Modeling and Simulation of Tape Libraries for Hierarchical Storage Systems

Jakob Lüttgau[1,2] and Julian Kunkel[2]

[1]University of Hamburg
[2]Deutsches Klimarechenzentrum GmbH (DKRZ)

Abstract

The variety of storage technologies (SRAM, NVRAM, NAND, Disk, Tape, etc.) results in deep storage hierarchies to be the most feasible choice to meet performance and cost requirements when handling vast amounts of data. Long-term storage systems employed by scientific users are mainly reliant on tape storage, as it remained the most cost-efficient option. Archival systems are often loosely integrated into the HPC storage infrastructure. With the rise of exascale systems and in situ analysis also burst buffers will require integration with the archive(Klasky et al., 2015). Exploring new strategies and developing open software for tape systems is a hurdle due to the lack of affordable storage silos and availability outside of large organizations and due to increased wariness requirements when dealing with ultra-durable data. Lessening these problems by providing virtual storage silos should enable community-driven innovation, and enable site operators to add features where they see fit while being able to verify strategies before deploying on production systems. The work contributes to the community at large by providing:

- Isolated workflows and models for software and hardware common in tape systems.
- A prototype simulator for hierarchical storage system within data centers.
- Proof of concept that simulation to optimizing data centers is feasible.

Model Overview

Overly abstract models may require extensive refinement to reflect the emergent behavior for complex systems. To adequately approximate the performance of real-world hierarchical storage systems when applying structural or algorithmic changes the full software and hardware stack has to be taken into account. Figure 1 illustrates a possible model overview as appears to be adequate for automated tape archive in combination with disk-based caches.

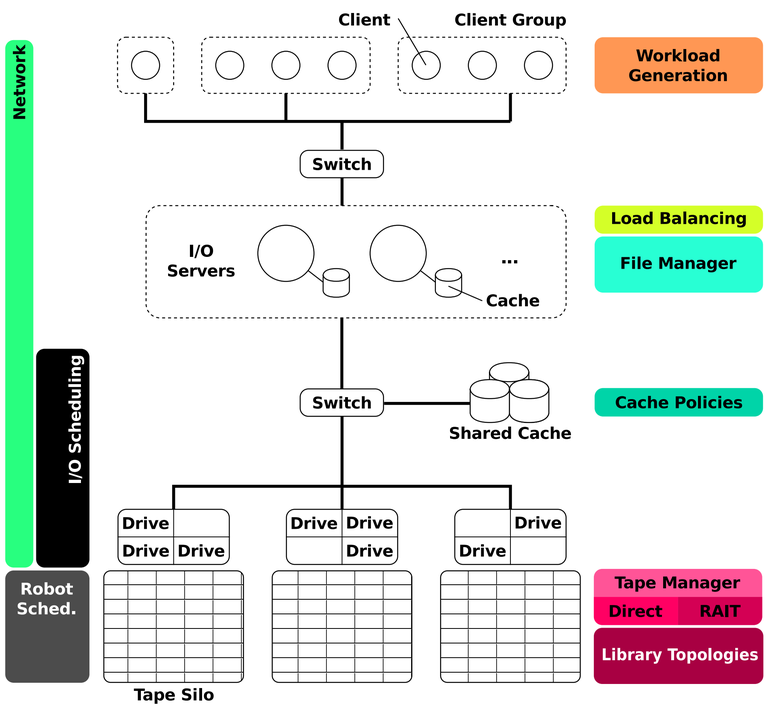

Fig. 1: A model overview for a tape system including hardware and software components which are required to allow meaningful system approximation.

Workload Models & Generation

Events within tape systems can be traced back to a variety of causes. The focus of this work is to improve the quality of service (QoS) when serving user induced workloads. Thus special workloads, e.g., service and migration workloads are common to tape systems are reflected in the coarse design of the simulator but no service workloads are generated.

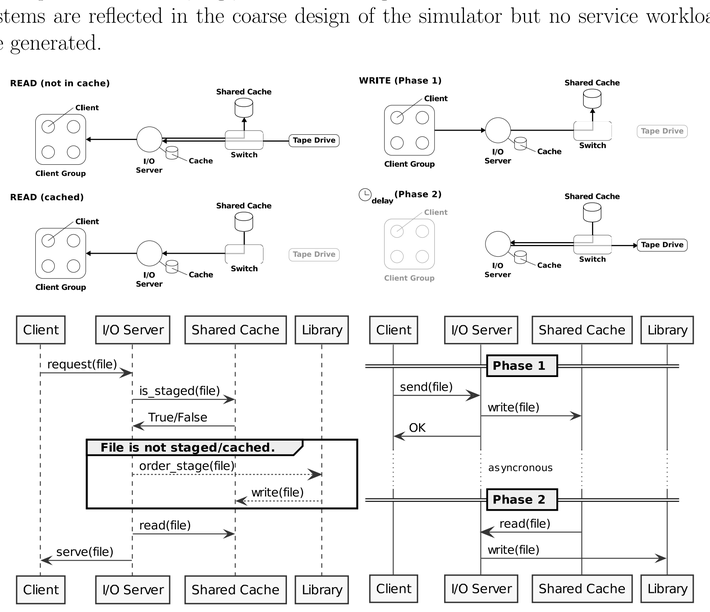

Fig. 2: Handling of `READ` and `WRITE` workloads.

Verification runs for the tested systems so far suggest that omission of service workloads may only introduces a minor approximation errors. Figure 2 illustrate the staging process triggered when read or write requests are issued to the system. Xferlog files as used by some FTP servers can be directly used as workload traces to stress the virtual system.

Hardware Models

Using discrete event simulation to approximate an actual system requires various models to calculate the time penalty associated with completing a certain action. Tape systems spent a lot of time waiting for mechanical systems. Since we would like to simulate weeks to months of activity on the system, amounting to the movement of multiple petabytes, models for the lower level electronics are kept very abstract.

Library Topologies

To calculate the times it takes to receive and mount a tape, additional models that reflect the dynamics within a tape library are required. How tape shelves, tape drives and robots are organized within library complexes can vary wildly. This reflects the fact that usage scenarios very from site to site.

Graph-based Topologies

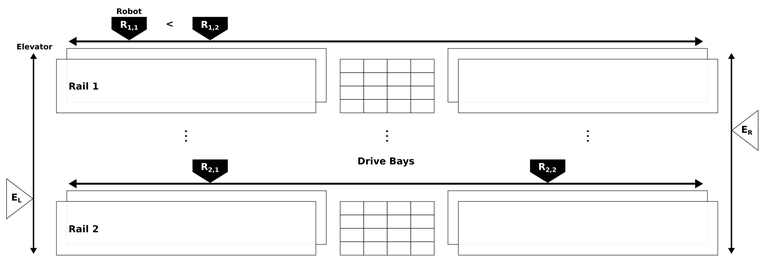

Fig. 3: Principal paths which tapes and robots can take on the railing system within a StorageTek SL8500 Module.

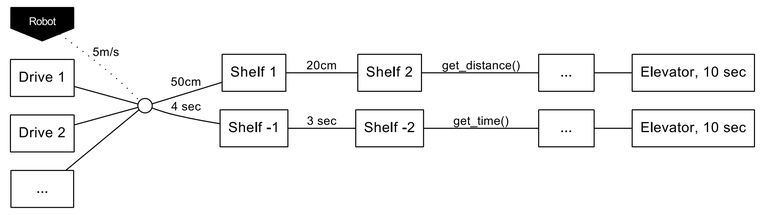

Fig. 4: Graph-based topologies for coarse system topology.

Library complexes can be easily modelled using a graph-based (Fig. 4) topology. The graph is influenced by the paths robots and tapes can take.

Specialized Models: 2D Topology

Specialized models allow to more accurately and more efficiently calculate time penalties. E.g. StorageTek SL8500, originally U-shaped, maybe projected onto a 2D Cartesian grid without risking to accumulate large approximation errors.

Software Models

Hardware devices and subsystems such as the library topology and the network are controlled by software. This section collects abstractions of relevant software components that could also be found in real deployments. As such there are mostly algorithms for load balancing, scheduling for resource management as well as for file- and tape system management. QoS is sometimes artificially limited by quota systems or provide multiple service levels. In addition, some requests may enjoy priority status and should be handled privileged. Finally, the robots in the libraries need to be orchestrated, so that they serve tape drives as fast as possible.

Scheduling/Replacement Strategies

The simulator was developed to allow to explore alternative scheduling strategies at different stages in the storage hierarchy. To do so a variety of queues is kept each which allows to replace the scheduling strategy with custom strategies.

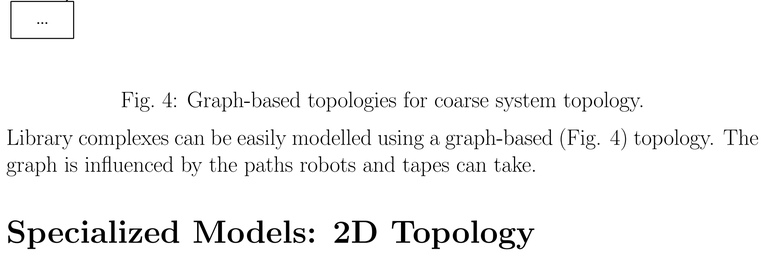

Fig. 5: Scheduling requests within a hierarchical archive.

Verification

For the evaluation of the simulator, workload traces of FTP traffic covering of roughly a month of FTP traffic on the production archive of DKRZ were used to stress the virtual system. Different metrics available through the original monitoring system where collected within the simulation to reproduce the plots for comparison (see Figure 6). The used trace in this case was chosen to also see how the virtual system recovers from a maintenance phase.

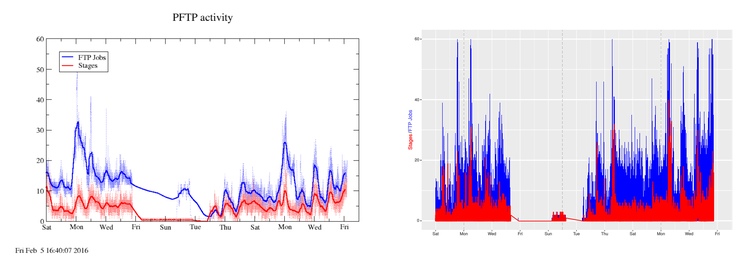

Fig. 6: FTP activity and the number of stages observed by the monitoring of a real system (left) compared to a simulation run (right).

Exploration & Optimization

For more educated decisions when acquiring new systems it would be interesting to compare different configurations in advance of buying a system. Similarly, when trying new scheduling or cache replacement strategies where robustness and performance improvements are hard to extrapolate from isolated test cases.

Use-Case: Drive Count Variation

By varying the number of drives (e.g. 30, 45, 60) we can verify if changes with expectable outcome are consistent with the simulation outcome. With more of these tests confidence in the simulation results increases and more experiments with more elaborate strategies can be pursued.

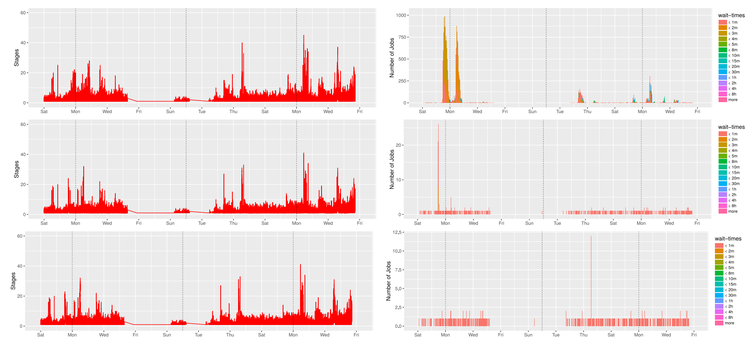

Fig. 7: Influence on the number of stages (left) and on the wait-time (right) when varying the number of tape drives.

Quality of Service (QoS)

A common question when dealing with tape systems concerns the number of tape drives to be used. For the use-case on drive count variation Figure 8 shows how the total time to complete requests improves as drives are added. Plots like this can then be used to quickly find which service guarantees can be given for a certain configuration.

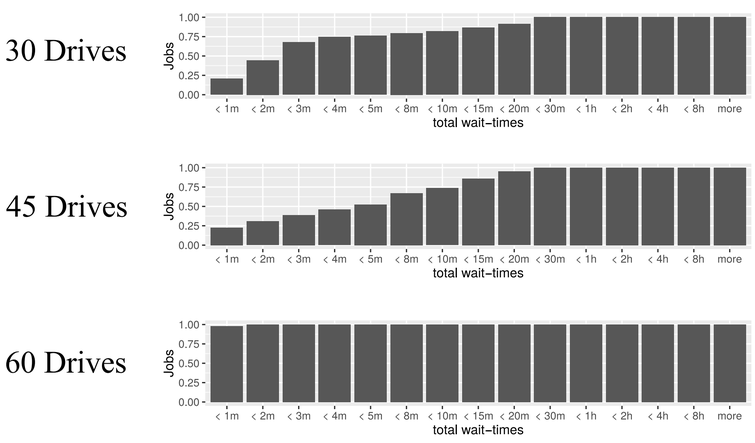

Fig. 8: Easy to read changes in QoS when varying the number of tape drives.

Summary & Future Work

Equipped with comprehensive models and a simulator to approximate tape archives within hierarchical storage systems, it is now possible to improve modern tape libraries without requiring a physical tape library for testing. Workflows to experiment and asses the performance directly influenced the design of the simulator. Consequently the next step is to put it to use in experiments. Also gradually turning the simulator into a open source tape library management solution for production systems is becoming an option.

- Improvement of the configuration process, for example through a GUI to setup library and network topologies.
- Mature architecture for physical tape library management.
- Porting core APIs to more efficient programming languages so they would be suitable for deployment in real system.

From a research perspective the interesting part of a simulation is to apply it to practical problems to learn and generate new insight. In particular, this might include parametrized Monte-Carlo methods to optimize for budgets or quality of service. In addition, it would be useful to have a comprehensive database for benchmarks that collect the characteristics of drives, libraries and other devices. The database should also include component prices, though utilizing online price comparison APIs to fetch prices on demand may also be an attractive option when combined with a way to apply a correction factor for discounts.

References

Klasky, S., Grider, G., Oldfield, R., Felix, E., Shipman, G., Gary, M., Wu, J., Richards, D., Neely, R., Glass, M., and Williams, D. (2015). DOE: Storage Systems and Input / Output to Support Extreme Scale Science.

Lüttgau, J. (2016). Modeling and Simulation of Tape Libraries for Hierarchical Storage Management Systems. Master's thesis, Universität Hamburg.

Acknowledgment

Special thanks belong to Wolfgang Stahl for providing invaluable insight into the production archive at DKRZ. This work is part of the ESiWACE project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DKRZ
DEUTSCHES KLIMARECHENZENTRUM