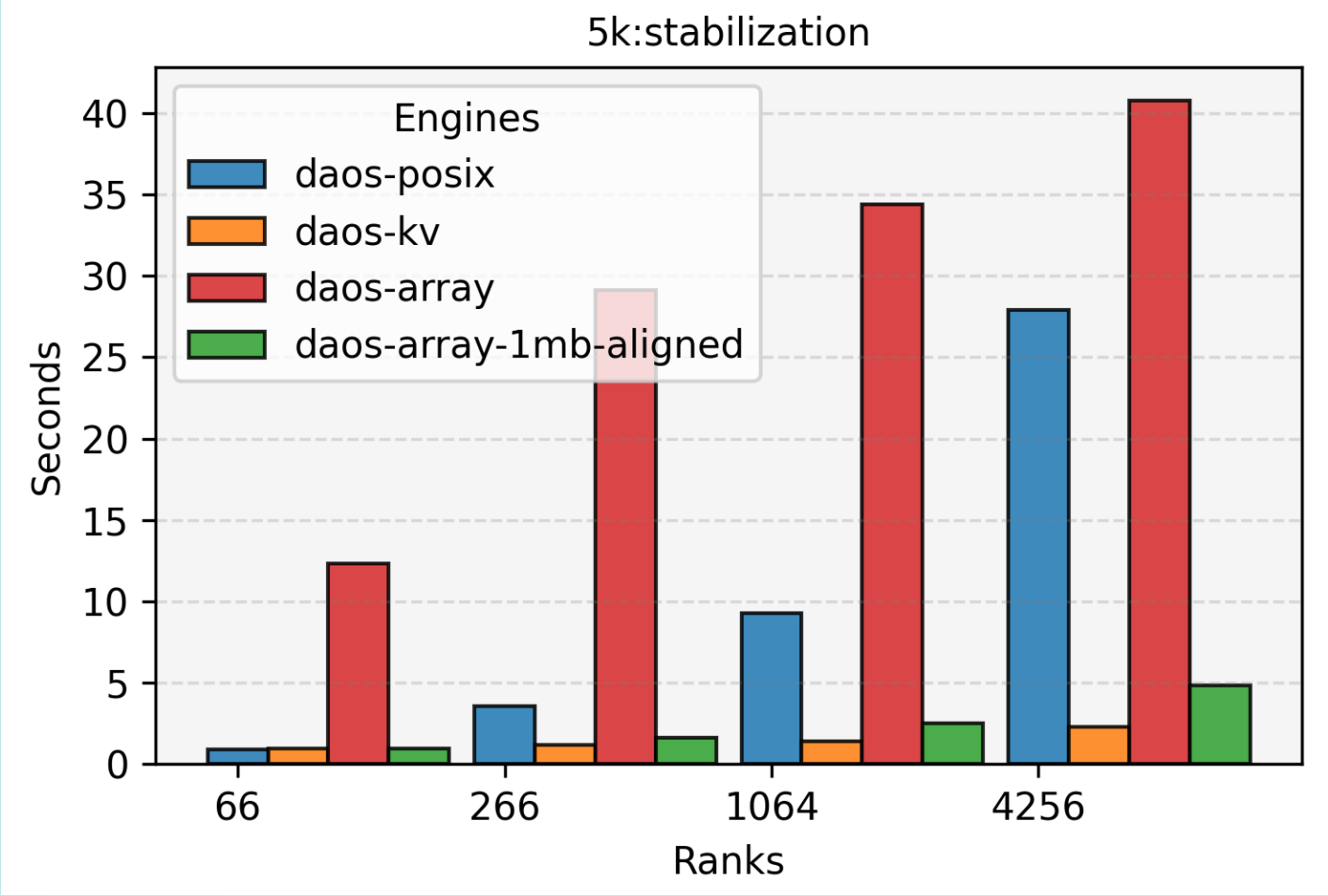
- Both KV and Array are 2.75X faster POSIX
- DAOS objects a win for E3SM

Small Metadata Size (5KB) – Stabilization Time



- Array 10X slower than KV and worse than POSIX
- Uneven distribution of of metadata across DAOS targets
- For Array – Data alignment matters

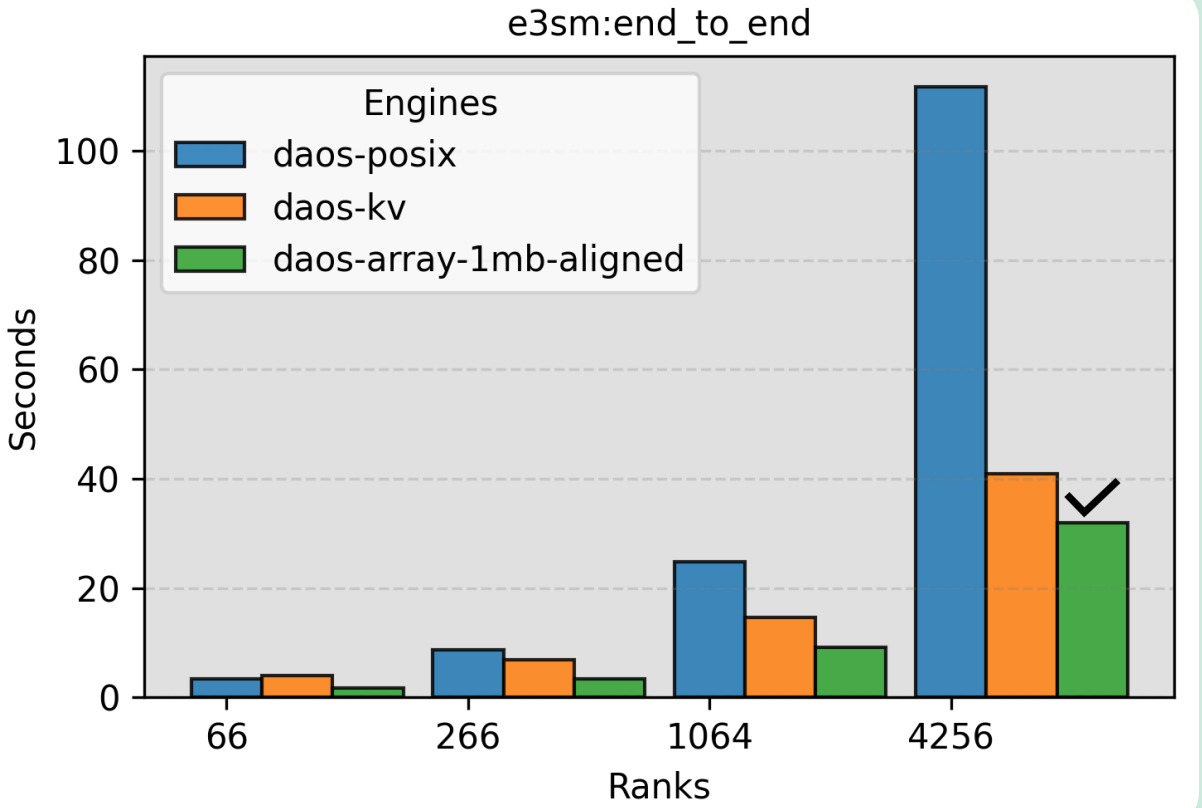
Small Metadata Size (5KB) – Stabilization Time(cont'd)

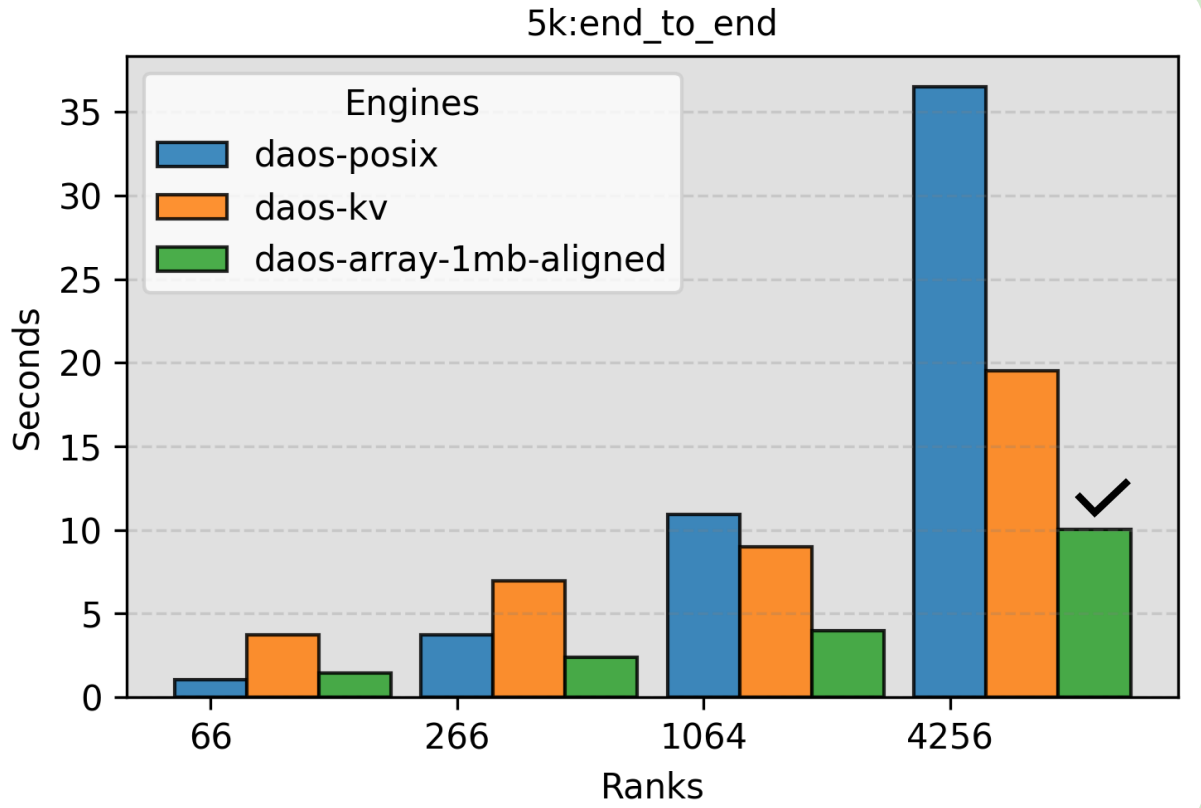


Array Chunk Aligned

- Careful alignment ensured improved metadata distribution
- 10X faster than Array
- No wasted Space

End to End transfer time





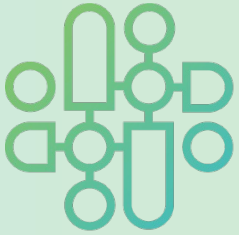
Stabilization Time: KV < Array Chunk Aligned

Acquisition Time: KV >> Array Chunk Aligned

Conclusion

Clear Opportunity

- End to End Time - DAOS Object APIs 2-3X faster than POSIX
- Preliminary work
- Multitude of Design Options
- More Design Options - How to provide ADIOS timestep?
  - Extend existing objects
  - Snapshot and reuse objects
- Ongoing work – Stay tuned!



SC23

Denver, CO | i am hpc.

Thank you