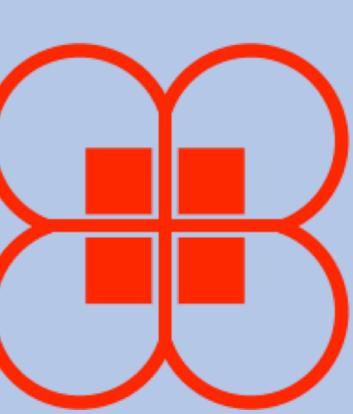


Cross-Geography Scientific Data Transferring Trends and Behavior



Summary

- Wide area data transfers play an important role in science applications but rely on expensive infrastructure that often delivers disappointing performance in practice.
- We present a systematic examination of a large set of data transfer log data to characterize transfer characteristics, including the nature of the datasets transferred, throughput achieved, user behavior, and resource usage.
- Our analysis yields new insights that can help design better data transfer tools, optimize networking and edge resources used for transfers, and improve the performance and experience for end users.
- Our analysis shows that (i) most of the datasets as well as individual files transferred are very small; (ii) data corruption is not negligible for large data transfers; and (iii) the data transfer nodes utilization is low.

1. Background, motivation, and data

By using Globus GridFTP, about 20 billion files, totaling 1.8 Exabyte between any two of 63,166 unique endpoints were transferred from 2014 to 2017. On average more than 25,000 files are transferred per minute in 2017.

We believe our findings can help:

- Resource providers to optimize the resources used for data transferring;
- End users to organize datasets to maximize performance;
- Researchers and tool developers to build new (or optimizing the existing) data transfer protocols and tools; Funding agencies to plan investments.

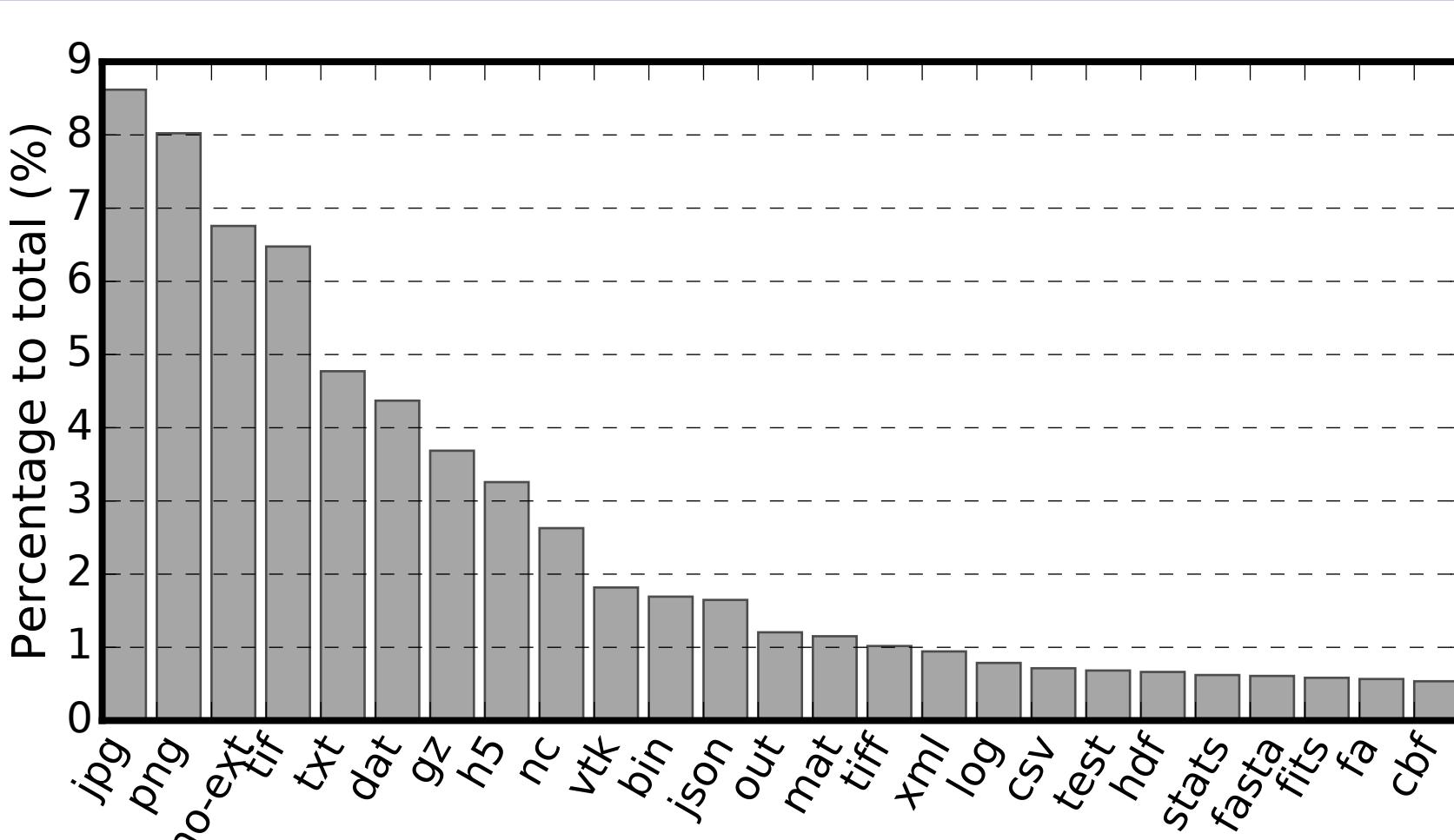
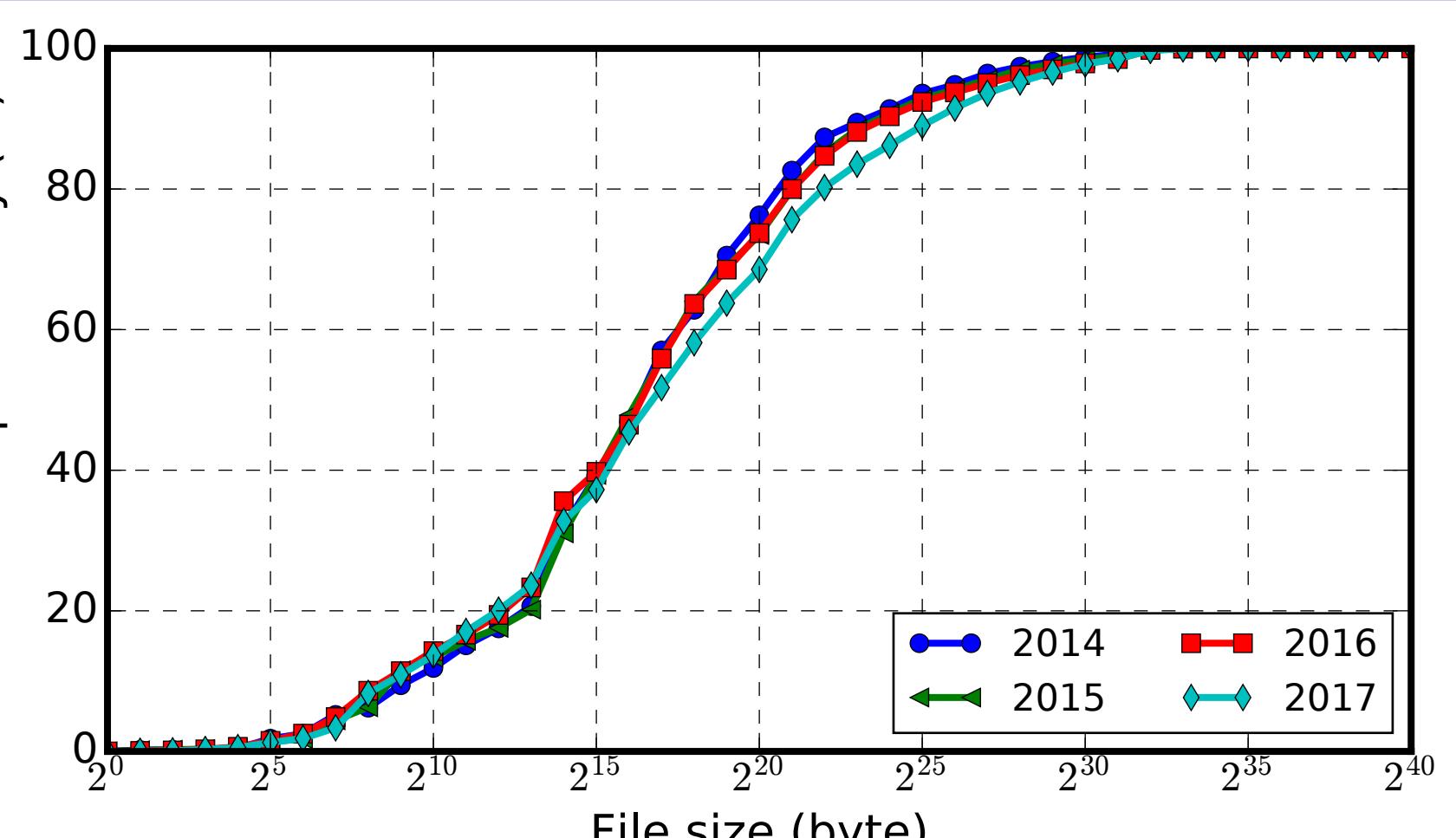
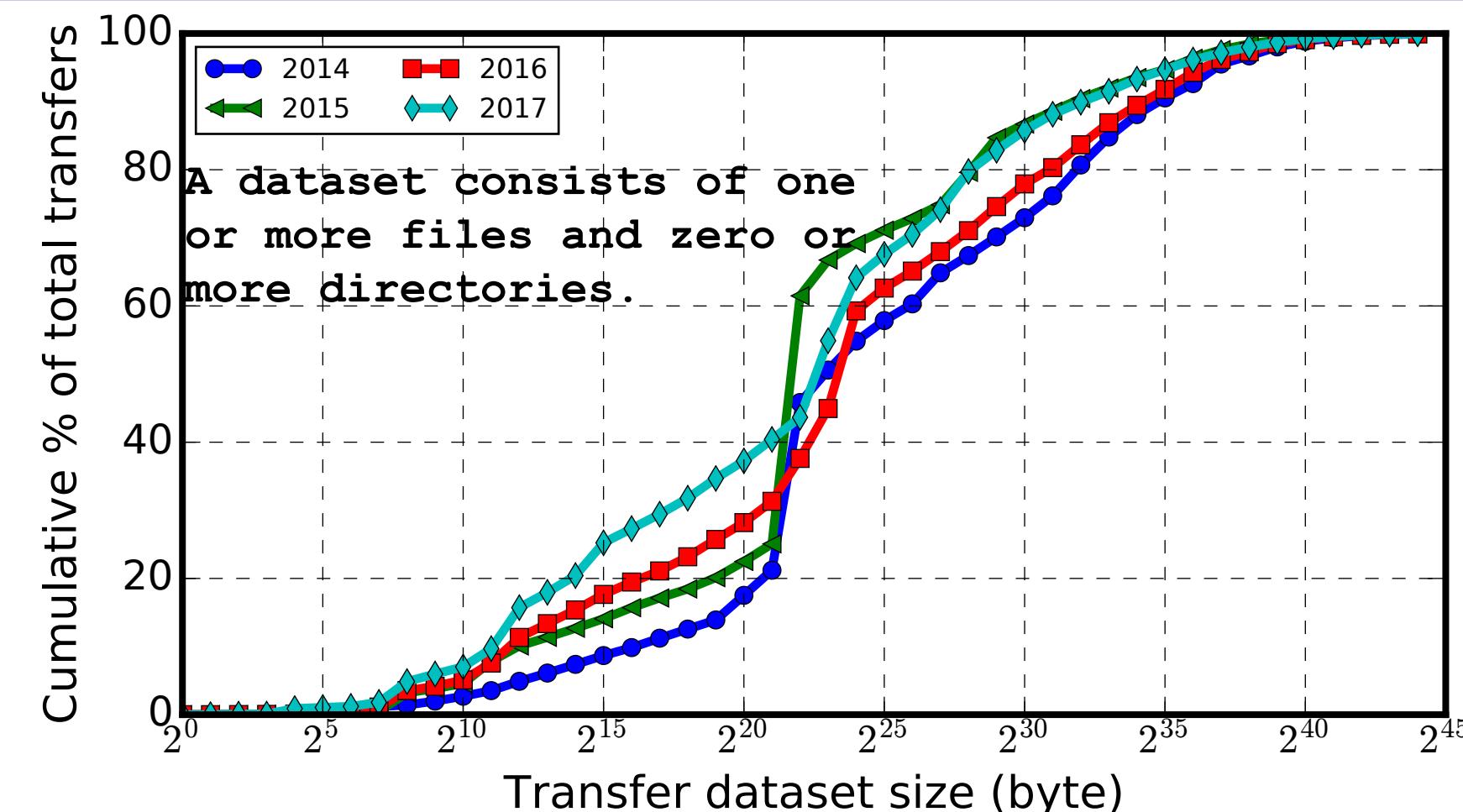
Petabytes and millions of files transferred via GridFTP using different clients.

| Year | fts_url_copy | | libglobus_ftp_client | | globusonline-fxp | | globus-url-copy | | gfal2-util | | Total | |
|-------|--------------|--------|----------------------|----------|------------------|-----------|-----------------|----------|------------|--------|----------|-----------|
| | PBytes | MFiles | PBytes | MFiles | PBytes | MFiles | PBytes | MFiles | PBytes | MFiles | PBytes | MFiles |
| 2014 | N/A | N/A | 111.23 | 746.59 | 39.81 | 1646.10 | 13.13 | 816.67 | N/A | N/A | 176.24 | 3431.78 |
| 2015 | 48.09 | 77.29 | 103.21 | 841.96 | 52.89 | 2424.58 | 19.27 | 947.78 | 0.93 | 6.70 | 267.33 | 4435.13 |
| 2016 | 244.46 | 295.67 | 105.75 | 998.96 | 88.56 | 3600.78 | 14.76 | 850.76 | 10.03 | 74.05 | 466.91 | 5922.83 |
| 2017 | 342.12 | 550.57 | 40.11 | 885.65 | 113.45 | 3901.27 | 16.89 | 898.14 | 45.93 | 234.65 | 585.01 | 6671.79 |
| Total | 634.67 | 923.53 | 360.3 | 3,473.16 | 294.71 | 11,572.73 | 64.05 | 3,513.35 | 56.89 | 315.4 | 1,495.49 | 20,461.53 |

Data transferred by Globus (i.e., globusonline-fxp)

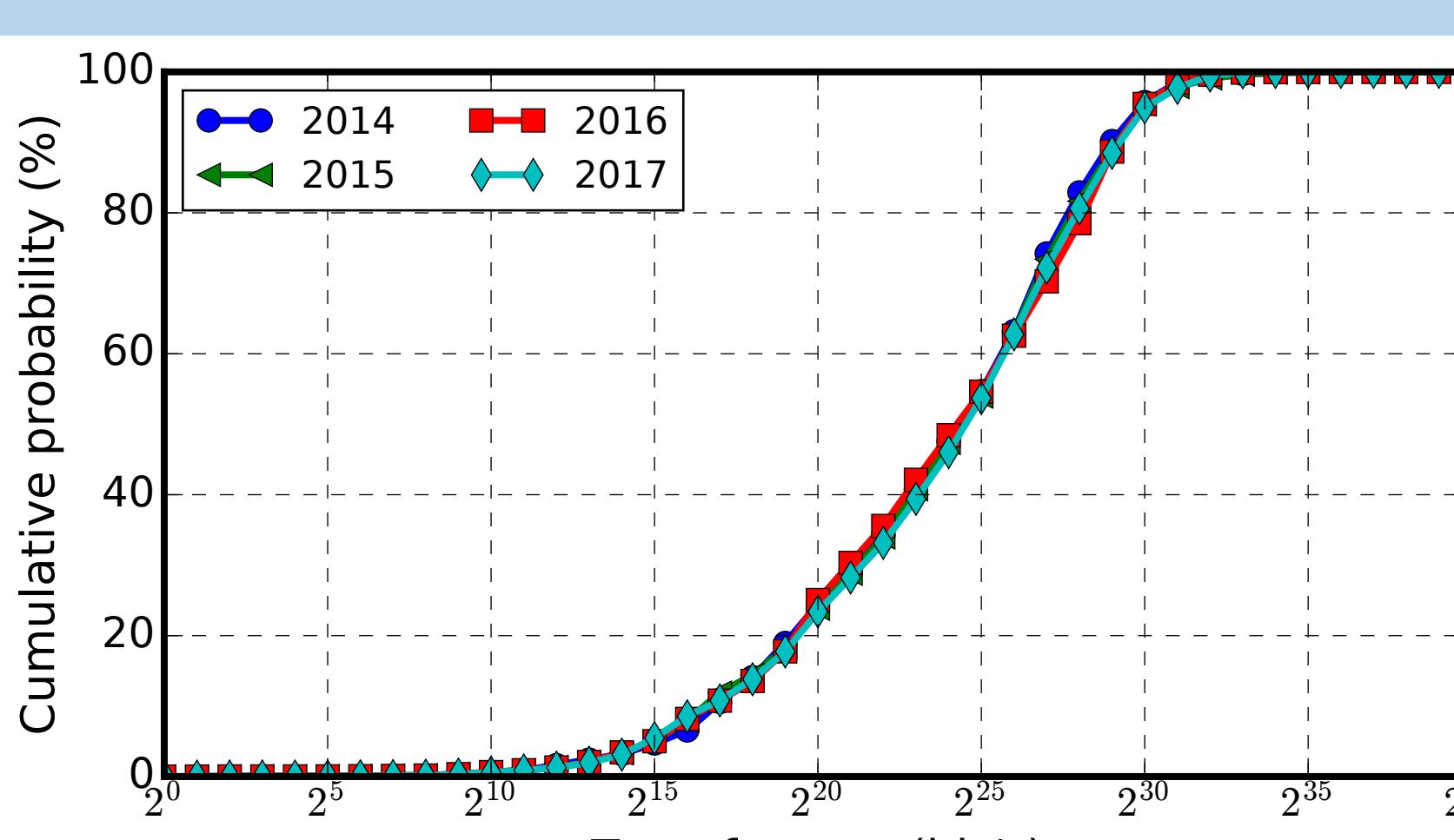
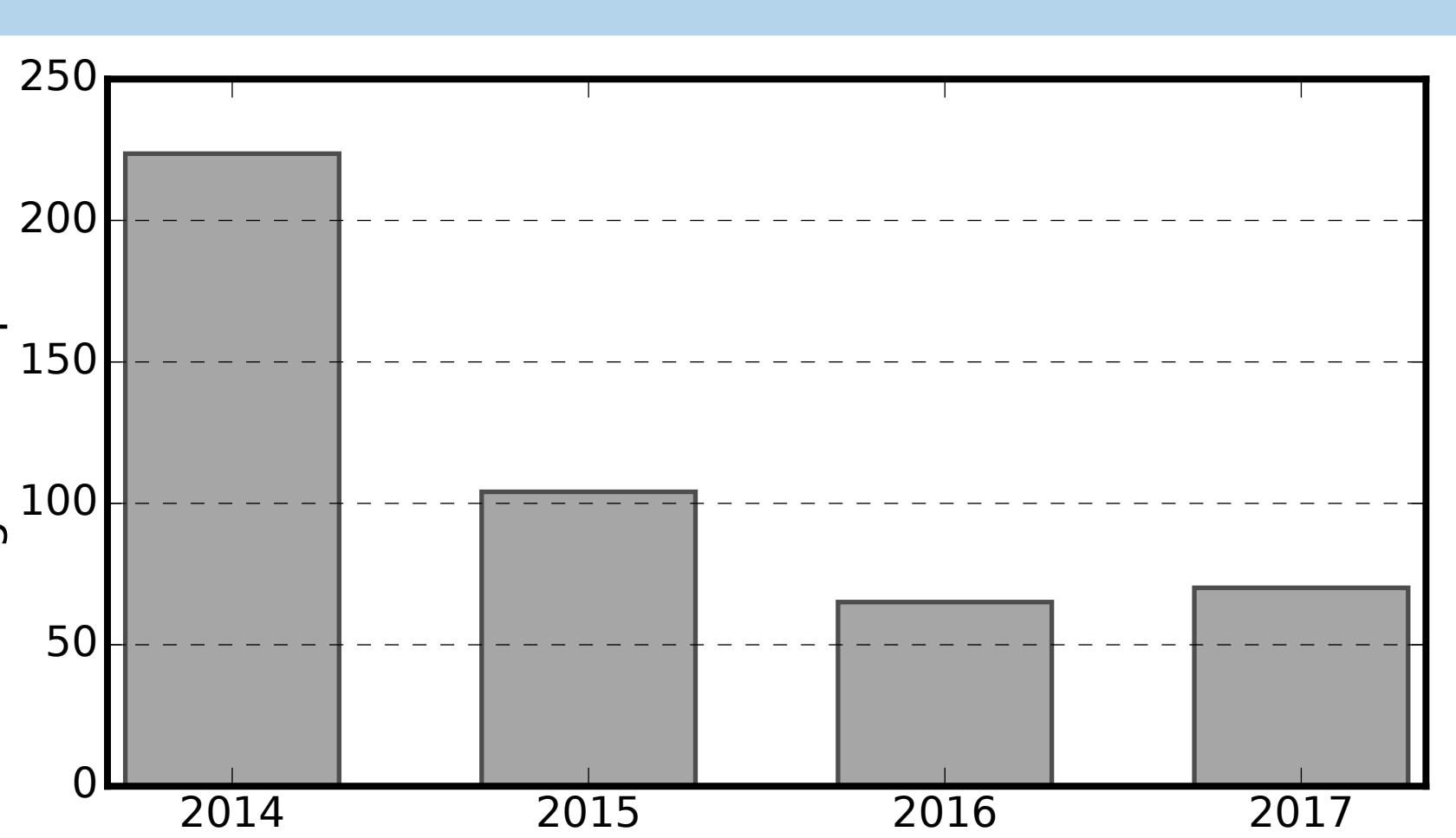
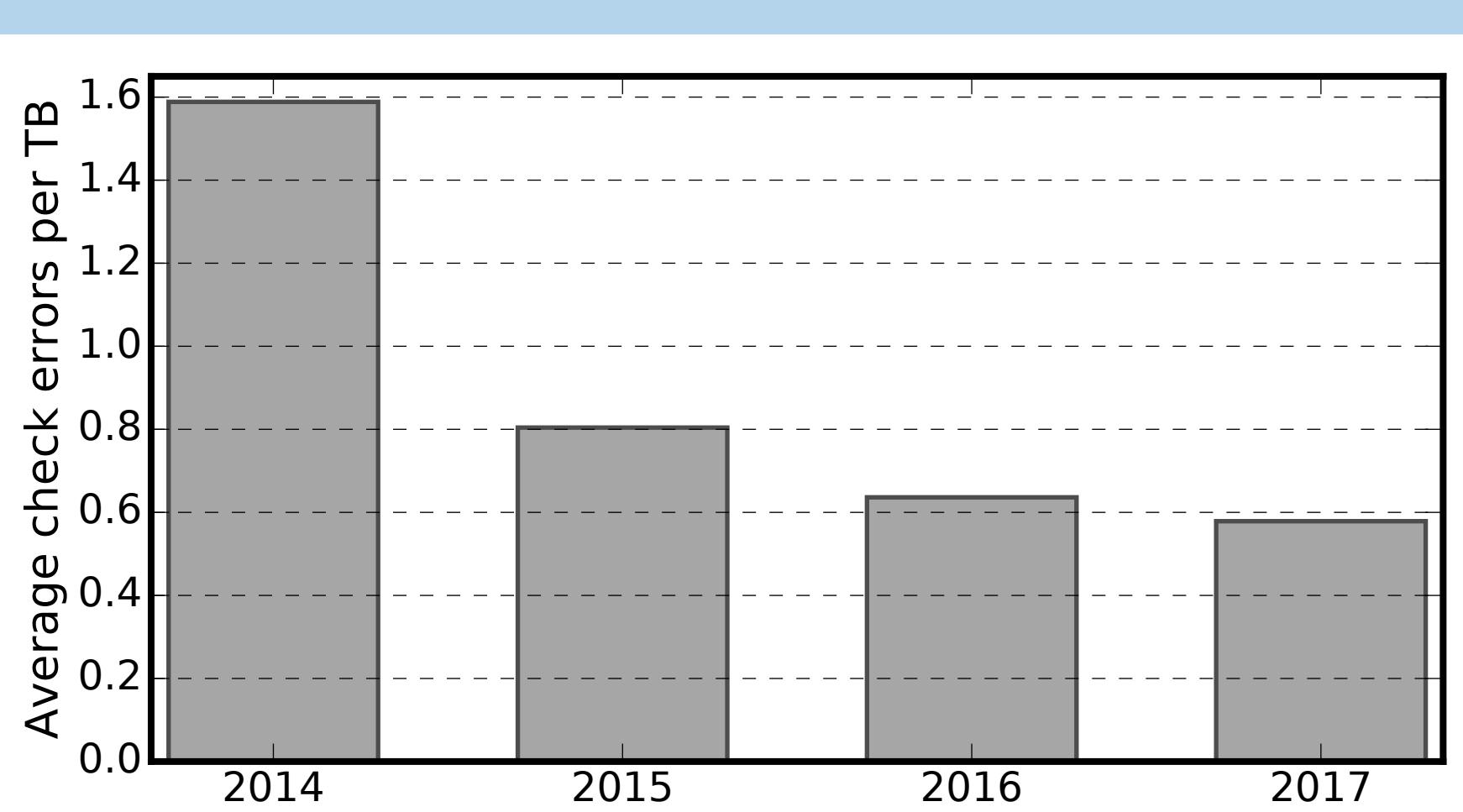
| Year | National | | International | | Total | |
|------|----------|--------|---------------|--------|--------|--------|
| | PBytes | MFiles | PBytes | MFiles | PBytes | MFiles |
| 2014 | 41.44 | 1,865 | 0.78 | 26.9 | 42.32 | 1,892 |
| 2015 | 53.45 | 2,763 | 2.55 | 94.3 | 56.39 | 2,873 |
| 2016 | 90.10 | 3,929 | 2.84 | 110.8 | 93.60 | 14,042 |
| 2017 | 109.16 | 4,162 | 3.23 | 94.3 | 113.50 | 4,264 |

2. Dataset characteristics



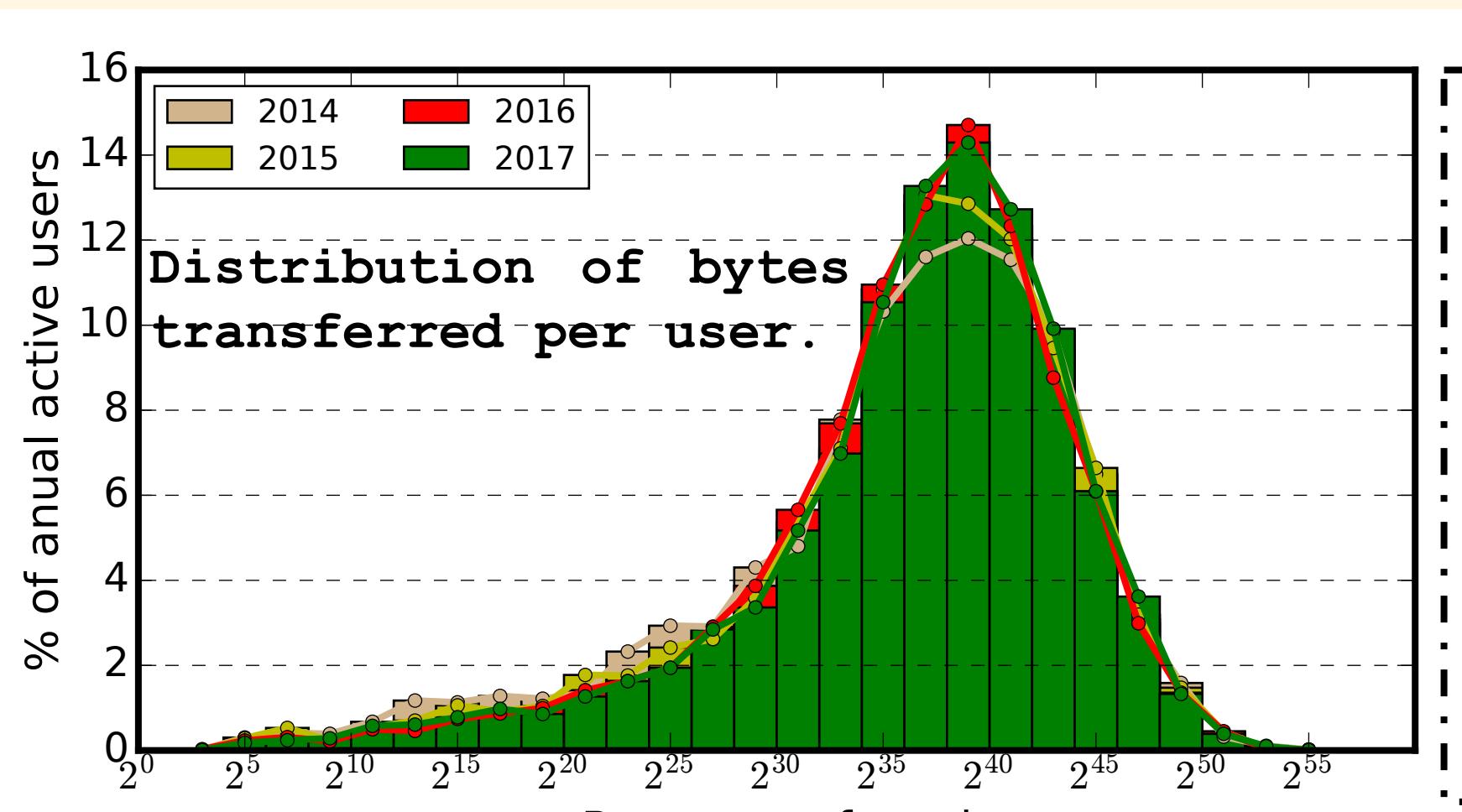
- Most of the datasets transferred by the Globus transfer service have only one file. And 17.6% of those datasets (or 11% of the total) have a file size > 100 MB, motivating the need for striping the single-file transfer over multiple servers.
- The average file size of most datasets transferred is small (on the order of few megabytes).
- Repeated transfers are not common, less than 7.7% of the datasets are transferred more than once. When they do occur, the datasets in question are distributed mostly from one (or a few) endpoints to multiple destinations.

3. Transfer characteristics

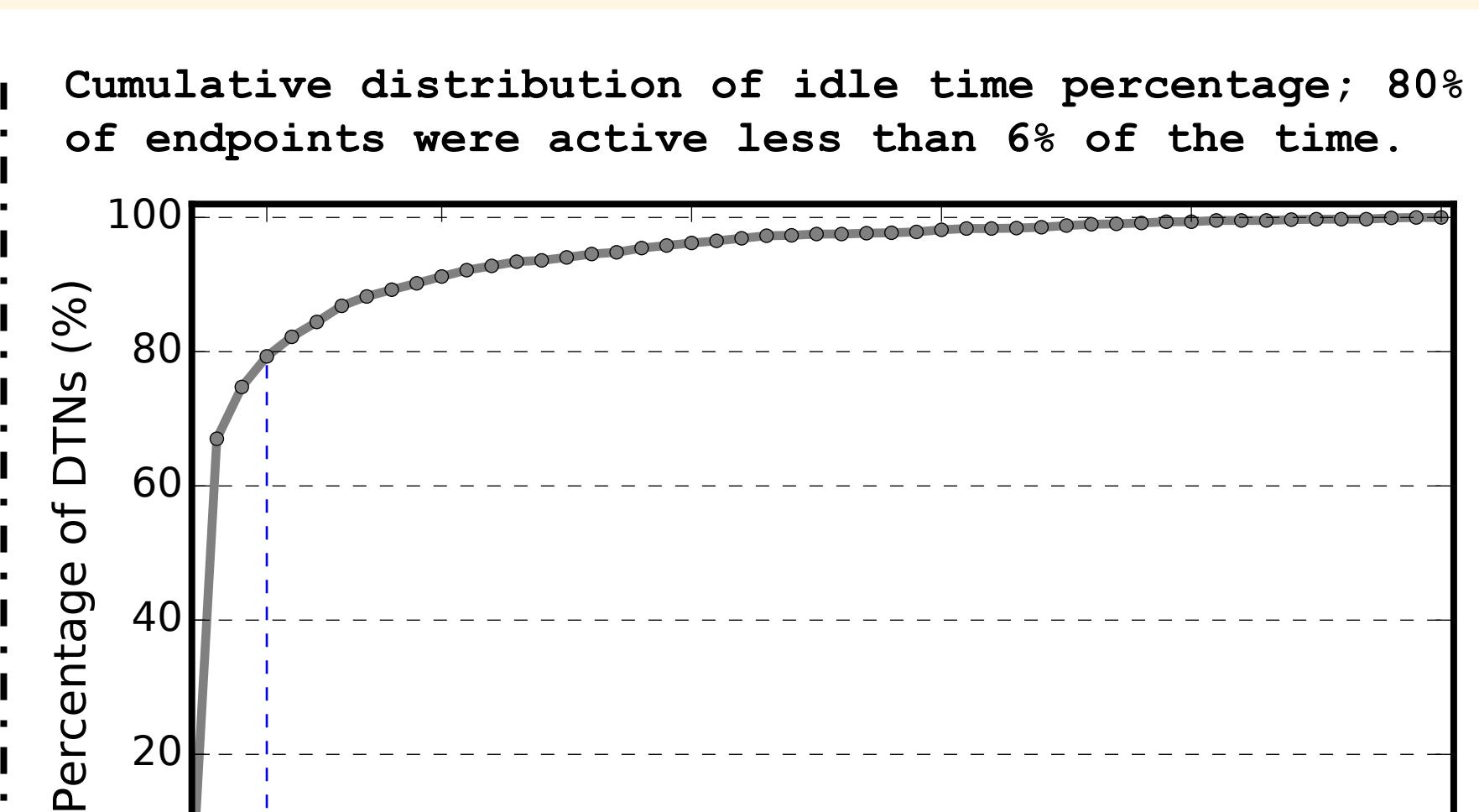


- At least one checksum failure occurs per 1.26 TB. The integrity checking is needed though it causes extra load.
- The failures are decreasing year by year. Overall, the service is becoming increasingly reliable.
- Although some server-to-server transfers achieve high performance (dozens of Gbps), most transfer throughput is low.
- There is no clear increasing trend in terms of transfer performance over time.

4. User behaviors

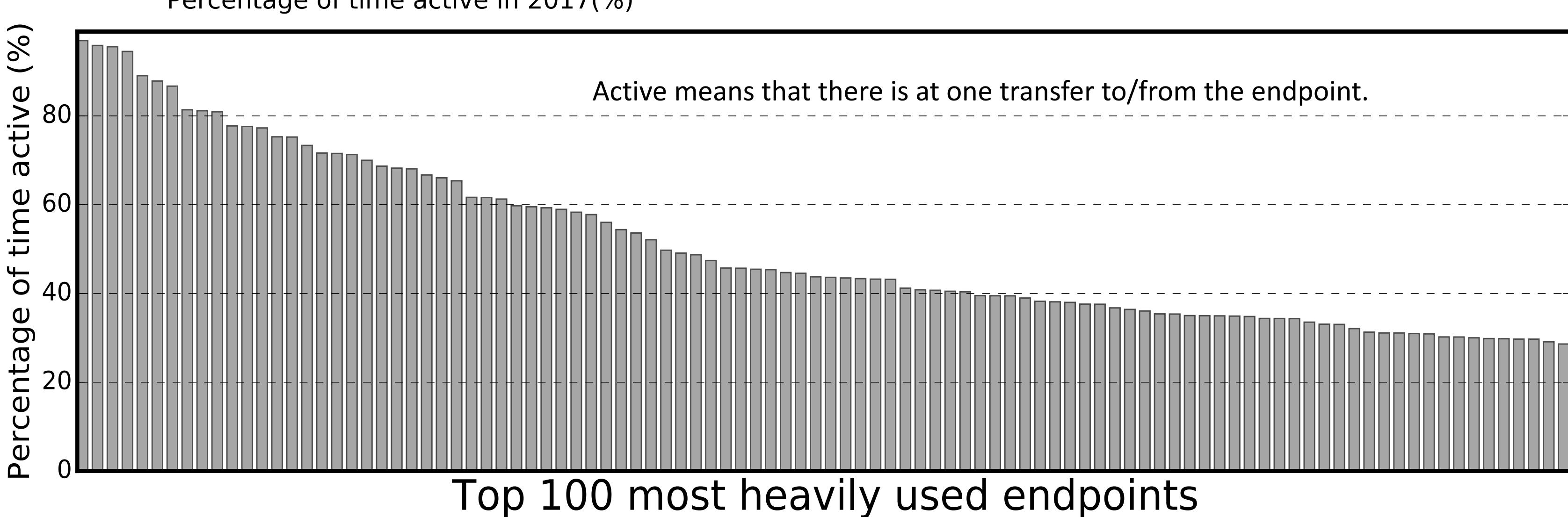
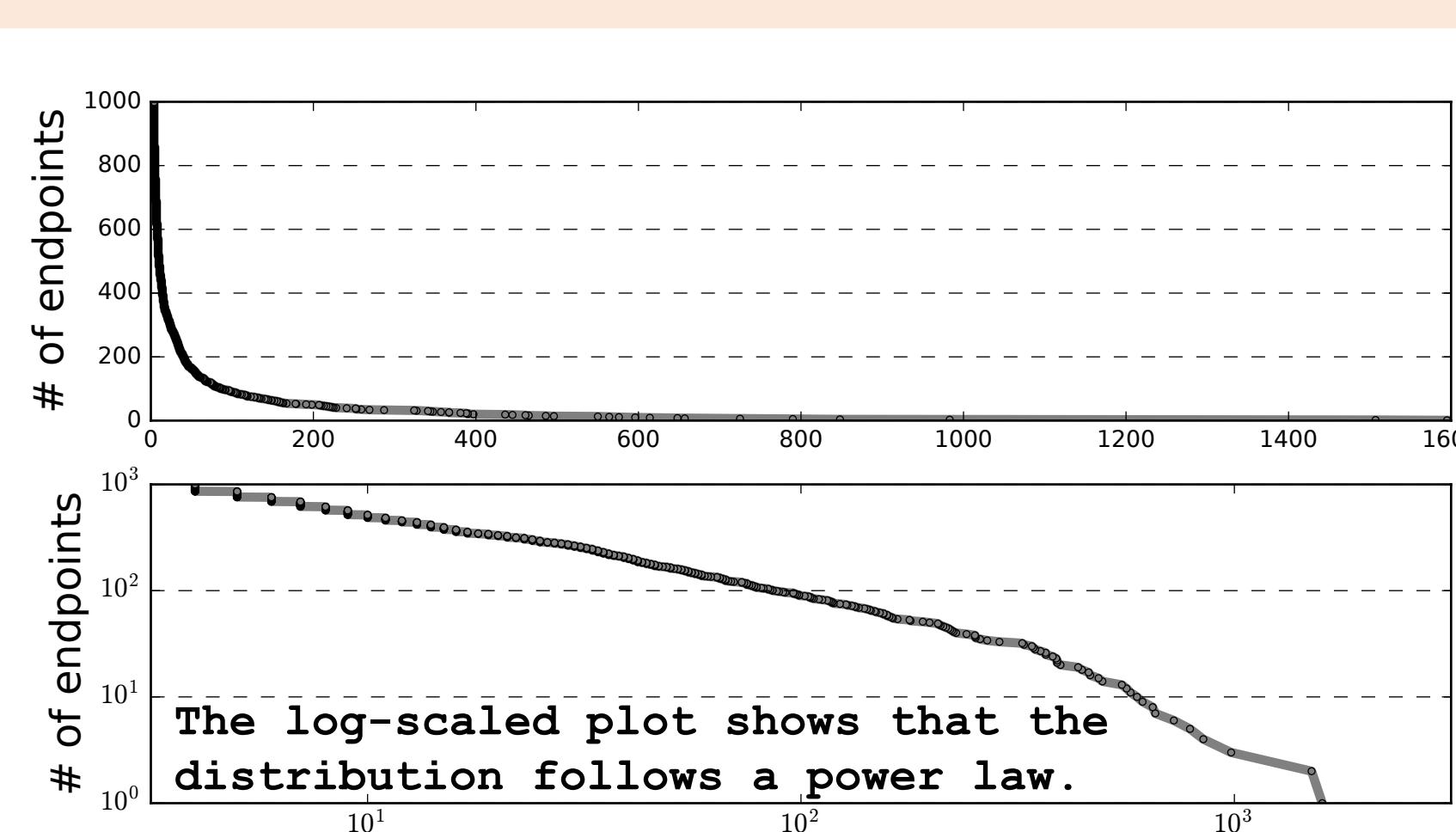
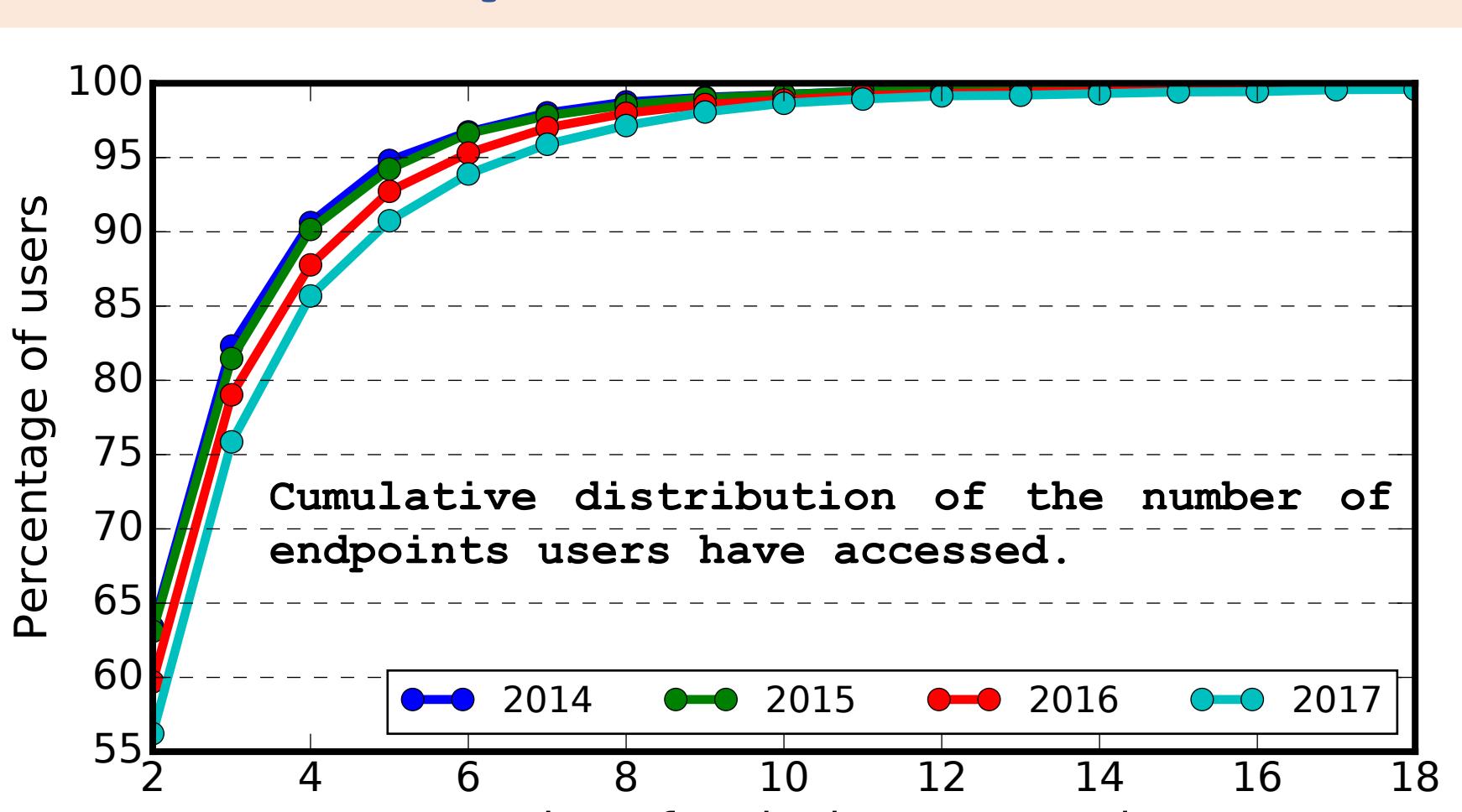


- Of all the bytes transferred, 80% are by just 3% of all users; 10% of the users transferred 95% of the data.
- The distribution of the number of users per endpoint follows a power-law distribution, similar to other real-world social network graphs.
- Most users do not manually tune the transfer parameters.
- Thus, transfer tools should be smart enough to choose the optimal parameters.



- DTN utilization is surprisingly low. Since the DTN requirement is high for high-throughput DTNs, some good topics for research would be the use of these computing resources:
 - for other purposes;
 - for complex encoding to deal with data corruption and;
 - to compress data to reduce the network bandwidth consumption.

5. Endpoint characteristics



- Slightly more than half of the users accessed two or fewer endpoints.
- The degree distribution of the number of users per endpoint follows a power-law distribution, similar to other real-world social network graphs.

Acknowledgements

This material was supported in part by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357 and the DOE RAMSES project fund by Scientific Workflow Analysis program managed by Richard Carlson.