About the need of 3 Tier storage

Dmitri Ozerov
Patrick Fuhrmann

CHEP 2012, NYC, May 22, 2012

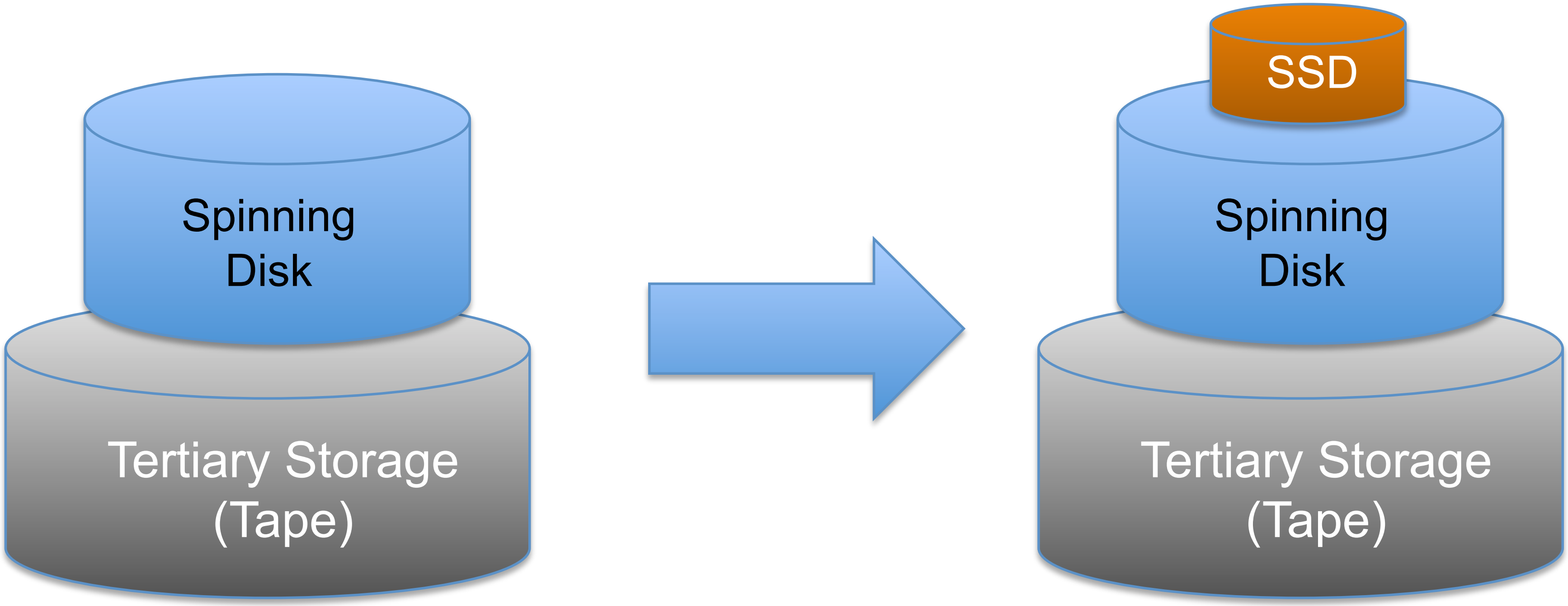
Partially evaluated by

DESY Grid Lab

Content

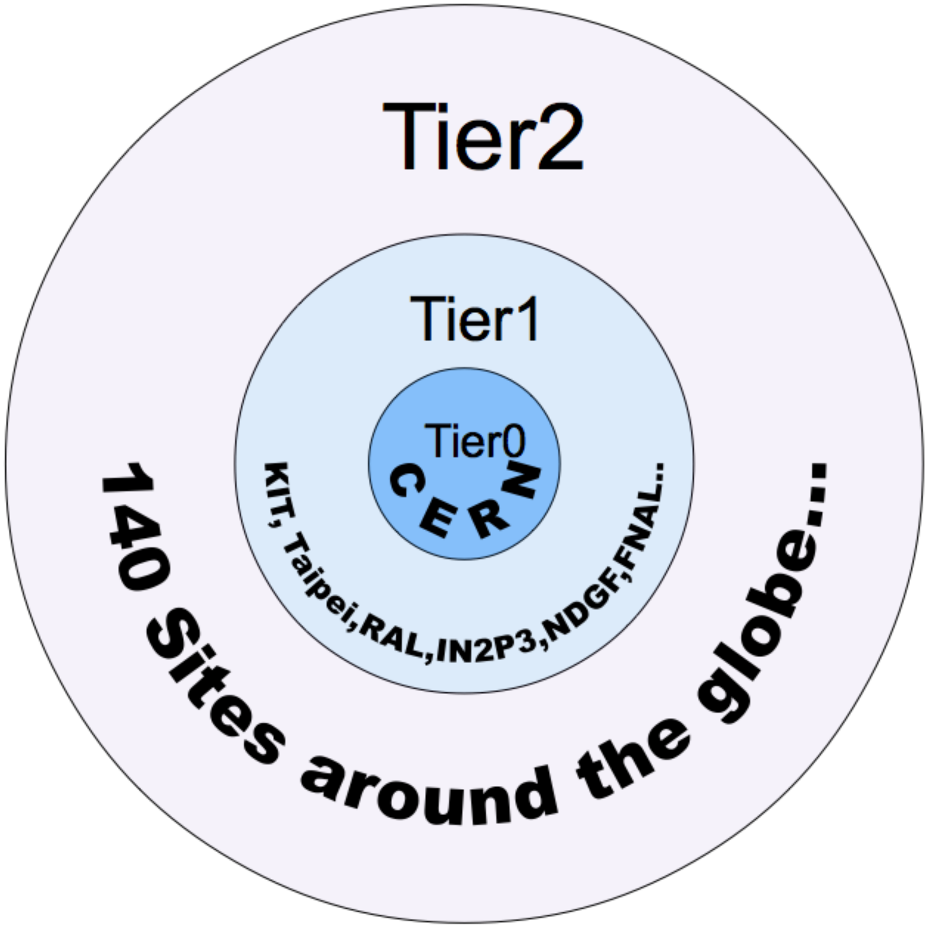
- Profile of a “large” WLCG Tier 2.
- Efficiency and performance of Storage Media.
- Replaying billing information.
- Simple cache filling mechanism.
- Improved mechanism.
- Summary

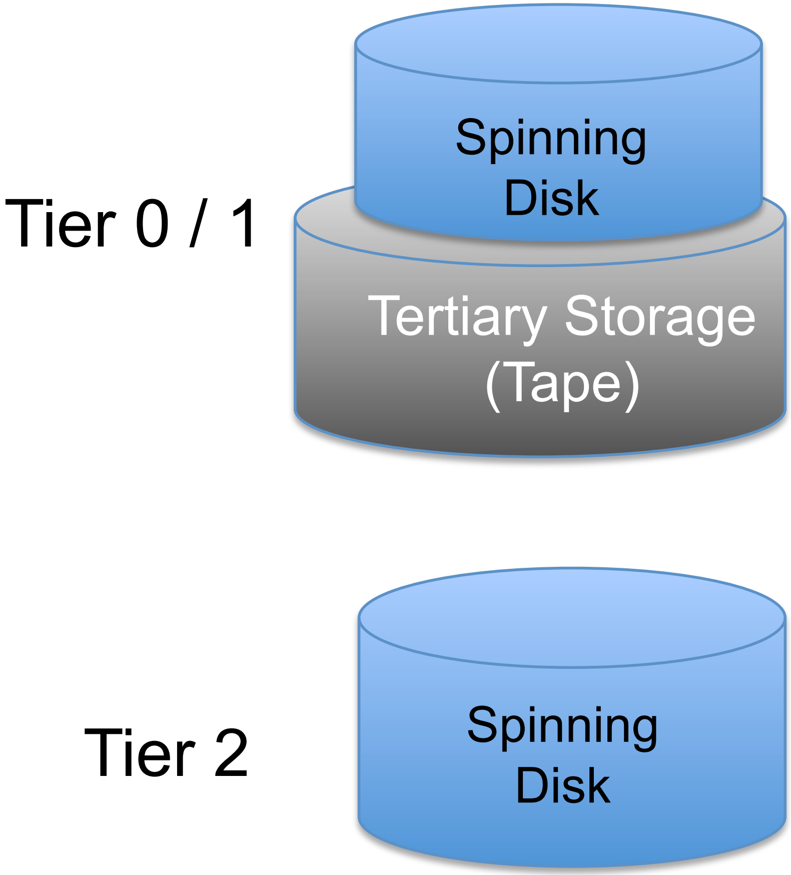
About the need of 3 Tier storage



Grid Lab

The WLCG 3 Tier Model

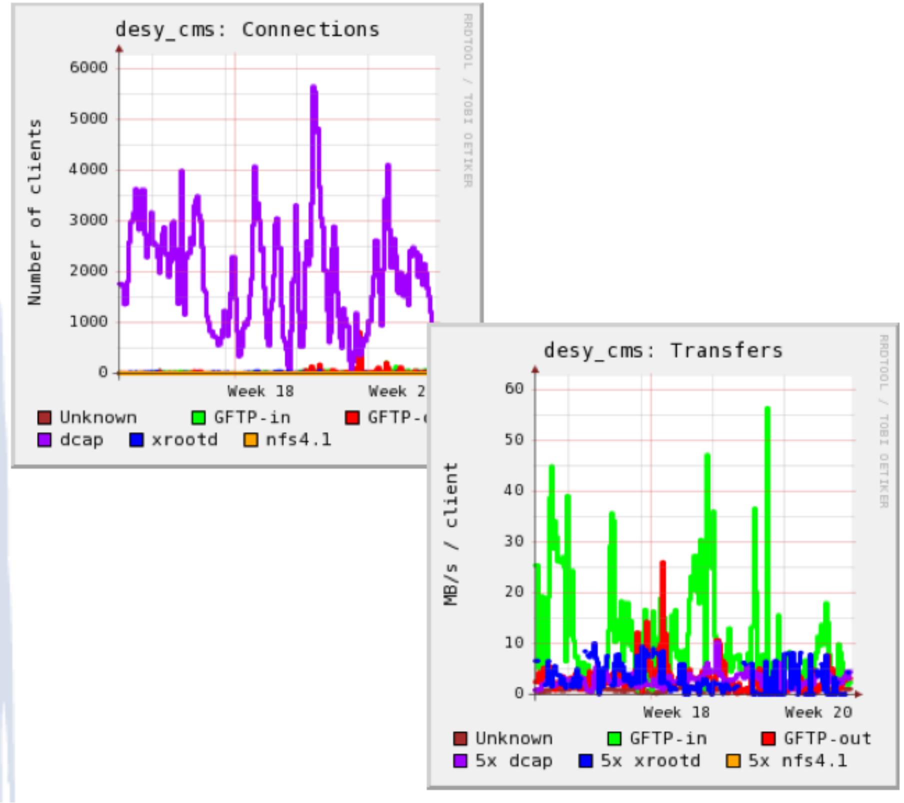

Tier2
Tier1
Tier0
CERN
KIT, Taipei,RAL,IN2P3,NDGF,FNAL..
140 Sites around the globe...

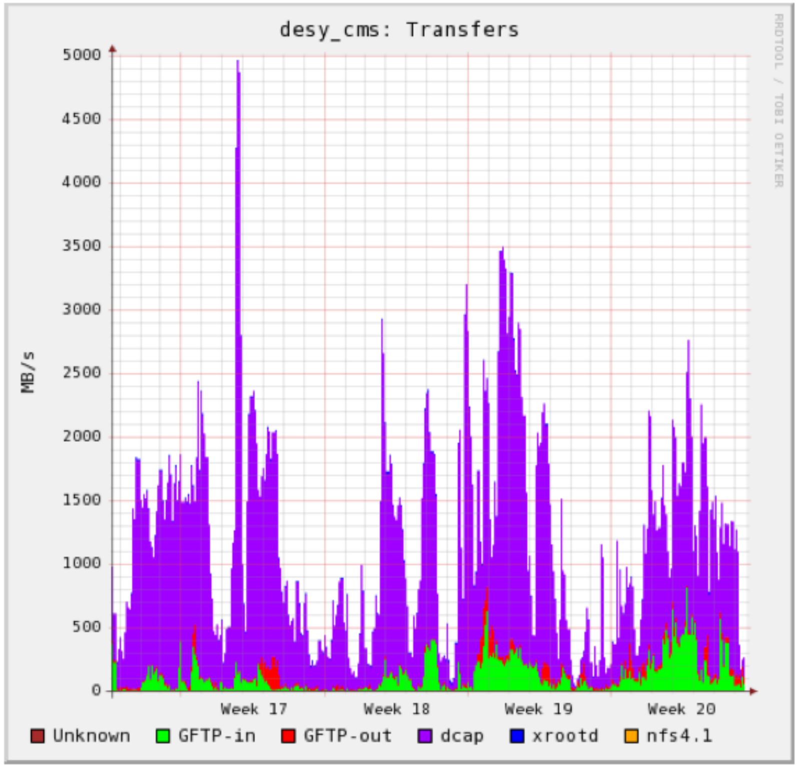


Our evaluation results are most relevant for storage providers of Tier 2/3.

Grid Lab

Activities of a large Tier 2





- Most of the clients are using dCap.
- Data is read directly from storage.
- We found a significant difference in transfer speed between dCap and gridFTP. (From few MB/Sec to 30 MB/s due to different access pattern)

Partially reading files

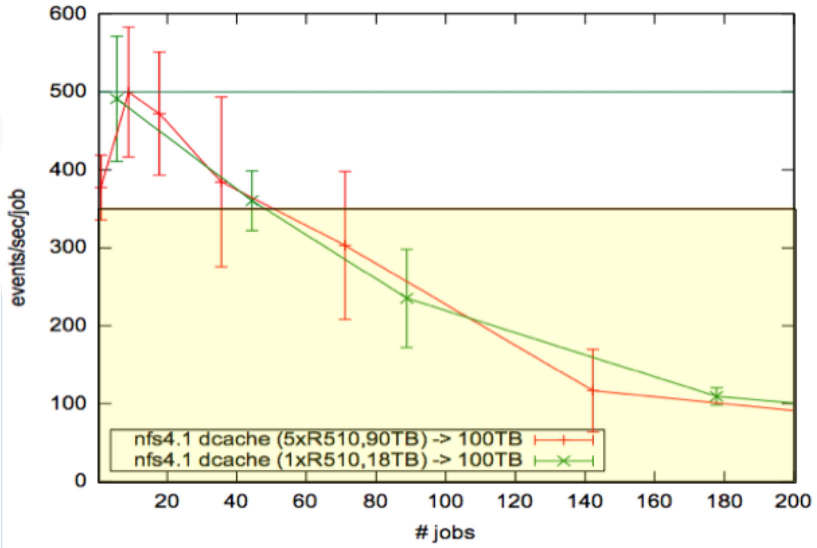
Is it more efficient to copy data to WN storage, to be processed locally ?

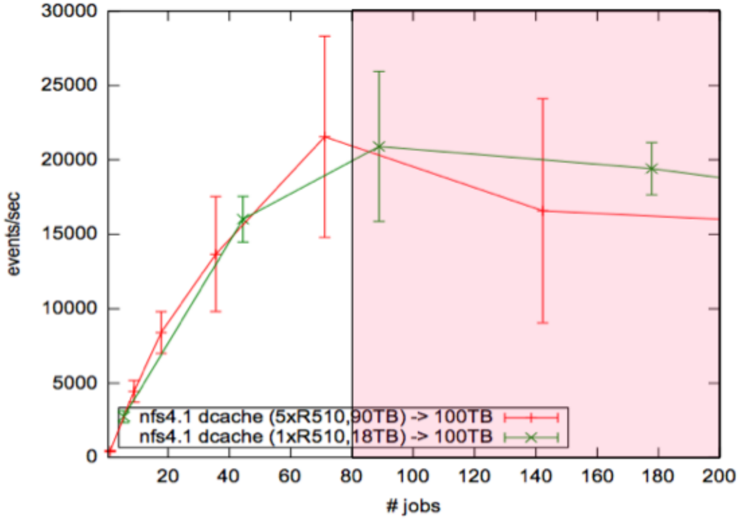
CMS User group statistics : Nov 2011

- 870 TB of data was transferred over the network; files were only partially read.
- The transferred data corresponds to 24 PB, if the entire file would have been transferred over the wire.
- The average network bandwidth: 0.33 GB/sec -> 9 GB/sec
- Load on the storage moves from dedicated effective storage pools to local space on WN

Grid Lab

Efficiency of storage hardware





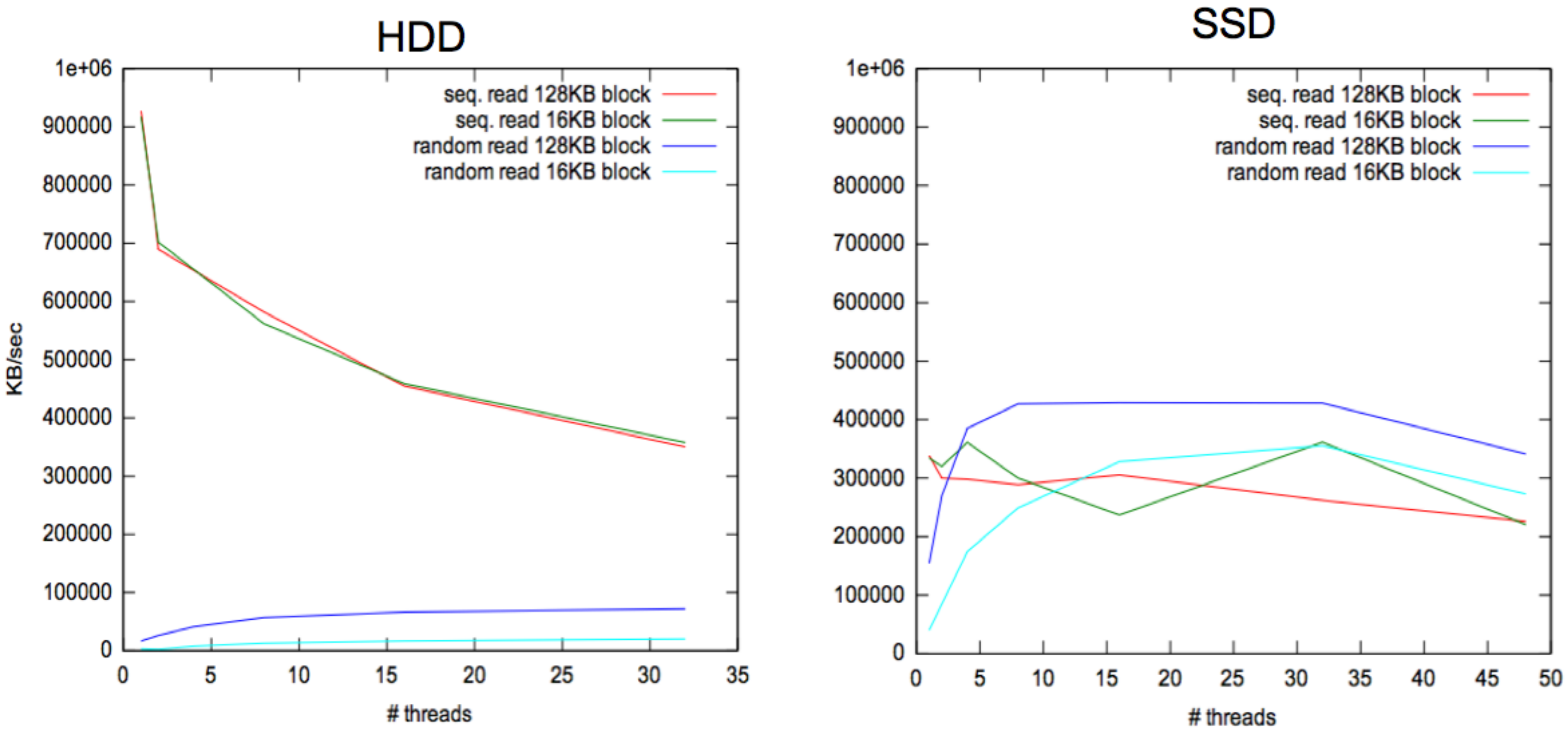
Investigations were made with the DESY Grid Lab facility

(For details, please see our poster 503)

Processing I/O intensive ATLAS Hammercloud jobs revealed two limitations:

- ✓ 50 Jobs / 100 TB – efficiency of jobs drops by 30%
- ✓ Running more than 80 Jobs / 100TB is a waste of CPU resources.

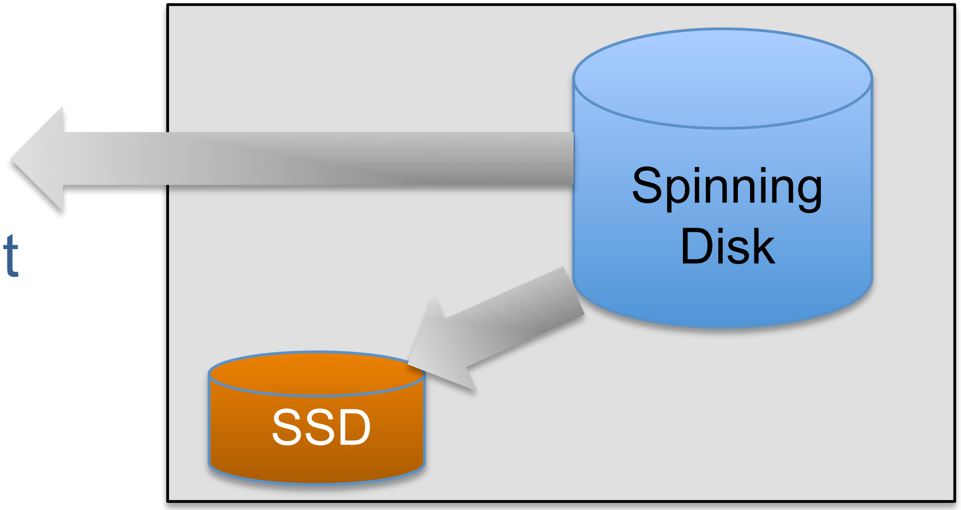
Storage media performance



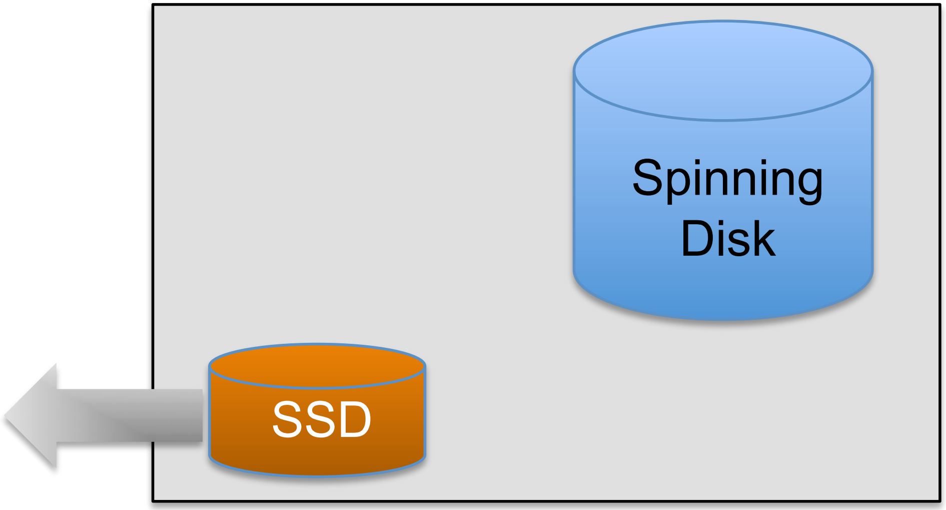
1. Traditional Disk (HDD) – good in streaming, bad in random reads SSD -> 10 times faster in random reading
2. Copying the entire file to the WN instead of directly accessing the data will be affected significantly due to the increased number of reads.

Simple Strategy

1. Get file-A from disk
2. Asynchronously copy the same file to SSD storage, removing "least recently used" files in case space is running short.

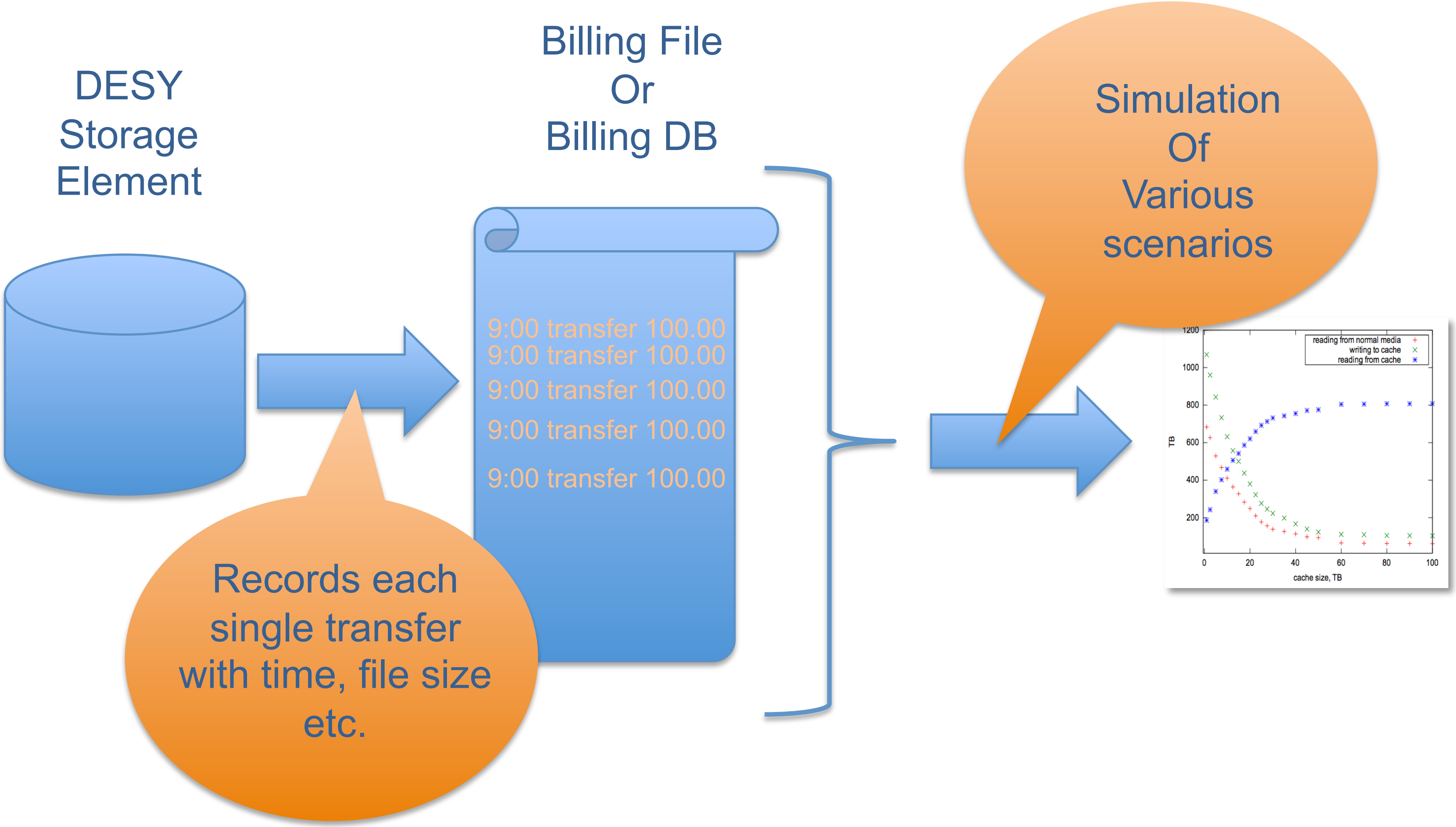


1. Next read will get file-A directly from fast SSD storage as long as file is still present.



Grid Lab

How was the analysis done ?



DESY Storage Element

Billing File Or Billing DB

9:00 transfer 100.00
9:00 transfer 100.00
9:00 transfer 100.00
9:00 transfer 100.00
9:00 transfer 100.00

Records each single transfer with time, file size etc.

Simulation Of Various scenarios

Grid Lab

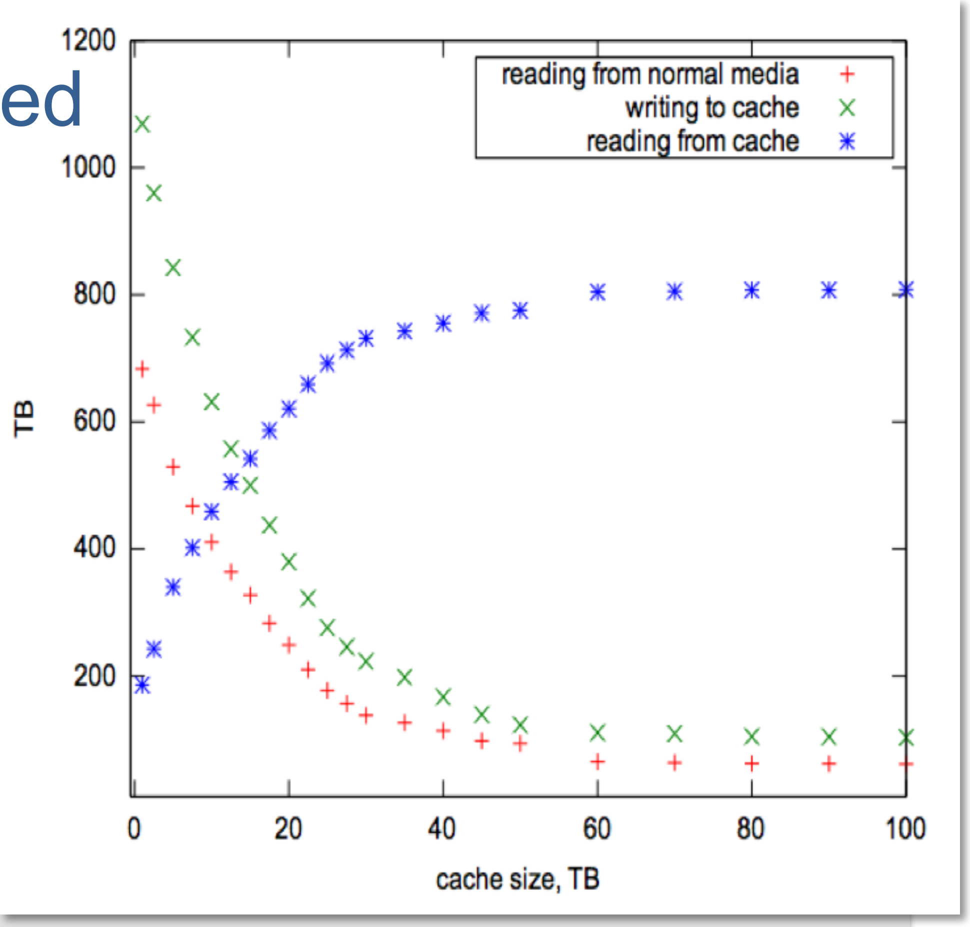
Re-play of Storage System billing info

Concerning bytes transferred

1st November 2011:

The cache is completely empty

- Read file from cache if present
- If not -> read from disk:
  - ✓ copy to cache
  - ✓ if no space left, remove "least accessed" file.



Result:

- 20 Tb Cache Size : 70% of traffic to clients is originated from SSD.
- 380TB was streamed in total (with 140 Mbyts/sec)

Grid Lab

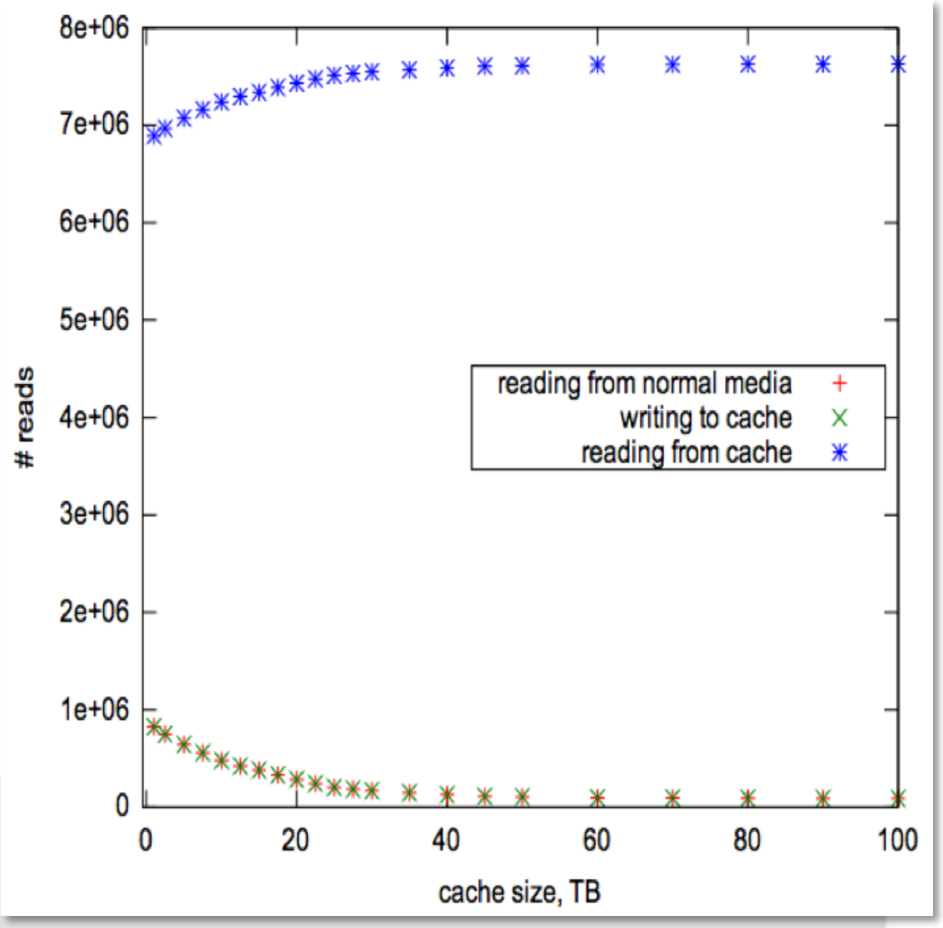
Re-play of Storage System billing info

Concerning # of reads

1st November 2011:

The cache is completely empty

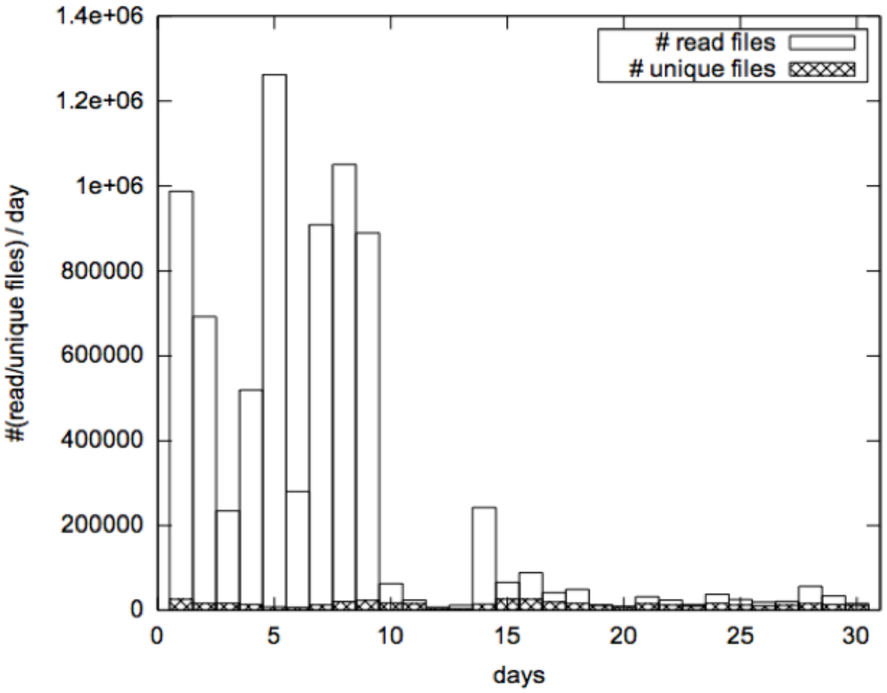
- Read file from cache if present
- If not -> read from disk:
  - ✓ copy to cache
  - ✓ if no space left, remove "least accessed" file.

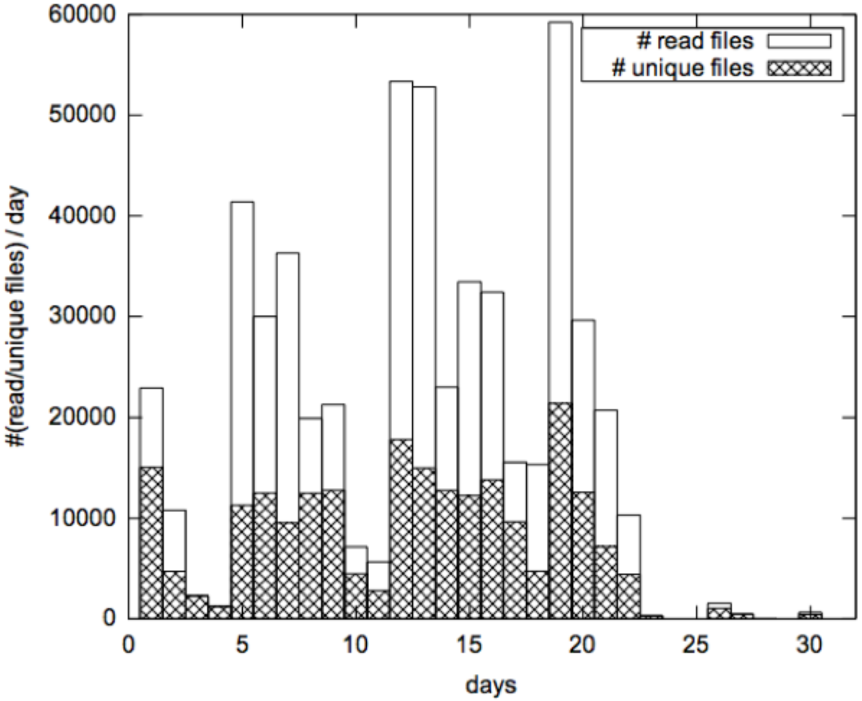


Result:

- 20 Tb Cache Size : 70% of traffic to clients is originated from SSD.
- 380TB was streamed in total (with 140 Mbyts/sec)
- Only 4% of reads from SSD

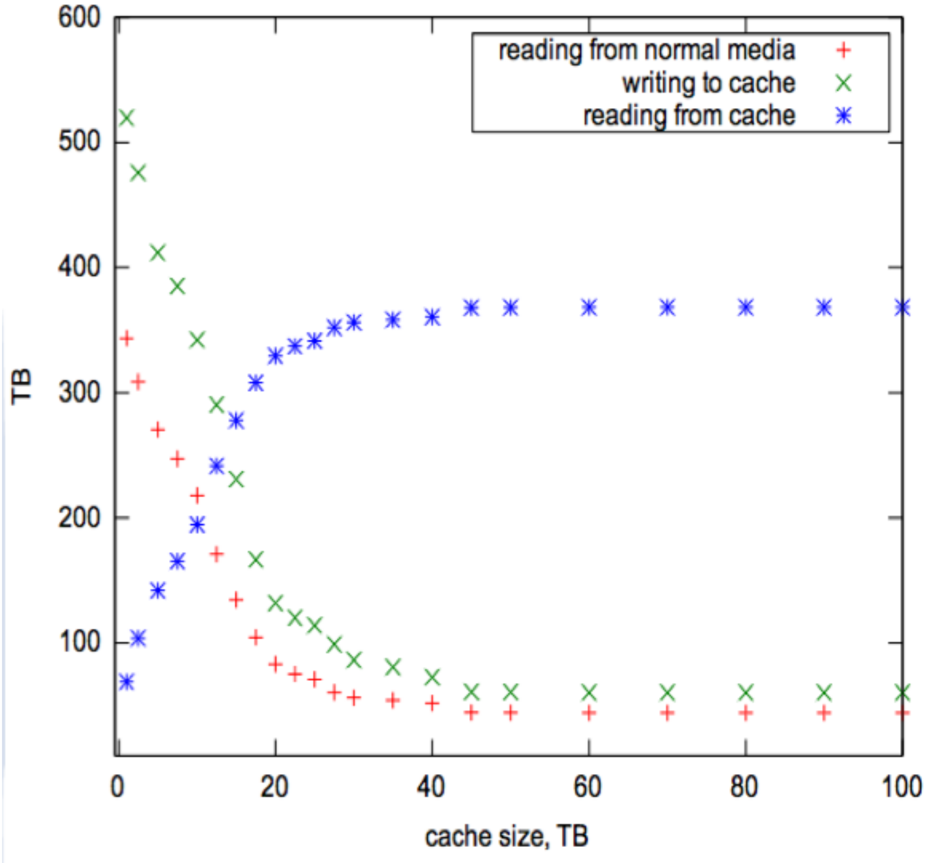
Agnostic against abnormal activity

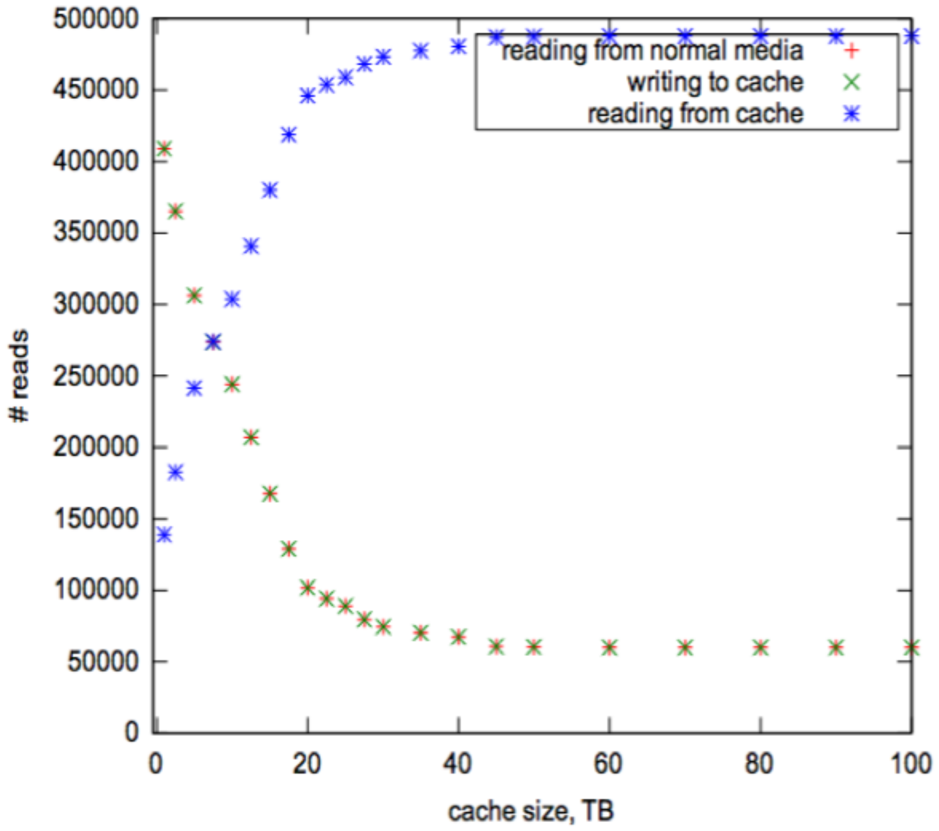




To prove that the evaluation has not been mislead by abnormal activities in the observed time frame (e.g. user code bugs or specific jobs only seen those months) an additional month was analyzed.

Results from a less active month

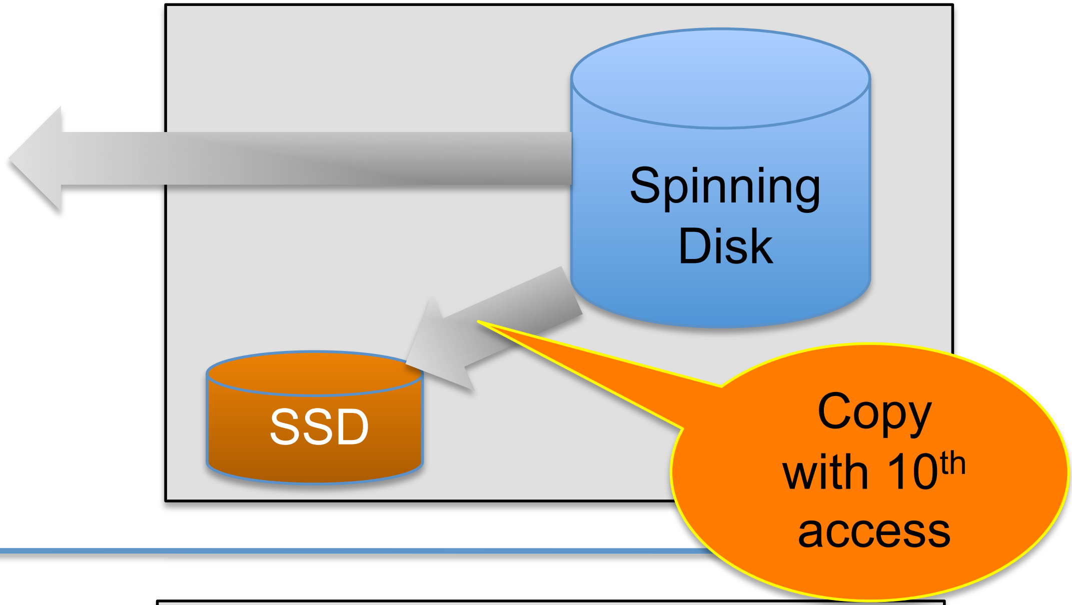




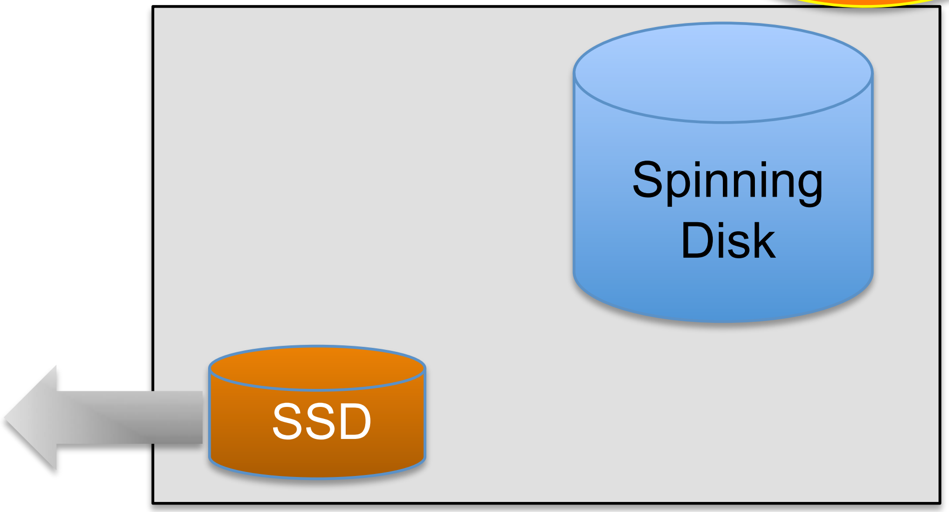
With 20 TB of cache, 80% of reads are redirected to the cache

Improved Strategy

1. Get file-A from disk
2. Read the file up to 10 times
3. Asynchronously copy the same file to SSD storage with the 10th read

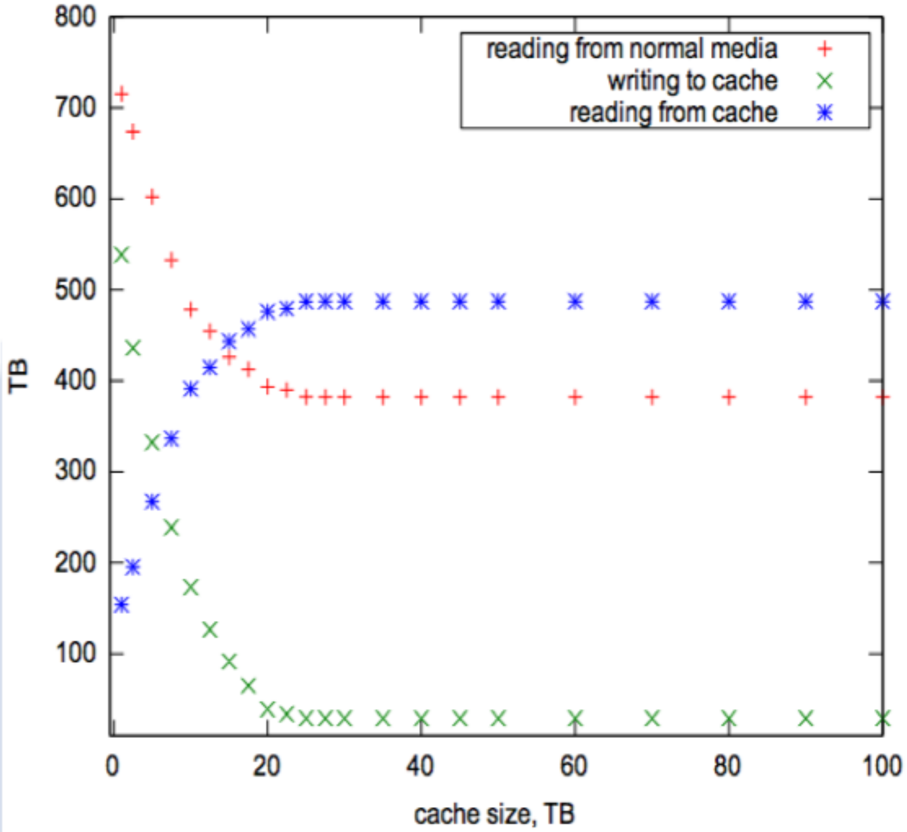


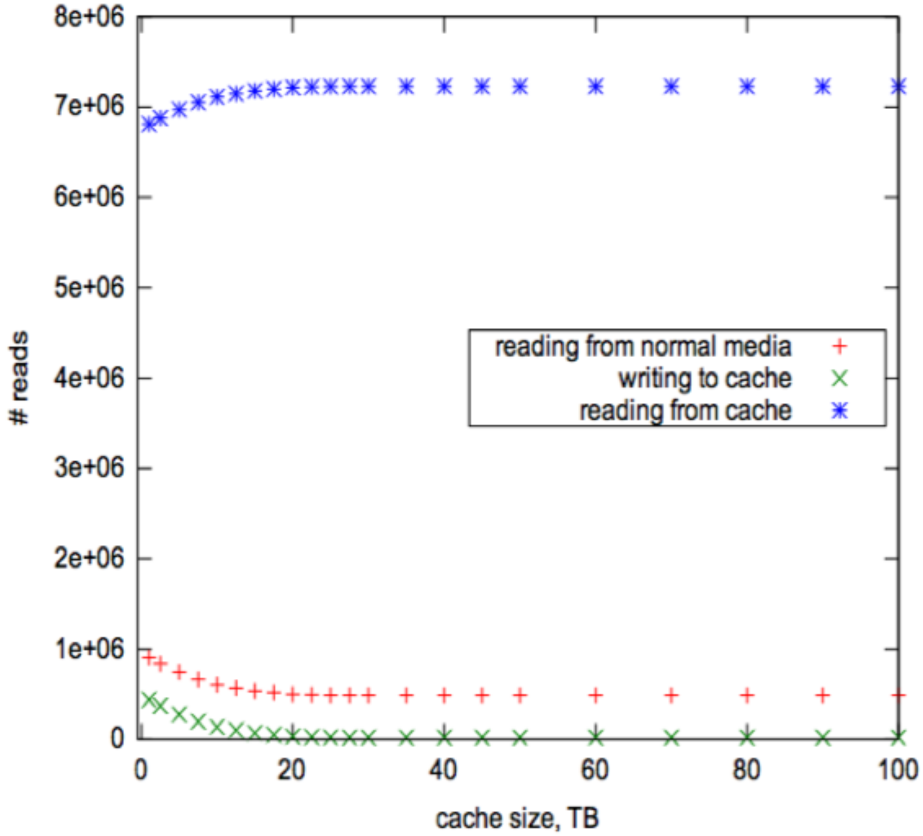
1. Next read will get file-A directly from fast SSD storage



Grid Lab

Copy to SSD cache after 10 file accesses





With the improved (tunable) algorithm, we are able to even reduce data transfer, while keeping high cache access rates.

Summary

- Storage systems based on traditional spinning disks have a limited I/O performance.
- Adding SSDs, as an additional cache layer in currently deployed storage solutions (DPM, dCache..) can boost I/O performance significantly.
- An analysis of CMS user activities has shown that only 10% of cache, results in 70% of random reads from the cache.
- A storage system with 3 Tiers is transparent for the user and cost effective for sites.