
# Analysis and Modeling of Storage Access Patterns and Caching Strategies

*Shreya Krishnan*

**Supervisors: Markus Schulz and Andrea Sciaba**

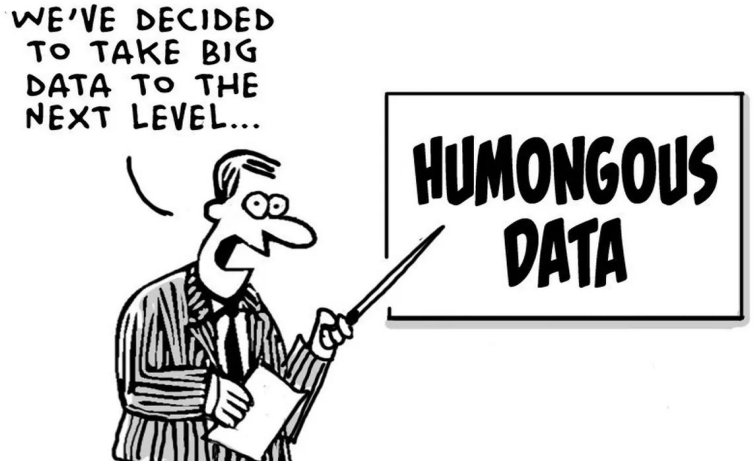


Lightning Talk - 13th August 2019

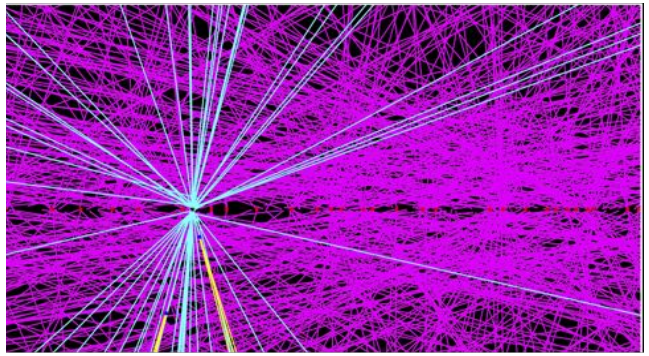
# Challenges of the Future: HL-LHC

## (~~More~~ data x higher complexity) + Flat Budget

Much, much, much more...




✖



=

## Reduce cost and complexity of storage!

# Replace storage by caches @ Tier 2s

*We need to Analyse & Model Storage Access Patterns & Caching Strategies*

Used CMS Logs

## Analysing Caching Strategies

- ★ Cache Size
- ★ File Eviction Strategy
- ★ Trade-off between Network Usage and Cache Size

## Modeling file access patterns

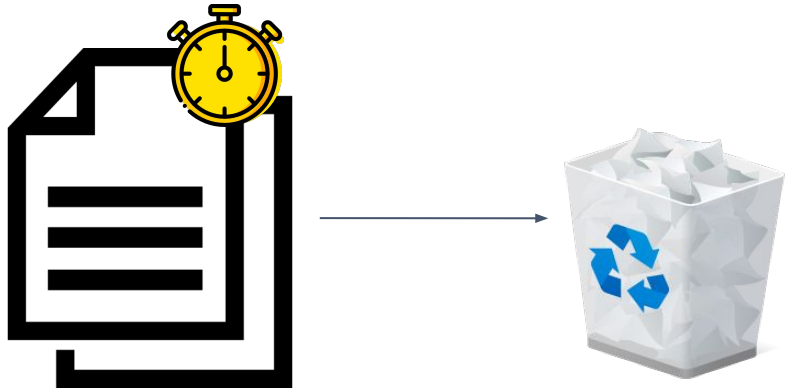
- ★ Generate a data set based on file access patterns
- ★ simulate the behavior of cache
- ★ See if the model works!

# File Eviction Strategies

*Key factor in the design of an optimal cache...*

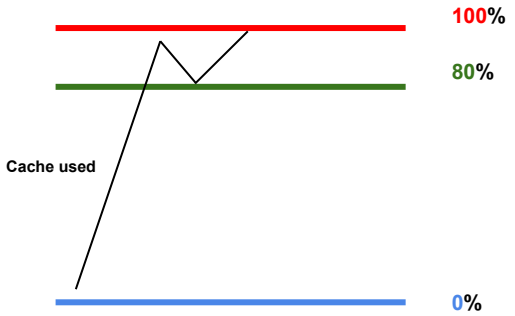
## UNUSED FOR N DAYS

**Delete files that haven't been accessed for more than N days.**



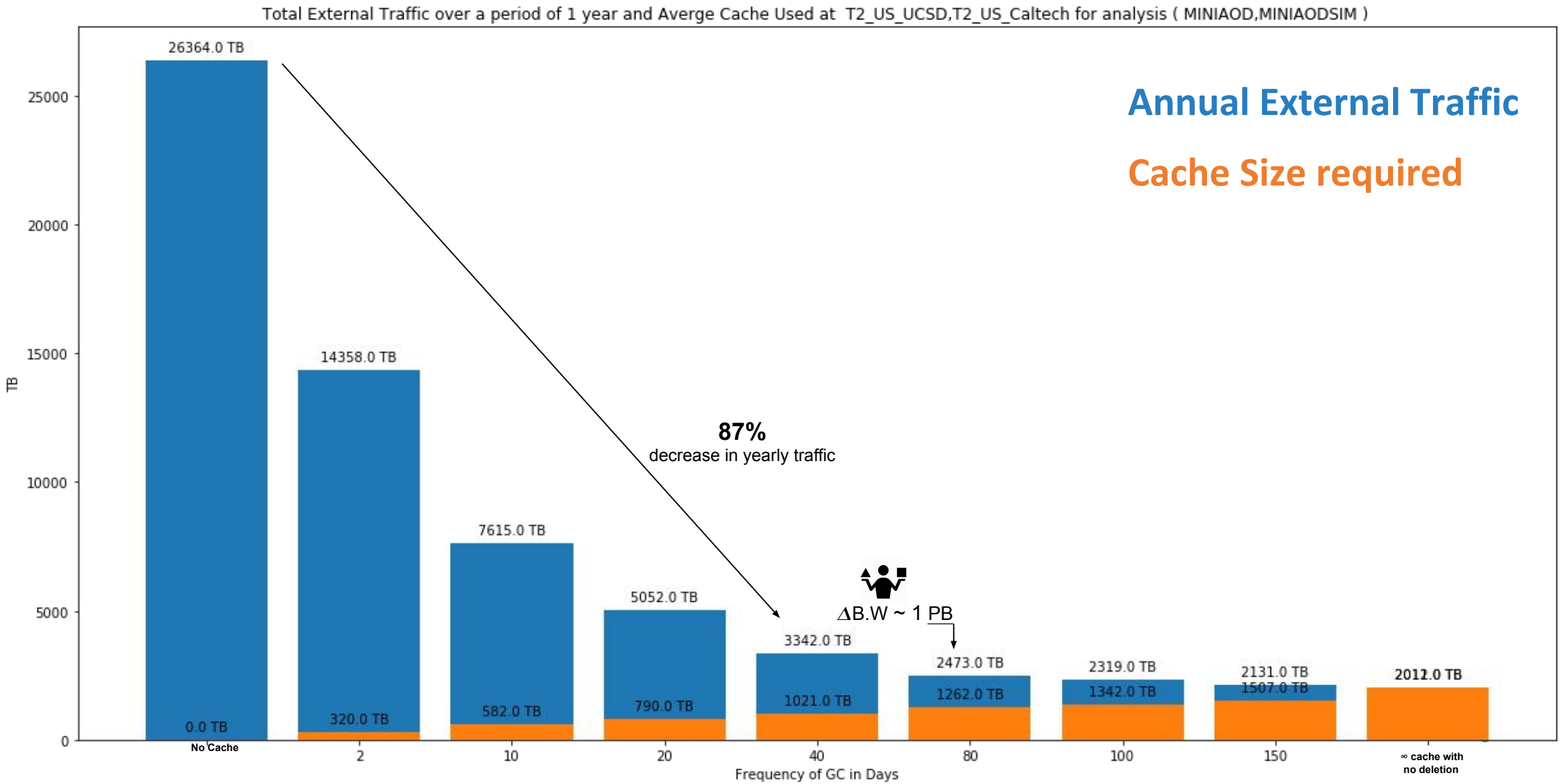
## HIGH WATER MARK

**Garbage collection is initiated when used space exceeds the size of cache. This is done by deleting the least recently used files until occupancy falls below 80%**

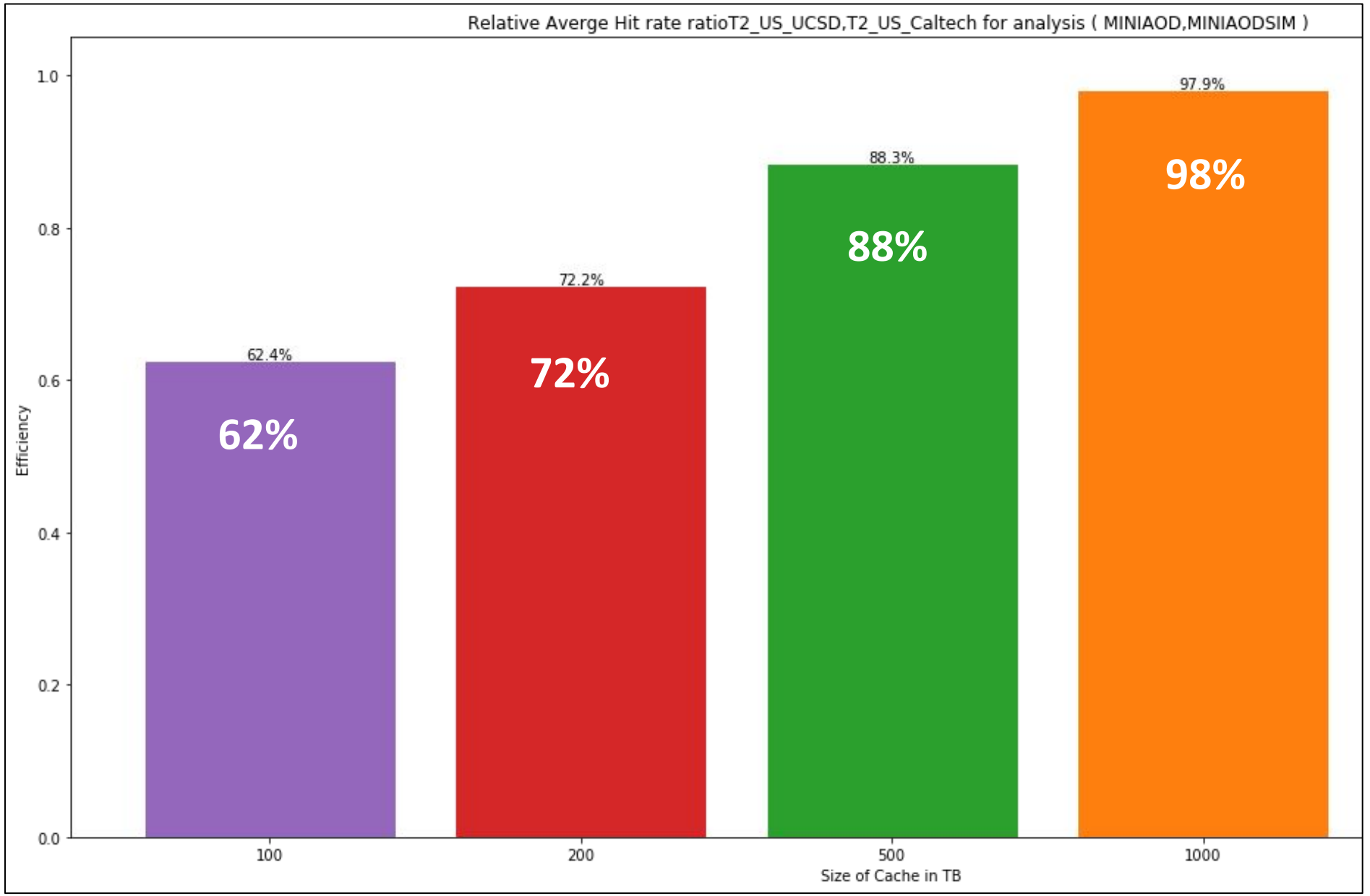



**I will now discuss some of my results...**

# N - Days Strategy

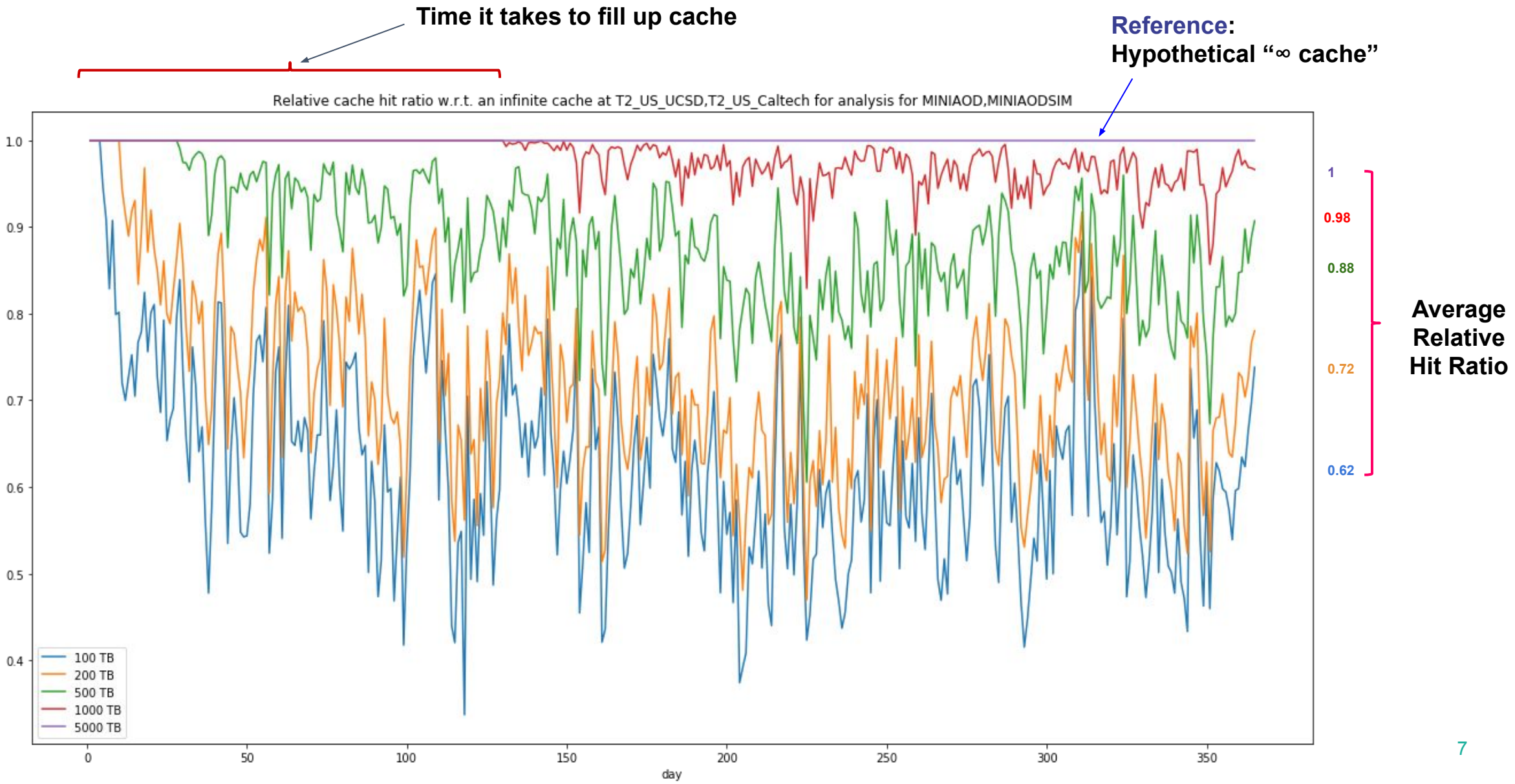


# High Water Mark Strategy - Efficiency of Cache




**BUT…**

# High Water Mark Strategy - Daily Fluctuations



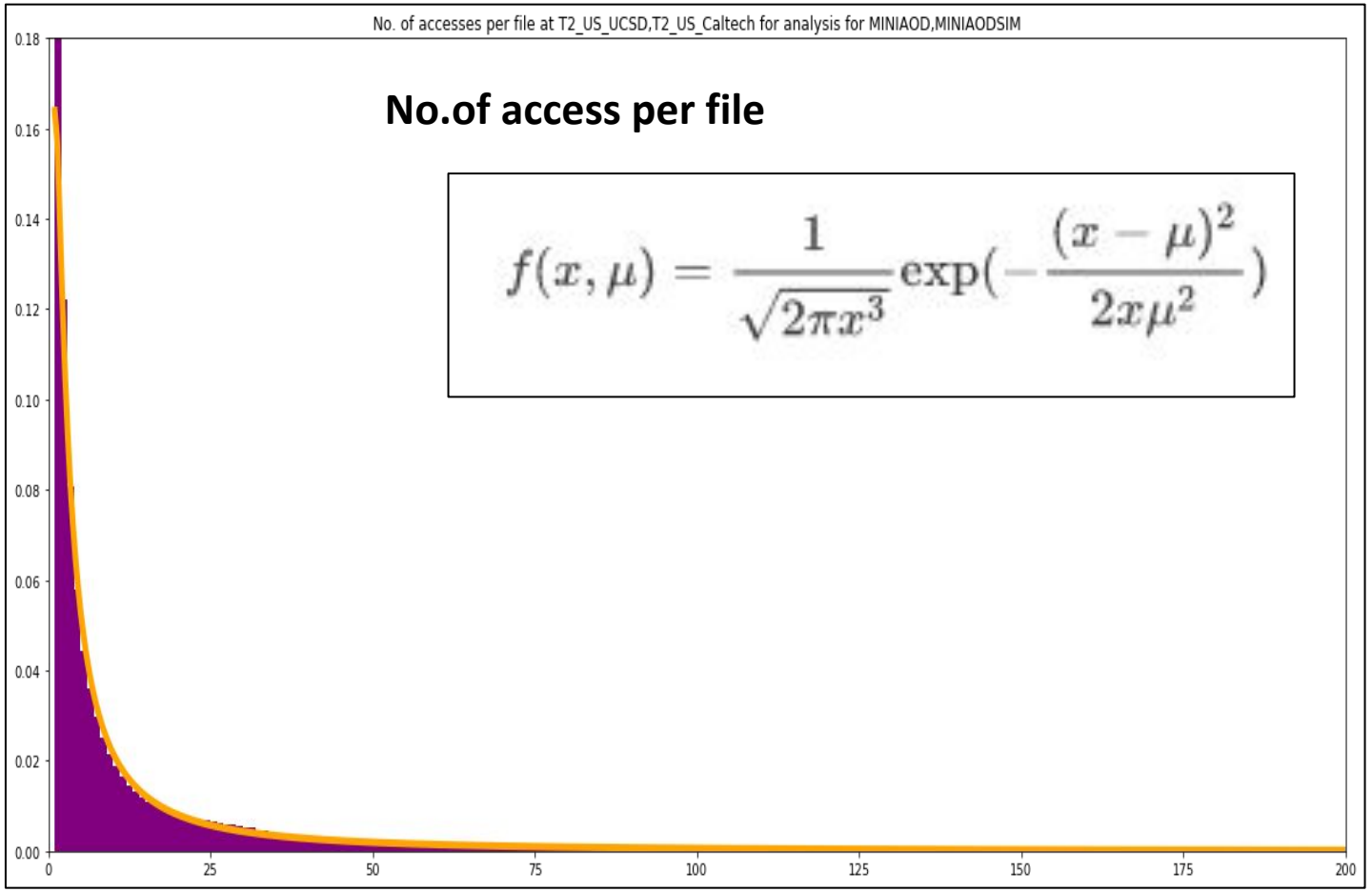
# Can we model the Access Patterns?

*Distributions of file accesses can be described by just a few parameters!*

## For Example:

**The number of access per file follows an Inverse Gaussian.**

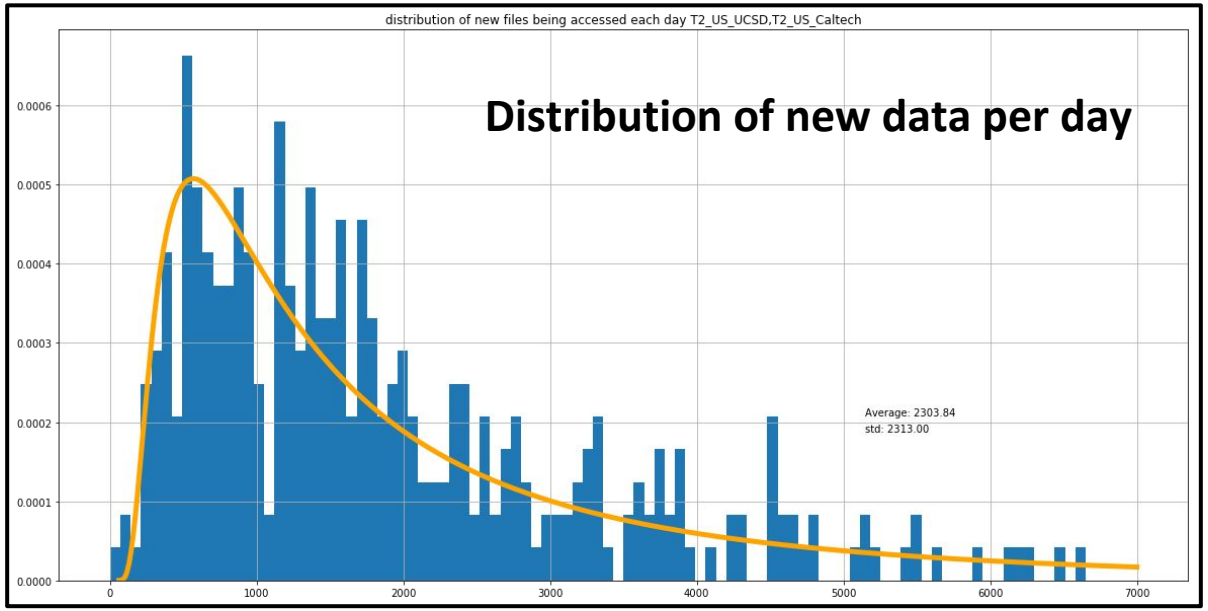
**Which is given by the probability density function:**



# Monte Carlo Model for file accesses

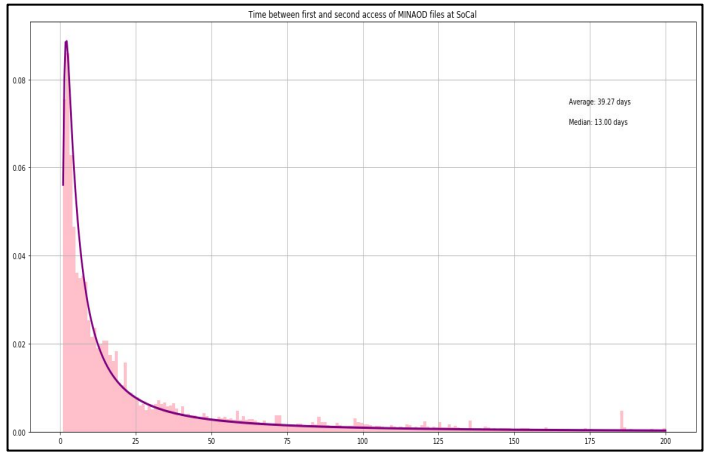
*We extract parameters from our previous analysis!*

- For every day:
  - **Generate new files**
    - Average and width of the distribution from data popularity
  - **Simulate accesses**
    - Probability based on number of accesses and time since last access
  - **Update** the status of files, write accessed files to simulated data popularity files
  - **Re-run** Cache Simulation using this data set and compare results

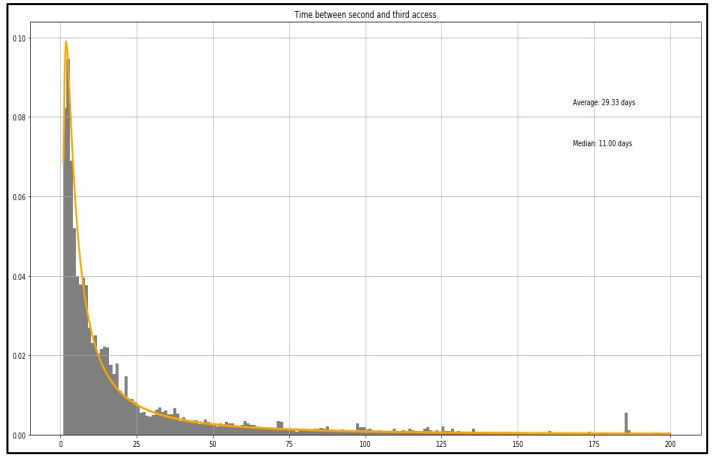


# Monte Carlo Model for file accesses

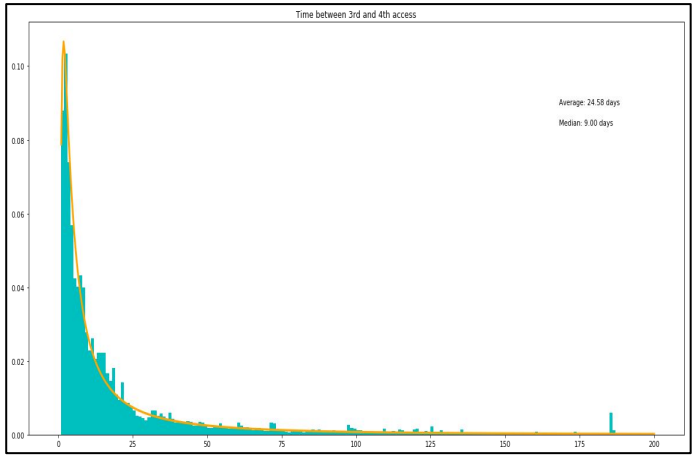
- **Simulate accesses**
  - Probability based on number of accesses and time since last access




Probability distribution of time between 1st and 2nd access




Probability distribution of time between 2nd and 3rd access




Probability distribution of time between 3rd and 4th access

• • •

f(x)

Interval in Days

**Probability that a file will be accessed for the $n^{th}$ time in a certain time interval**

# Conclusion

- *These studies have helped us understand the impact of site caches as an alternative to storage **quantitatively** and we know that this **will bring a significant improvement** in the way we store and access data across the WLCG.*
- *These studies can be used by site managers for resource planning*

## If you are interested to know more...

**Please reach out to me at:**


sshreyakrishnan@gmail.com

sarada.shreya.krishnan@cern.ch

https://www.linkedin.com/in/shreya-krishnan/