Modeling and Simulation of Tape Libraries for Hierarchical Storage Systems

Jakob Luettgau
Deutsches Klimarechenzentrum GmbH
Bundestraße 45a
22455 Hamburg
Email: luettgau@dkrz.de

Julian Kunkel
Deutsches Klimarechenzentrum GmbH
Bundestraße 45a
22455 Hamburg
Email: kunkel@dkrz.de

***Abstract*—The variety of storage technologies results in deep storage hierarchies to be the only 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. 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. Different models for the individual components in tape systems are developed. The models are then implemented in a prototype simulation using discrete event simulation. It is shown that the simulation can be used to approximate the behavior of tape systems deployed in the real world and to conduct experiments without requiring a physical tape system.**

I. Introduction

With the increasing demand for long-term storage, automated tape libraries will likely remain a integral part of the storage hierarchy for many years to come. Tape as a storage medium has many attractive properties. It es fairly resilient and provides high data densities, but by far the most important factor is that tape is very affordable in comparison to other storage technologies. Standardization efforts such as LTO make tape attractive and future proof, thus protecting investments. Despite tapes long history the technology is still advancing[1], [2], though on a artificially slowed schedule. In an effort to speed up innovation and enable also talent without access to large scale tape systems to contribute, the objectives of this work were to develop a simulator, the tools, and primarily the appropriate models required to reproduce the dynamics of hierarchical storage systems and tape libraries. Modeling a complete tape system is a complex task because so many different components are involved. It was possible to identify a number of key components that are essential to any tape system. It was further possible to provide comprehensive models to describe the dynamics of many of these key components. In particular models for hardware and software components where proposed and isolated in such a way that turning to a more accurate model is possible.

II. Related Work

Efforts to improve tape storage system often focus on advancing the technology that is used to read and write tape. This is mostly in the domain of vendors and not much of the research conducted is published to protect a business advantage. More openly discussed are strategies for data placement on tape [3], [4], [5] or the magnetic representation [6]. Such strategies maybe exploited by higher level algorithms, but tape drives and hardware generally does not provide fine grained control to the users. Another form of placement which was researched but has not yet found it's way into many production system is RAIT [2] or TapeRAID and combinations of RAID and Tape[7]. Pure tape systems cease in relevance and hybrid and hierarchical storage systems promise to provide cost efficient solutions with the best properties of all technologies. [6] stresses the opportunities of automation, which enabled scalable solutions that seamlessly integrate into storage hierarchy at large. [8] focuses especially on the integration of disk and tape. [9] explored different object placement strategies with in tape libraries to optimize tape switch, data seek and transfer times using a simulation.

III. Hardware Models

The problem with computer systems is the complexity that unfolds because of the large number of possible combinations for hardware and software. Modeling hardware is particular cumbersome because in the real world the performance of a device emerges as a result of the laws of physics, but for a virtual model the dynamics have to be understood and abstracted. For standardized components it often is relatively easy to find a model that is adequately applicable for the whole class of of components. Composite components, such as the library topologies turn out to be harder to generalize in a simple way than expected. By mixing mostly 2D and a graph-based topology approaches very good approximations of the library dynamics can be achieved. Another problem occurs with proprietary designs for which detailed information are hard to find. The same is true for benchmarks and a comprehensive catalogue of performance parameters. The network is

an integral part to hierarchical storage systems and can be used to model and simulate even low-level components and communications. A flow-based and a package-based approach were discussed. The flow-based approach appeared elegant on first sight, in retrospect it was a suboptimal design choice for the type of system because network and I/O scheduling became more entangled with other components that they would have been in a real system.

IV. Software Models

Hardware devices and subsystems such as the library topology and the network are controlled by software. For a proof of concept prototype it was often sufficient to turn to "naive" implementations. But because parts of the virtual software stack already begin to stabilize and because of the modular structure more sophisticated algorithms can be integrated into the already existing virtual software stack.

V. Evaluation

We could show that the footprint of the current simulation is fairly moderate at 400-600 MB main memory and 3-4 hours runtime. The library and network topologies, that differ from data center to data center, can be configured simply by providing a XML file. Adding or removing components is always possible through the topology APIs, so we can load a configuration and change a few parts by executing additional startup scripts that e.g., "installs" additional drives. A combination of commend-line arguments, configuration files and well designed APIs has proven to make experimenting much easier. Besides starting a simulation, collecting results is important. Many tools scatter result files all over the current working directory. The simulator creates a unique and timestamped directory for every run to store generated data. In addition components can register named CSV reports which can be used for logging of performance and debugging information. This structure proved useful, as it made aggregating results and generating reports much more user and machine friendly. Verifying the accuracy of the simulation required additional tools and reporting capabilities. In particular it is now possible to directly use FTP *xferlog* files found in real systems as workload traces for the simulation. Feeding these workload traces to simulations prepared to use a library configuration similar to the one at DKRZ showed that the models can be used to approximate the behavior of a real system. Two key performance metrics were measured and compared to records from the DKRZ's tape library monitoring as a reference. For both metrics, the number of staged files and wait-times in the systems stage queue, the simulator could demonstrate a behavior analog to the one observed by the monitoring. While fine-tuning is still required, the approximation is good enough to conduct first experiments. As a proof of concept, the number of tape drives installed to the virtual system was varied. Consistent with our expectations a reduction of tape drives degraded the tape systems performance and increased wait-times. Likewise an increase of tape drives reduced the wait-times and files were staged quicker.

VI. Conclusion

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 now within reach. From a research perspective the interesting part of a simulation is to apply it to practical problems to learn and generate new insight which was out of the scope the thesis. The logical next step is to carefully construct experiments with the current system and iteratively improve the tools required to conduct more experiments. In particular this might include parametrized Monte-Carlo methods to optimize for budgets or quality of service. The foundation to perform these kinds of experiments is provided with the work of this thesis. 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.

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.

References

[1] R. E. Fontana, G. M. Decad, and S. R. Hetzler, "The Impact of Areal Density and Millions of Square Inches (MSI) of Produced Memory on Petabyte Shipments of TAPE , NAND Flash , and HDD Storage Class Memories," *Proceedings of the 29th IEEE Symposium on Massive Storage Systems and Technologies*, 2013.

[2] C. J. Hughes, D. Fisher, K. Dehart, B. Wilbanks, and J. Alt, "HPSS RAIT Architecture 07/01/2009," 2009.

[3] A. Dashti and C. Shahabi, "Data placement techniques for serpentine tapes," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, pp. 1–10, 2000.

[4] D. Pease, A. Amir, L. Villa Real, B. Biskeborn, M. Richmond, and A. Abek, "The linear tape file system," *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010*, vol. 4, 2010.

[5] A. Pantazi, S. Furrer, H. E. Rothuizen, G. Cherubini, J. Jelitto, and M. A. Lantz, "Nanoscale track-following for tape storage," pp. 2837–2843, 2015.

[6] R. H. Dee, "Magnetic tape for data storage: An enduring technology," *Proceedings of the IEEE*, vol. 96, no. 11, pp. 1775–1785, 2008.

[7] D. F. Lingfang Zeng, "Hybrid RAID-Tape-Library Storage System for Backup," *Second International Conference on Embedded Software and Systems (ICESS'05)*, pp. 31–36, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1609854

[8] I. Koltsidas, S. Sarafijanovic, M. Petermann, N. Haustein, and H. Seipp, "Seamlessly Integrating Disk and Tape in a Multi-tiered Distributed File System," pp. 1328–1339, 2015.

[9] X. Zhang, D. He, D. Du, and Y. Lu, "Object Placement in Parallel Tape Storage Systems," *Proceedings of the 2006 International Conference on Parallel Processing (ICPP'06)*, pp. 0–7, 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1690610