Tuning and Analyzing Sonexion Performance

CUG 2014

Mark Swan, Cray Inc.

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Performance considerations

Where data is on the OST

How data gets to the OST

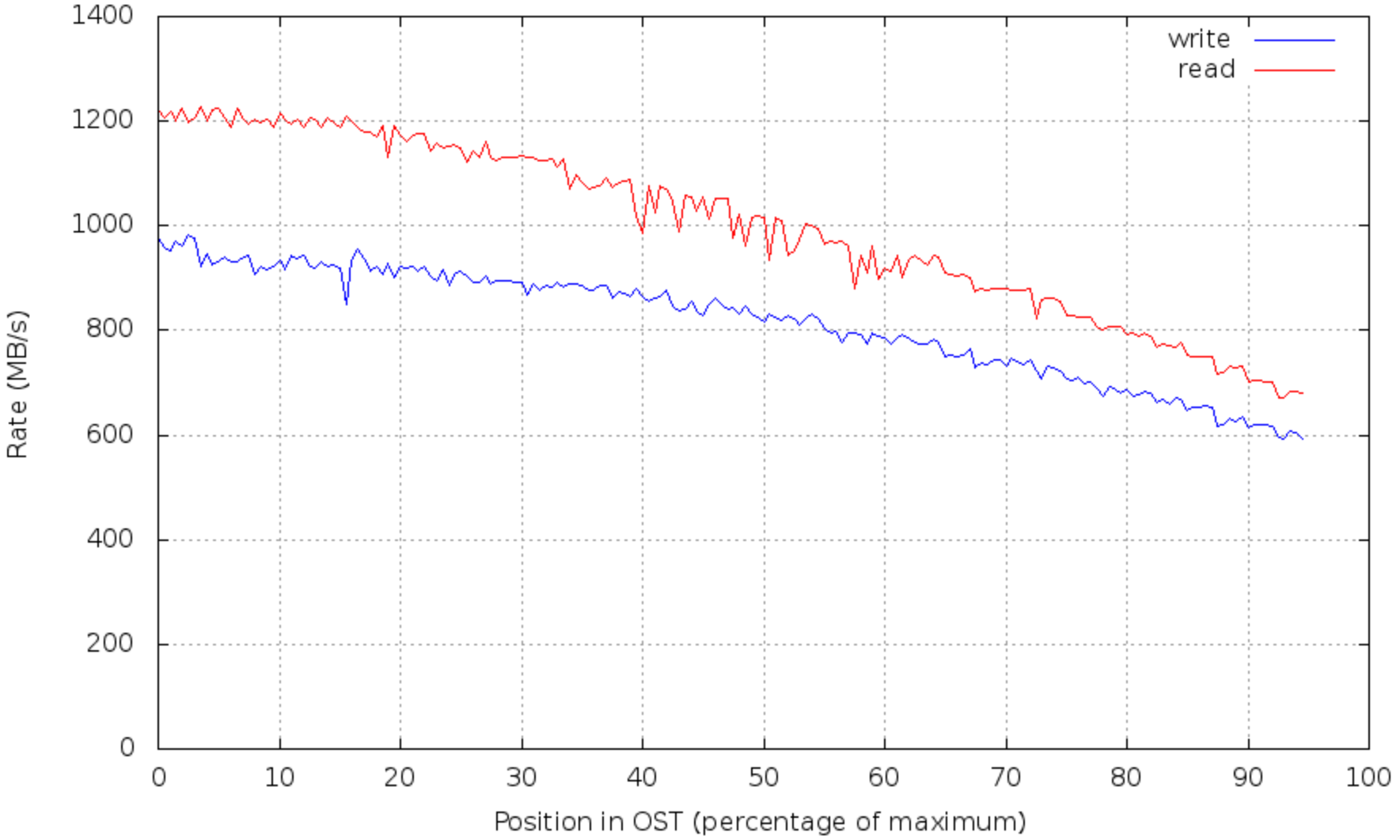
How data is arranged on the OST

Where data is on the OST

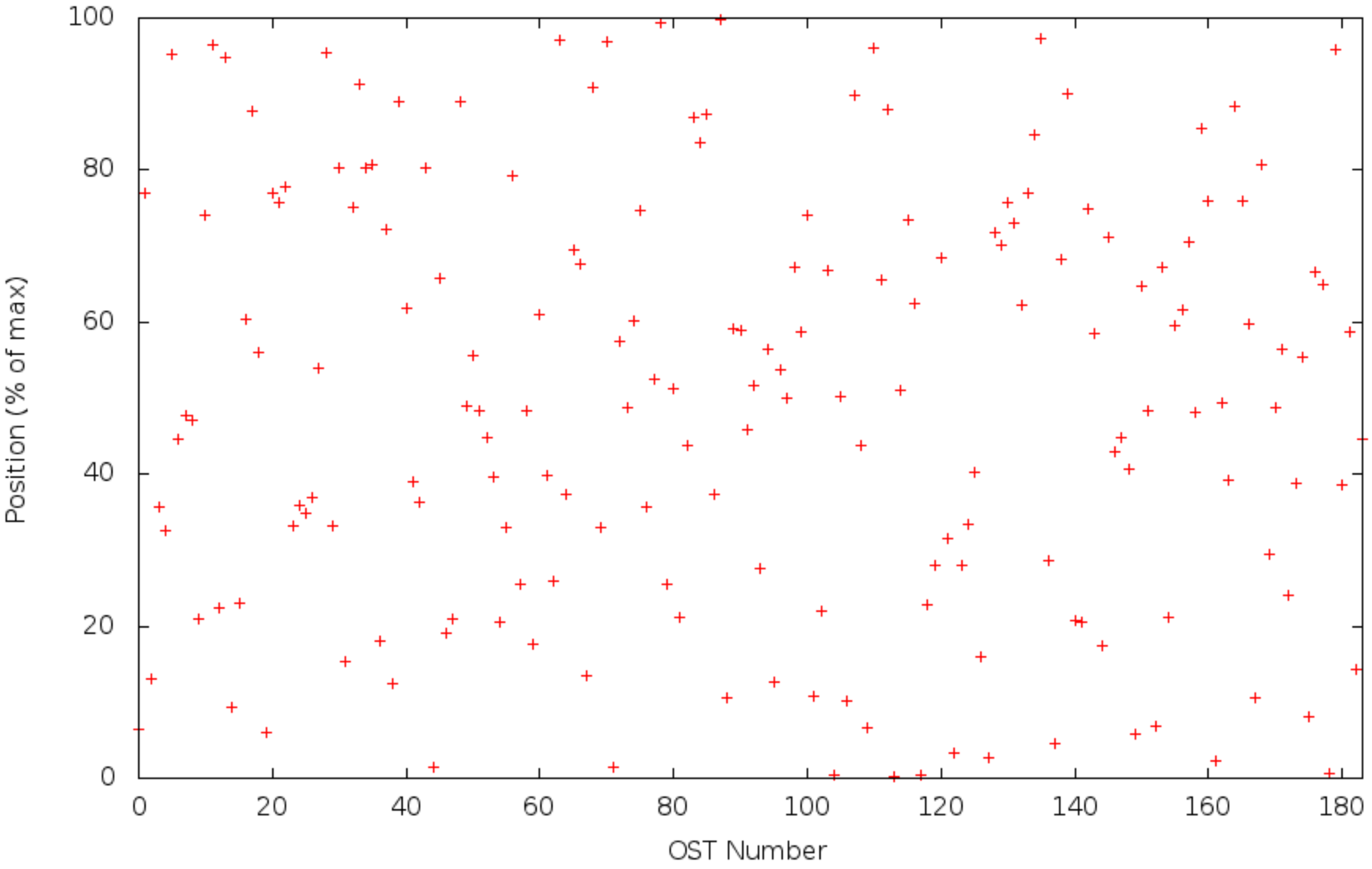
Fast edge and slow edge

Pre-existing data

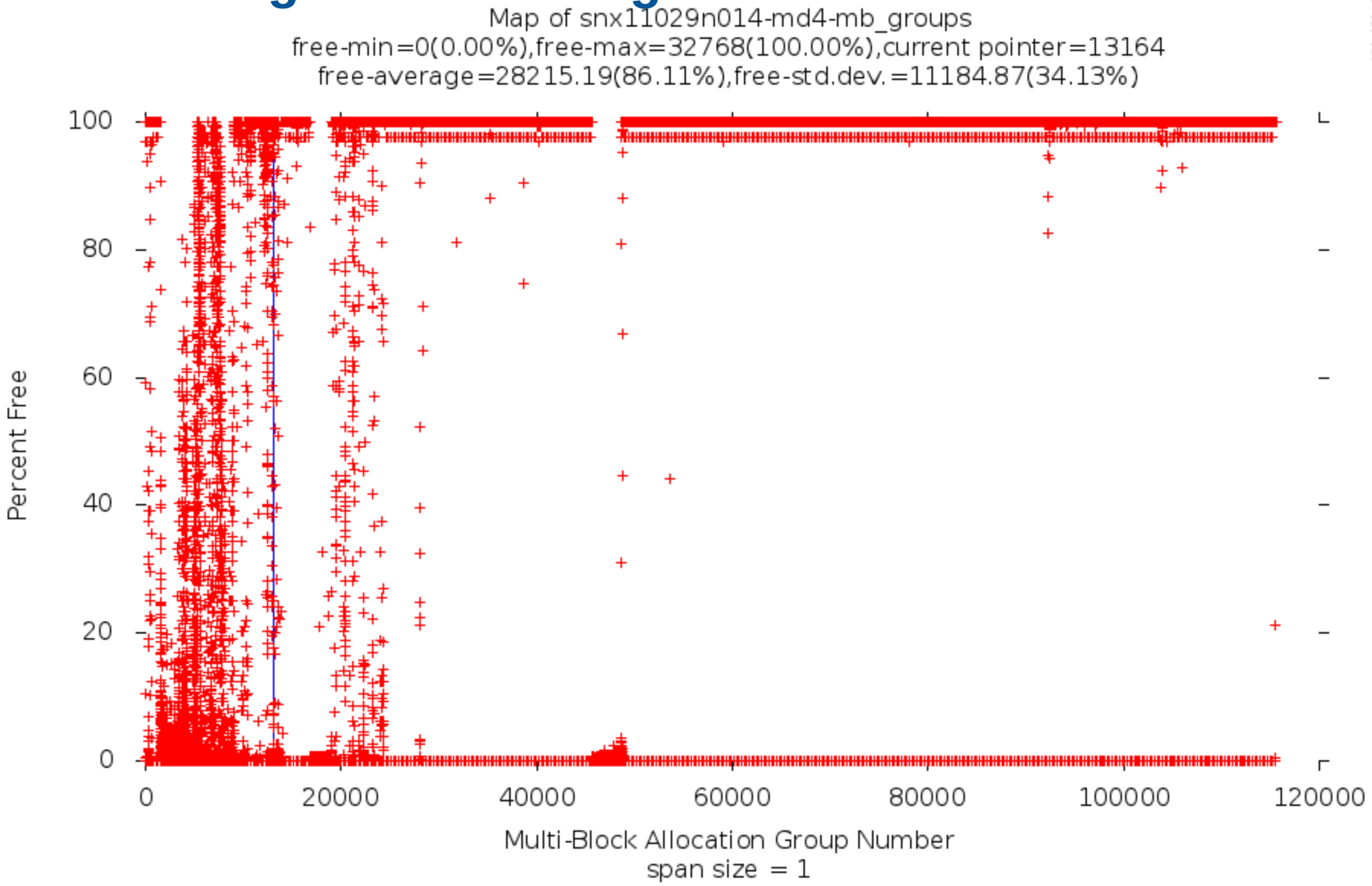
Edge to edge performance curve obdfilter-survey results single MDRAID OST, 3 TB Hitachi drives



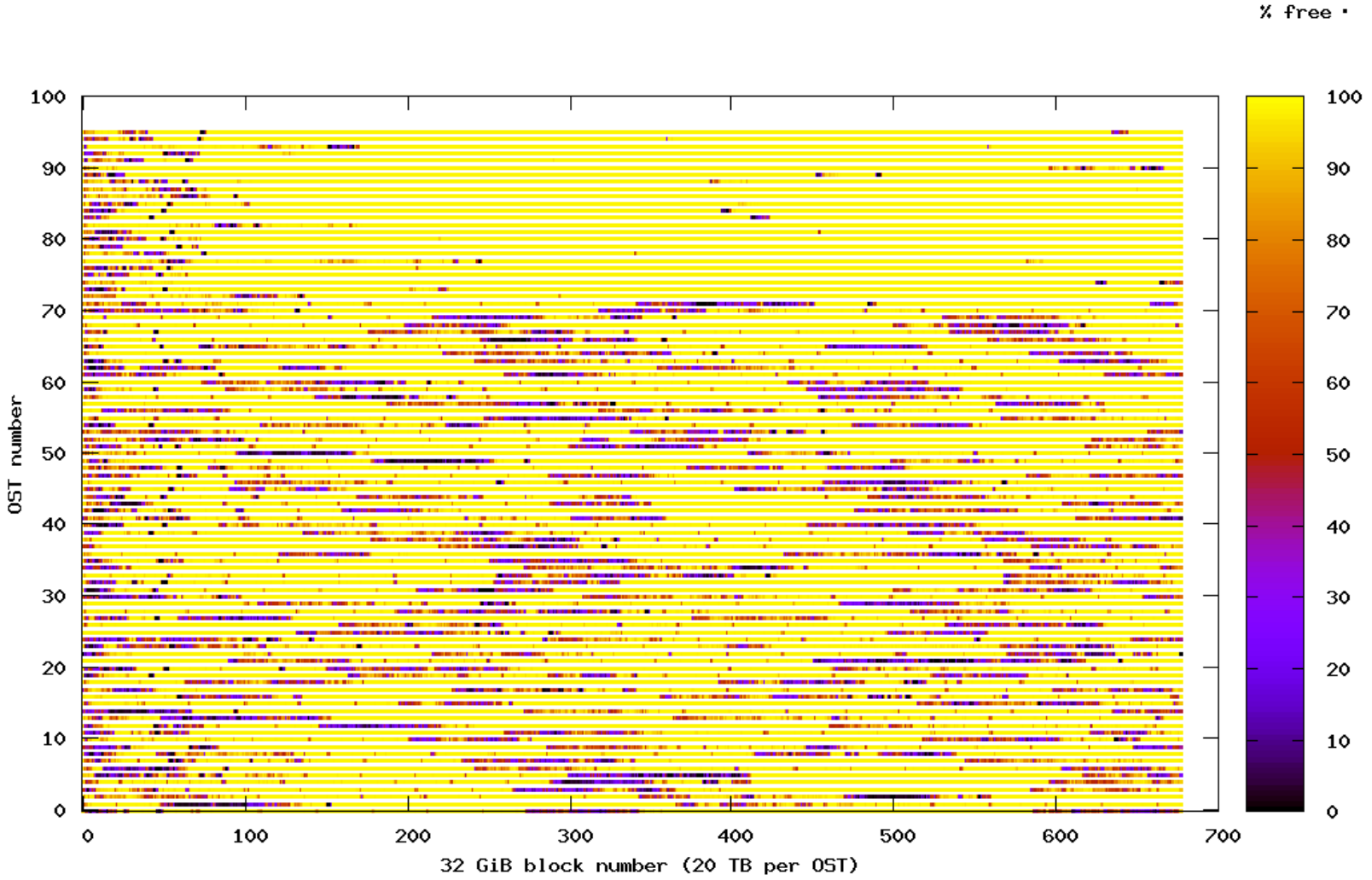
Pointers to “new data”



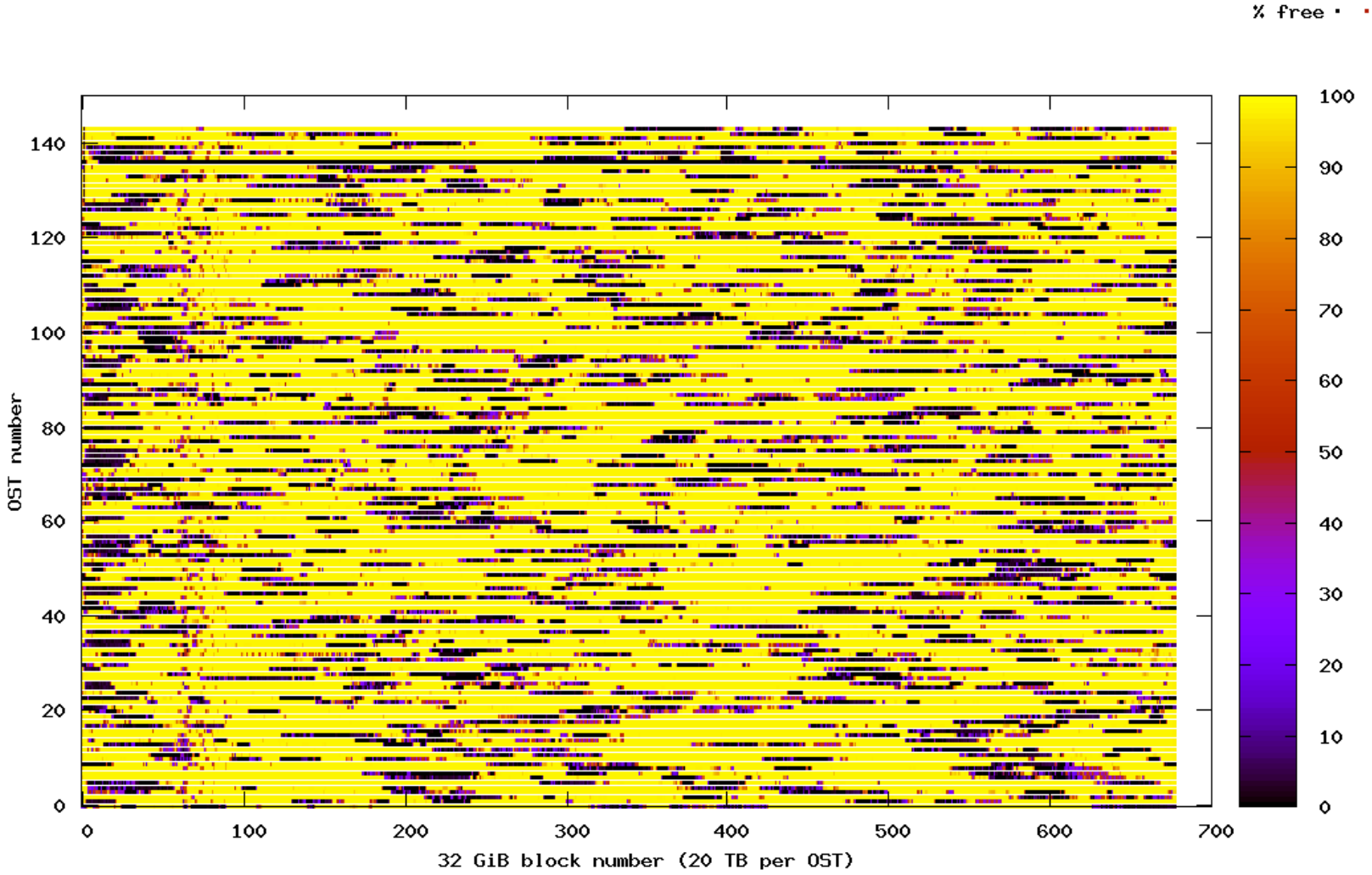
Pre-existing data on a single OST



Pre-existing data on an entire file system (1)



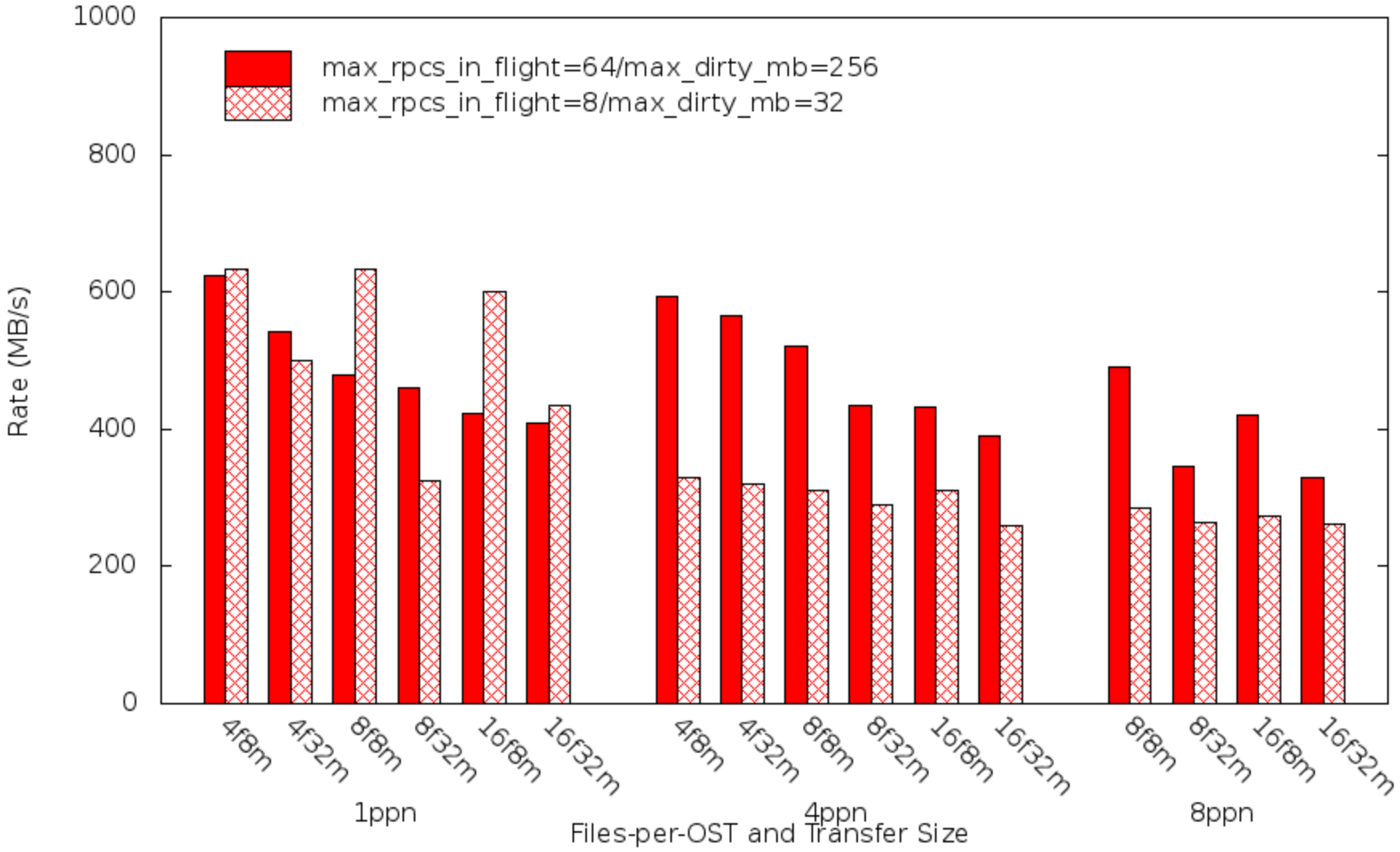
Pre-existing data on an entire file system (2)



How data gets to the OST

Client-side tuning

Effects of client-side tuning single MDRAID OST, IOR direct read

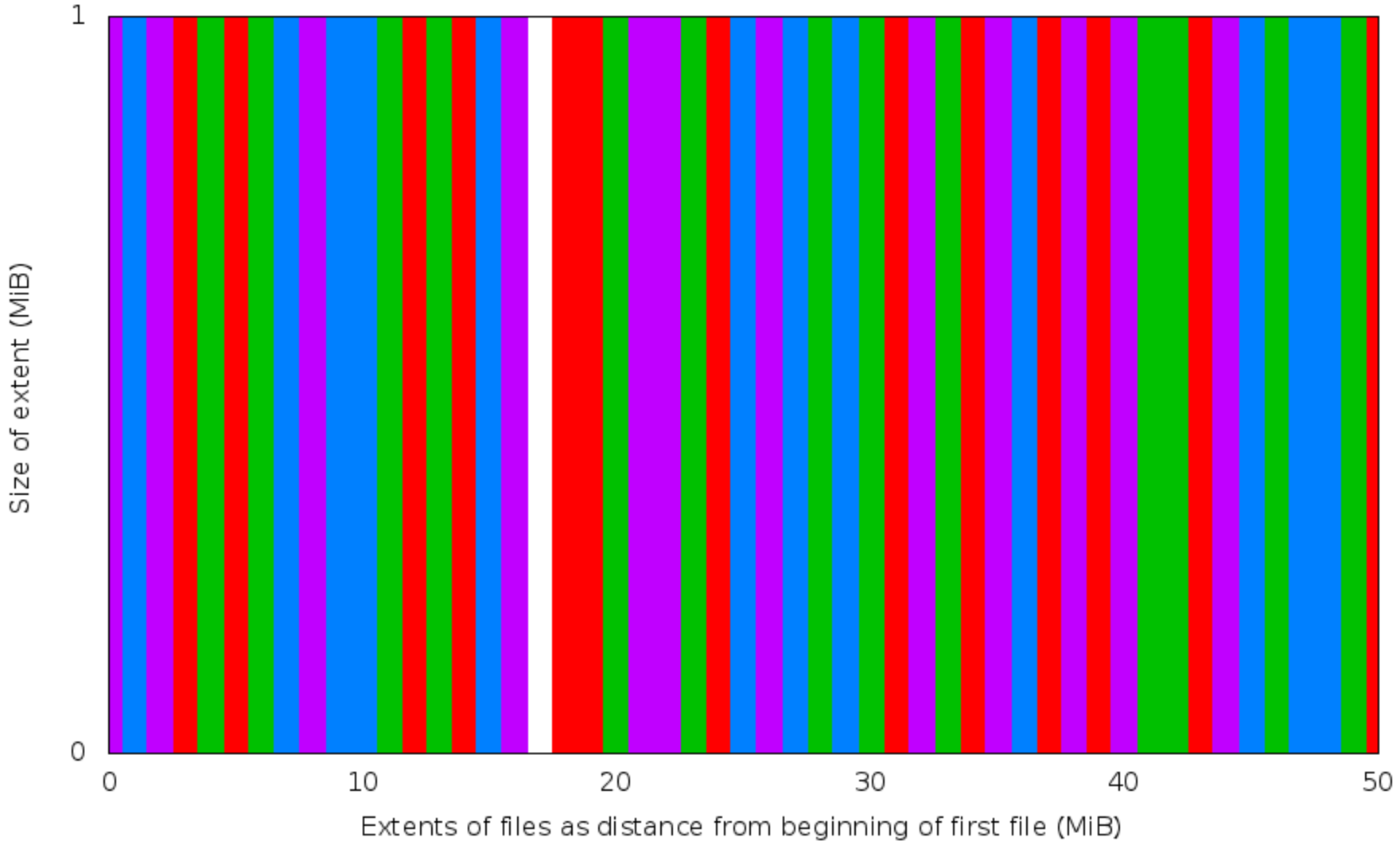


How data is arranged on the OST

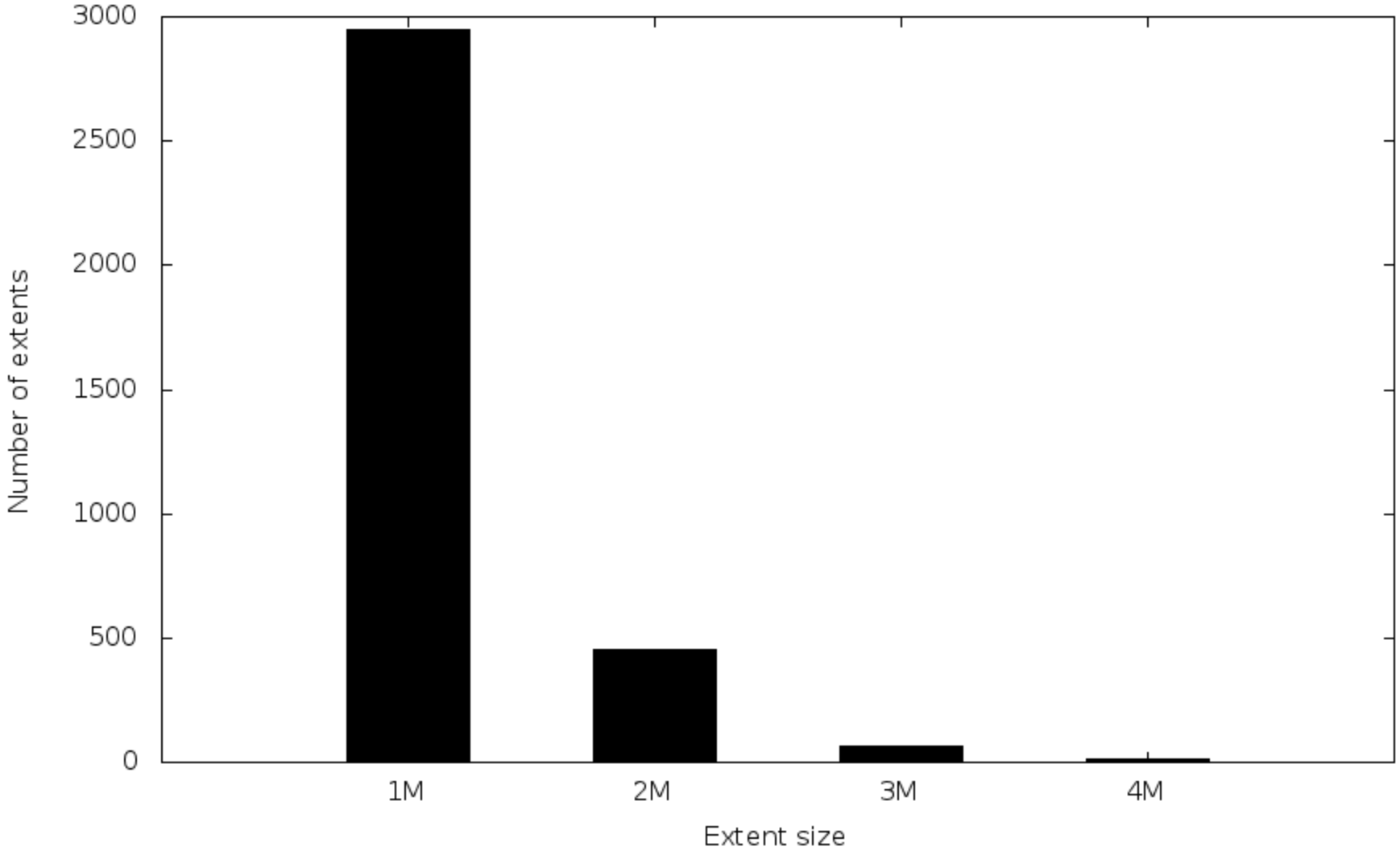
Interleaved data from multiple files

OST pre-allocation effects

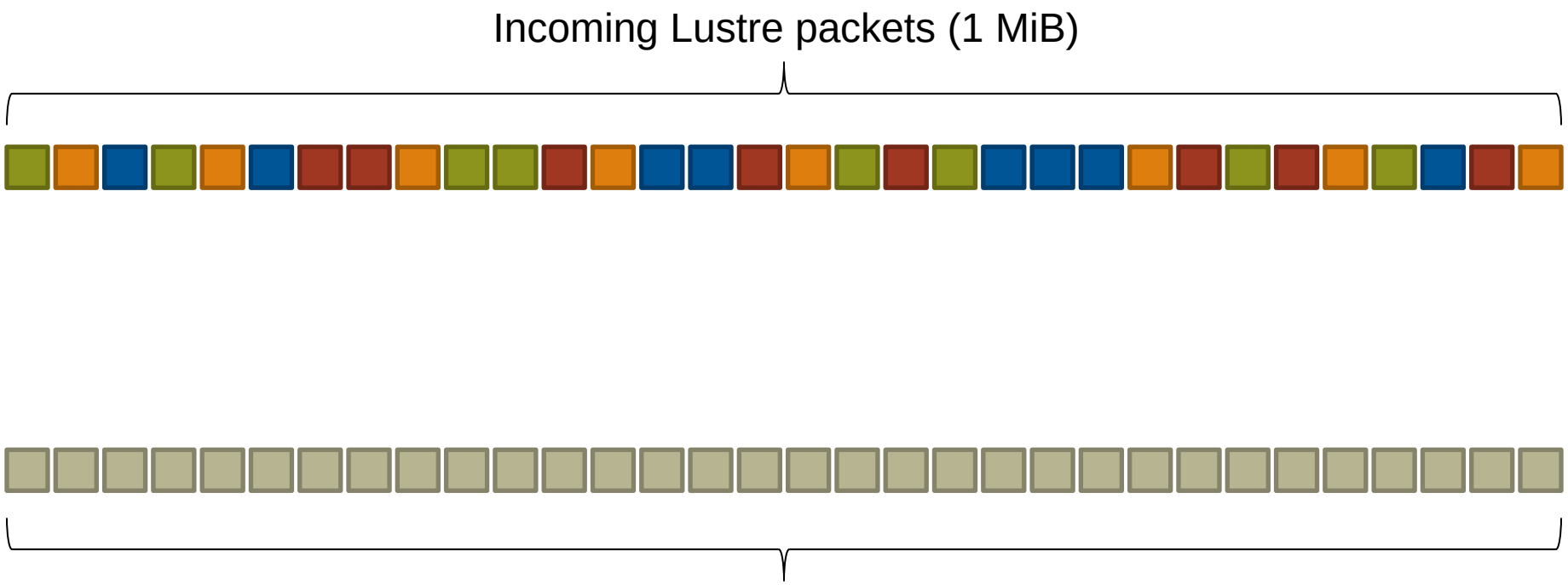
Interleaved data from multiple files (subset) direct I/O, 4 files per OST, 1 GiB per file



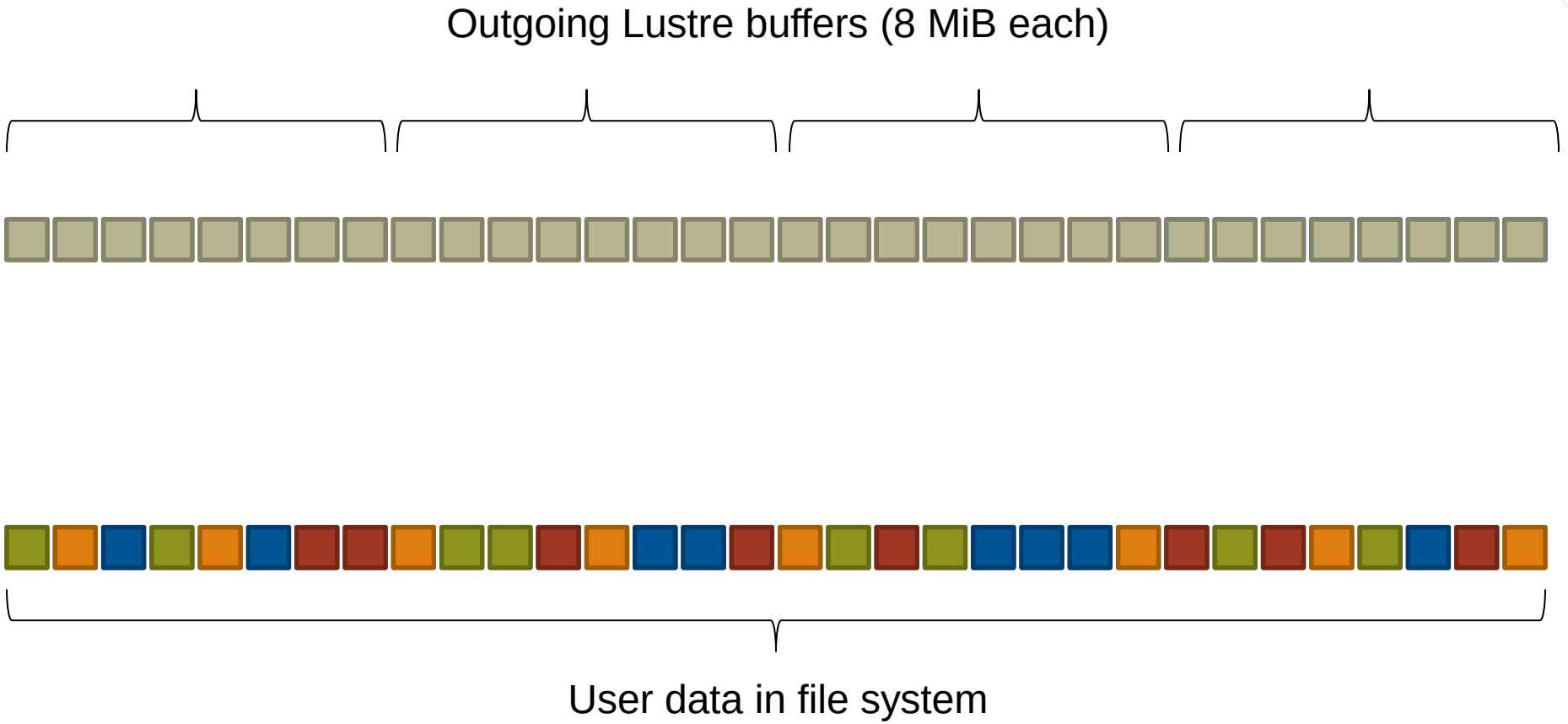
File fragment distribution
buffered I/O, 4 files per OST, 1 GiB per file



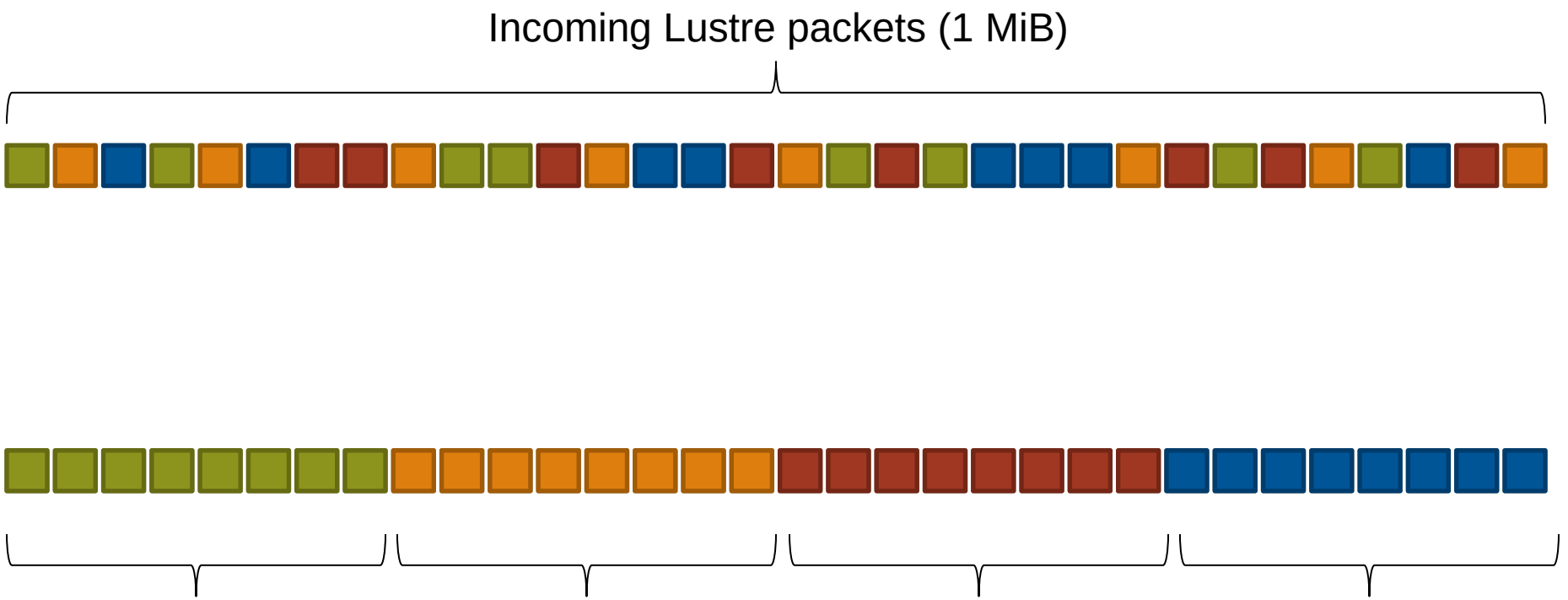
1M OST pre-allocation
4 files, 8M per file, write



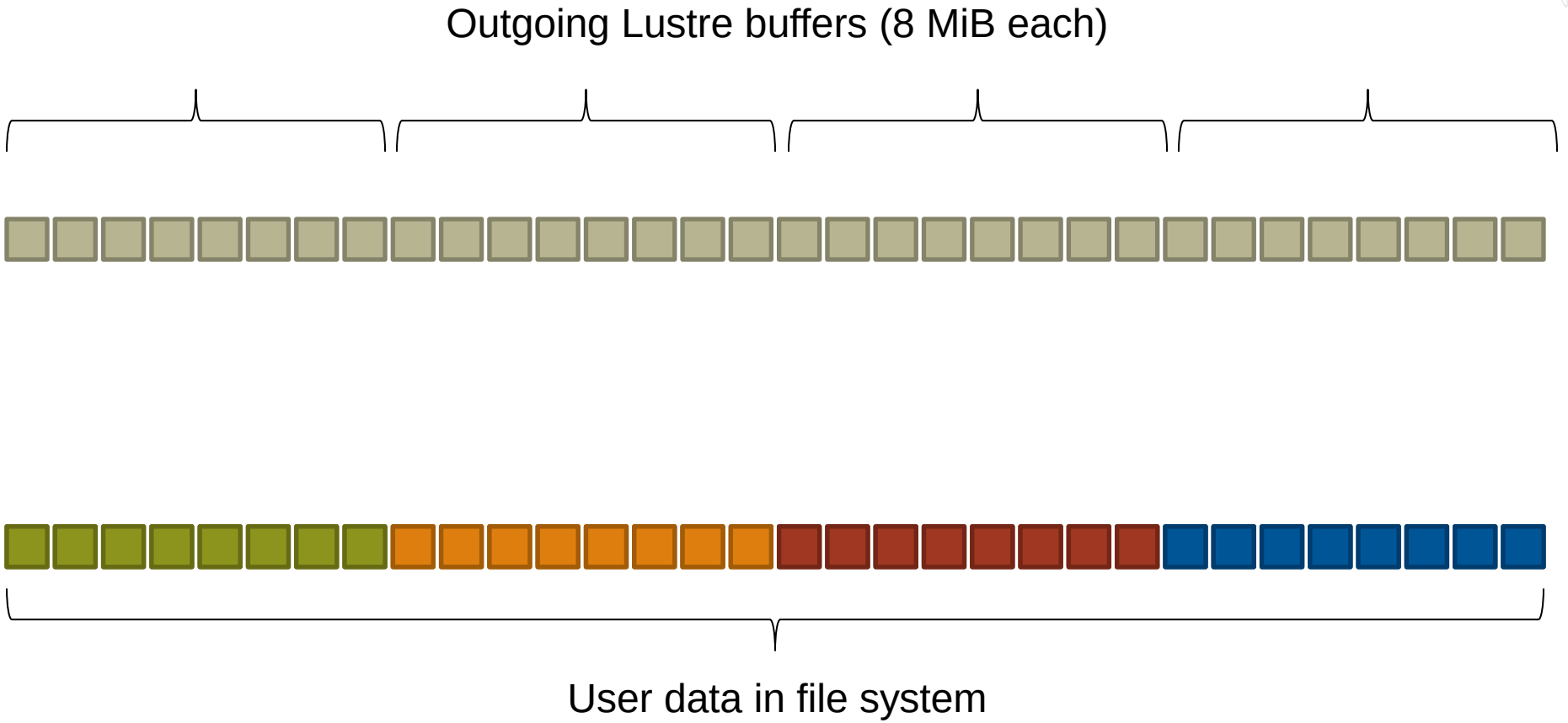
1M OST pre-allocation
4 files, 8M per file, read



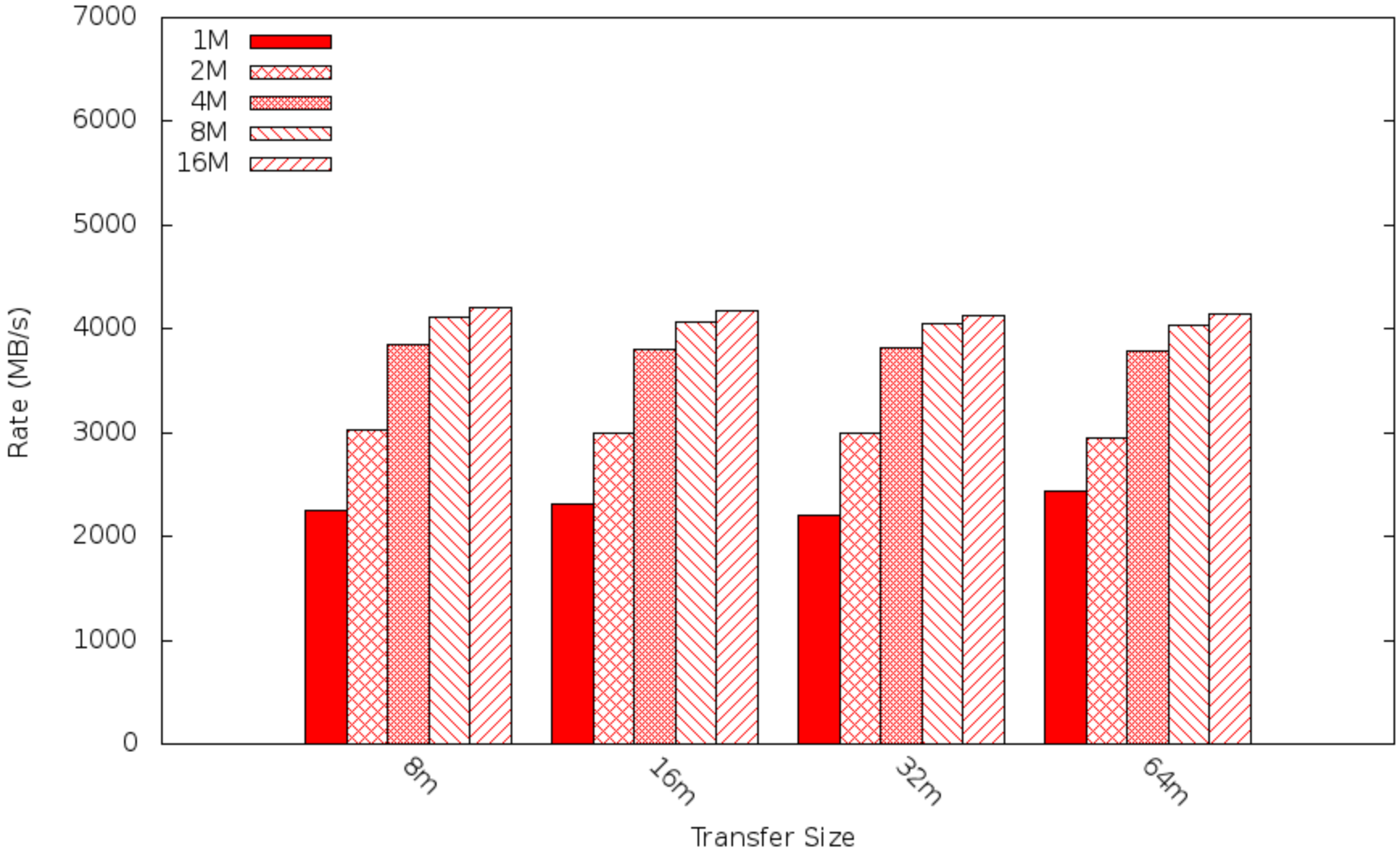
8M OST pre-allocation
4 files, 8M per file, write



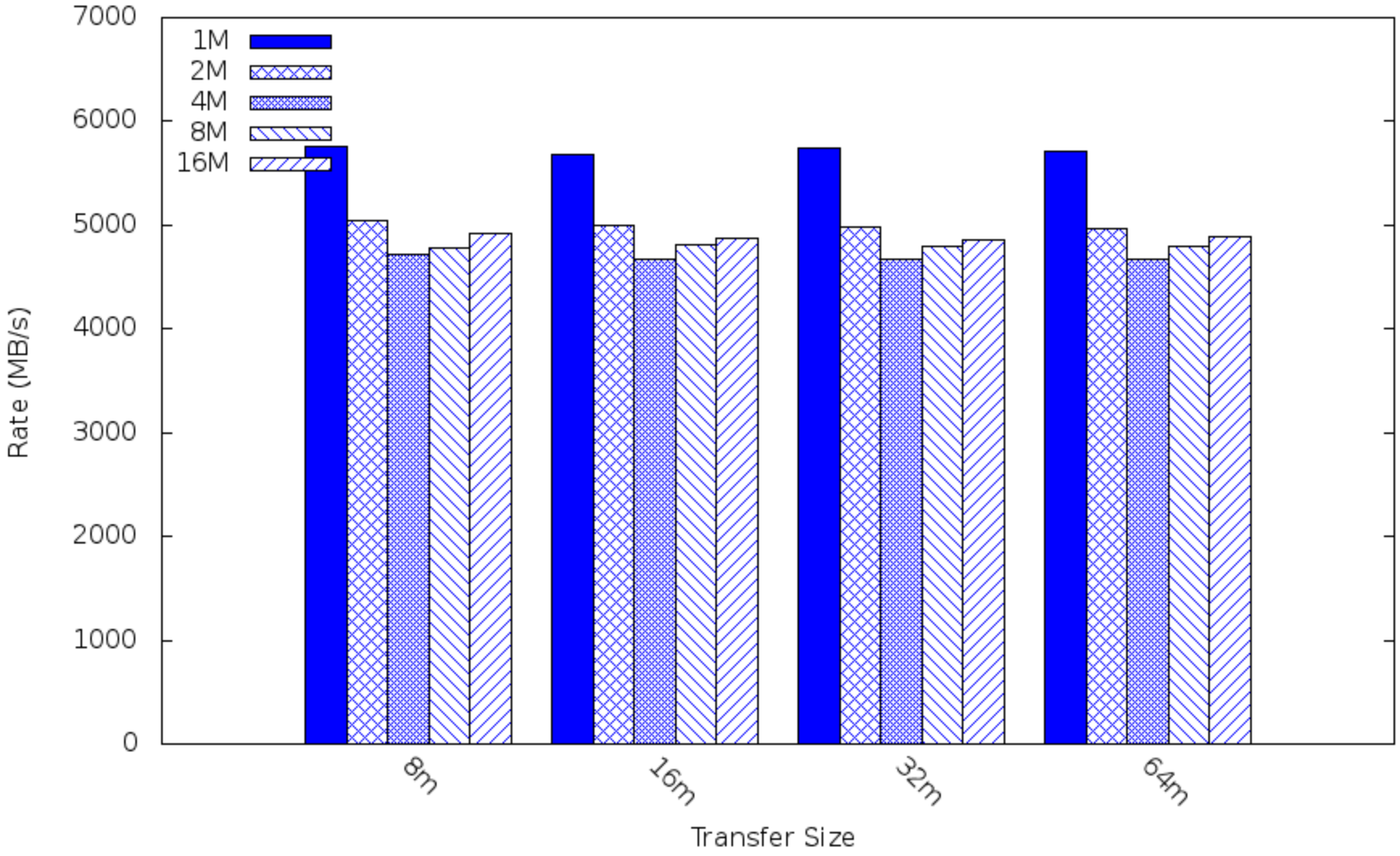
8M OST pre-allocation
4 files, 8M per file, read



Effects of OST pre-allocation on reads
single MDRAID SSU, buffered I/O, 8 files per OST



Effects of OST pre-allocation on writes single MDRAID SSU, buffered I/O, 8 files per OST



Summary

Understand where data is on the OSTs

- Every spinning disk has a fast edge and a slow edge
- The OST "new data" pointer moves across the disk
- Every OST's "new data" pointer moves independently

Understand client tuning

- Number of outstanding requests
- I/O transfer sizes

Understand how data exists on the OSTs

- Fragmentation is going to happen
- No good tools to pack data
- Methods to create more contiguous data

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.