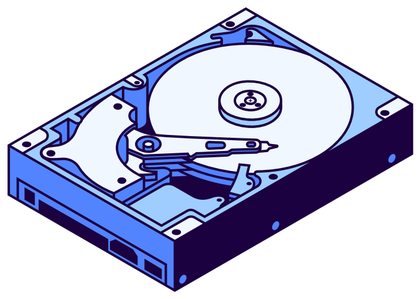
Risk Factors of Drive Failure



Introduction

With the rapid growth of cloud computing and large-scale databases, the cost of drive failure is only growing. Even to the average consumer, knowing what to look for when buying physical drives could be of use. S. Shah et al. found that the capacity dependent failure mechanisms on higher capacity hard drives significantly increased overall hazard rates, while others found no significance, citing temperature, model type and assembly quality as better predictors of failure.[1,2]

Objective

The goal of this analysis was to evaluate the risk factors (brand and capacity) associated with time to drive failure.

Results

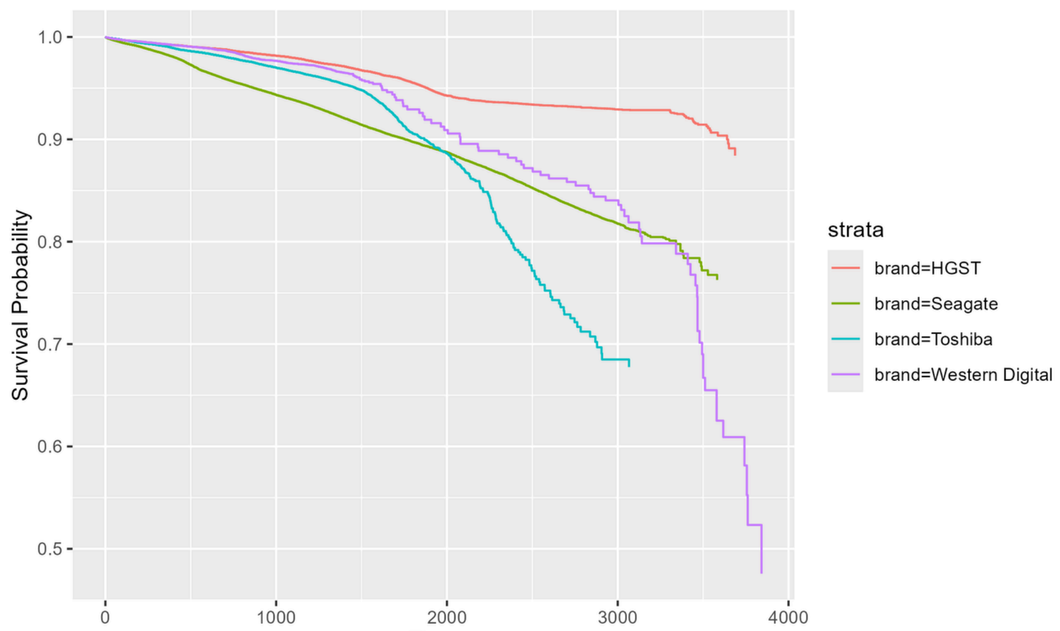
A significant difference in survival probabilities across brands and capacities ($p < 0.0001$).

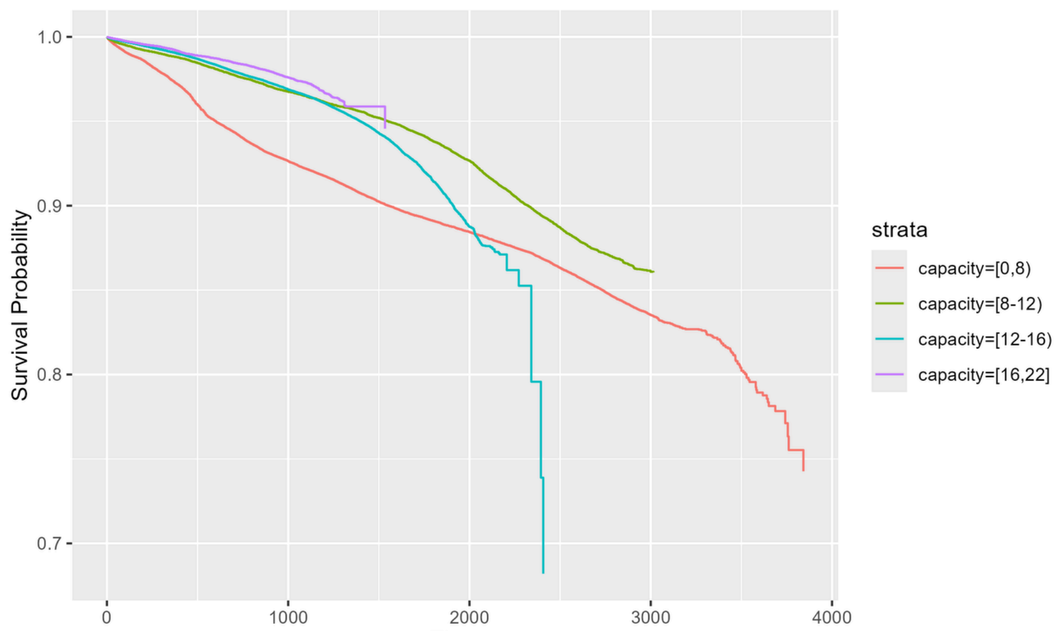
Brands

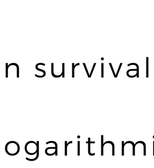
- Seagate & HGST showed gradual, linear declines in survival overtime
- At ~1500 days, Toshiba and Western Digital survival drop more rapidly
- Steep drop-off past 3400 day-mark for Western Digital

Capacity

- Gradual, linear declines in survival for all groups except 12-16Tb
- 12-16Tb strata showed a logarithmic decay in survival







Conclusions

- Brand type and capacity may be associated with survival probability
- Seagate & HGST KM curves were generally stable, while that of Western Digital & Toshiba were more logarithmic in their decline. The differences in stability is likely partially attributed to the differences in sample size. Lower sample sizes lead to higher volatility.
- Western Digital & Toshiba may be increasingly more susceptible to failure past the 3400 & 2000 day-marks respectively compared to competing brands, however further testing is required.
- The steep drop-off in survival for 12-16Tb drives past 2000 days may support the notion of higher drives having higher failure rates, however interpretation is limited due to unadjusted estimates.

Materials & Methods

Data source:
Backblaze, an open cloud storage company[3]

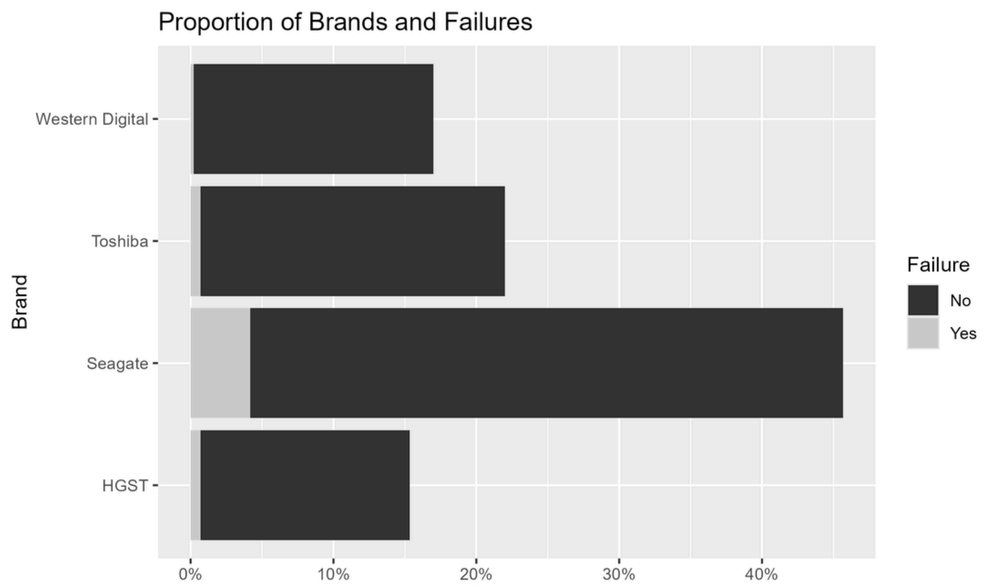
Sample:
Daily logs of 442110 total drives from 2013-2023 Q3[4]

Parameters of Interest:
- Brand
- Capacity (Tb)

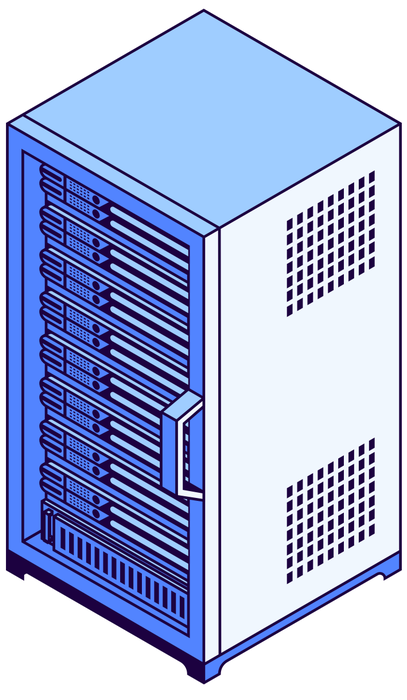
Statistical Techniques:
- KM Estimates
- Wilcoxon & Log-rank tests

Limitations

- Omitted brands with low sample size & high censoring
- Grouped capacities & models by brand for ease of interpretability at the cost of higher variability
- Parameters compared in isolation
- Limited interpretability due to violated AFT and Cox-PH assumptions (transformations to time-varying coefficients were found to be ineffective or computationally infeasible)



Group	$\chi^2_{3,\ Wilcoxon}$	$\chi^2_{3,\ Log\text{–}rank}$	p-value
Brand	3174	3201	p<0.0001
Capacity	1539	1641	



References

1. S. Shah and J. G. Elerath, "Reliability analysis of disk drive failure mechanisms," Annual Reliability and Maintainability Symposium, 2005. Proceedings., Alexandria, VA, USA, 2005, pp. 226-231, doi: 10.1109/RAMS.2005.1408366.
2. S. Sankar, M. Shaw and K. Vaid, "Impact of temperature on hard disk drive reliability in large datacenters," 2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN), Hong Kong, China, 2011, pp. 530-537, doi: 10.1109/DSN.2011.5958265.
3. https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data
4. The daily log csv's were compiled and uploaded onto a SQL database using python, and then queried such that each row contained a distinct drive with their corresponding capacity, model, observed time (i.e. number of days between first and last log), and indicator for failure. The rest of the data cleaning was performed in R. 2 entries were omitted due to an ambiguous model type, likely a typo. Capacities were grouped into 4 categories & a brand covariate was added in place of model for ease of interpretability.