VHS: Virtuous data accesses in High-capacity Storage systems

VHS – Common 2024 ANR call – JCJC (Jeunes chercheuses/Jeunes chercheurs) – Duration: 42 months

Axe E.03 (CE25) : Sciences et génie du logiciel - Réseaux de communication multi-usages, infrastructures numériques

Valentin Honoré – École Nationale Supérieure d'Informatique pour l'Industrie et les Entreprises

Contents

Abstract

Magnetic tapes have been used to store computer data since the 1950s, so the layman now often considers it as an outdated technology. However, tape storage is still and will remain essential in many fields such as academic research, international organisations or cloud companies for its strong practical benefits: low cost per TB, low energy consumption, longevity etc... This dependency on tapes has motivated industrial efforts in technology improvements, resulting in much faster data density progression on tape rather than on disk. After recent breakthroughs in materials used, tape capacity is expected to witness a massive leap in coming years, increasing to several hundreds of TB per tape. This evolution will amplify the main benefits of tape storage.

However, tapes have often been primarily considered for archiving cold data, because of their main drawback: it takes around a minute to mount a tape from its shelf into a drive and position the reading head before starting reading data. This explains the current lack of academic effort to optimize relatively frequent data accesses. Nevertheless, more and more research projects require to handle tremendous volumes of data, which are not only destined to be archived but also regularly accessed for scientific analysis. Budget constraints impose the usage of tape storage, and optimizing tape data access therefore becomes more and more significant, and not limited to improving archive retrieval.

The general idea of the VHS project is to propose new interactions between resource management and tape systems. Using filesystem on tapes, we plan to design novel data placement strategies that will propose efficient data accesses by considering tapes at the level of the storage hierarchy by optimizing its operational cost. Our methodology starts from the tapes themselves, to better understand the physical processes involved in the different operations. Then, we will leverage this knowledge to derive interactions between tape and disk storage systems in order to improve data placement.

Evolution as compared to the first-round proposal

The current proposition is in line with the pre-demand on the research directions considered. However, we made some modifications in order to take into account the evaluations:

- We completed the related work in the "Positioning" section so that the positioning of the VHS project and the scientific objectives are more clearly stated.
- We provided more perspectives on the impact of the VHS project, especially future directions that would benefit from (1) a better knowledge of tape internal behavior and (2) a better integration of tape in the Hierarchical Storage Management (HSM) system.
- We added the cost of a 6-month research intern that will start working on the performance evaluation on tape system operations, that we did not count in the pre-demand. In addition, our institution increased the cost of a Ph.D. funding since last October. Overall, we adapted the requested resources to be in line with these requirements, with a 4% increase in comparison with the pre-demand (see Section "Requested Resources").

Summary of people involved in the project

Partner	Name	Position	Involvement	Role
ensIIE	Valentin Honoré	Associate professor	28 p.m (66 %)	Project leader: he will ensure that everything goes according to the plan. He will co-supervise the Ph.D student.
Télécom SudParis	François Trahay	Professor	2 p.m	He will help supervising the Ph.D student.
Télécom SudParis	*To be hired*	Ph.D. student funded by ANR	36 p.m	He/she will work on the performance evaluation and software interaction between the OS and the tape drive.
Télécom SudParis	*To be hired*	Intern funded by ANR	6 p.m	He/she will work on the performance evaluation of tape system operations.

Context, positioning and objectives

Magnetic tapes have been used to store computer data since the 1950s, so the layman now often considers it as an outdated technology. The name of the project, VHS, is a nod to the vintage K7 VHS, now totally outdated by the digital evolution.

However, tape storage is still and will remain essential in many fields. We can cite many actors making intensive usage of tapes in our close neighbourhood. The CERN is a major institution, that stores exabytes of data from high-energy physics projects over a wide network of data-centers, hence heavily relying on tapes. The Commissariat à l'Energie Atomique et aux Energies Alternatives(CEA) is also making an intensive usage of tapes for the storage of simulation data related to nuclear fusion. Finally, the European Centre for Medium-Range Weather Forecasts also store meteorological and climate data on tapes for daily computations.

All these actors use tape for the many advantages it offers: they do not consume energy when they are not used, offer a strong security (offline storage in shelves), and exhibit a better longevity for a much lower cost per TB than disks. This dependency on tapes has motivated industrial efforts in technology improvements, resulting in much faster data density progression on tape rather than on disk. After recent breakthroughs in materials used, tape capacity is expected to witness a massive leap in coming years, increasing to several hundreds of TB per tape. This evolution will amplify the main benefits of tape storage. Overall, tape is a reference storage medium that answers the strong constraints on sustainability for large-scale systems.

However, tapes have often been primarily considered for archiving cold data, because of their main drawback: it takes around a minute to mount a tape from its shelf into a drive and position the reading head before starting reading data. This explains the current lack of academic effort to optimize relatively frequent data accesses. Nevertheless, more and more research projects require to handle tremendous volumes of data, which are not only destined to be archived but also regularly accessed for scientific analysis. Budget constraints impose the usage of tape storage, and optimizing tape data access therefore becomes more and more significant, and not limited to improving archive retrieval. This denotes the central role of tapes in the present and future of storage.

The general idea of the VHS project is to propose new interactions between resource management and tape systems. Using filesystem on tapes, we plan to design novel data placement strategies that will propose efficient data accesses by considering tapes at the level of the storage hierarchy, and optimizing its operational cost. Our methodology starts from the tapes themselves, to better understand the physical processes involved in the different operations. Then, we will leverage this knowledge to derive interactions between tape and disk storage systems in order to improve data placement.

Positioning

Hierarchical Storage Management (HSM) is an old yet still central paradigm, used in HPC since the late 80s. It strongly evolves in the Exascale era as new challenges appear and new technologies become available. In particular, the storage stack changes from a tape-to-disk hierarchy to more complex hierarchies involving three to five levels, including different kinds of SSDs or rotative HDDs.

Those new hierarchies aim to benefit from advantages of each technology: storage capacity offered by tapes, which are slow, speed and high throughput offered by rather small SSDs. In this context, optimizing the throughput when accessing each piece of hardware is crucial. The VHS project aims to study the interactions between the tape drive's firmware and the tape drive support provided in the operating system. By doing this, IO operations may be smartly reorganized, leading to a better usage of the available IO bandwidth to tapes and an higher aggregated IO throughput on the overall HSM.

Contrarily to disks, very few works have focused on the tape technology itself, hence optimizing the data accesses at the tape level. Some studies [1] at CERN have focused on benchmarking timings when moving along the tape, to derive a time model and ordering schedule of operations. However, such work seems difficult to generalize in any hardware and software setup, and suffers lack of knowledge on hardware behavior. Other approaches [2] have studied hardware issues on tape, but are outdated. Some theoretical studies [3] have proposed algorithms to schedule reading operations on simple tape models, that are hardly adaptable in a real setup due to the complex and unknown mechanical behavior of tapes. To the best of our knowledge, no public dataset or performance evaluation of tape system have been made publicly available. On the theoretical side, different approaches [3–5] have proposed algorithms to schedule reading operations on simple tape models. In [3], we demonstrated that elemental objectives in scheduling problems are not trivial when applied on tapes. Tape vendors use undisclosed algorithms (Recommended Access Order) to manage the different operations on tapes. These algorithms are not publicly available, and based on the use of information from the firmware of the tapes that acts like a black-box. The VHS project does not strictly focus on algorithm development, but rather adopts a systems perspective by using the firmware of the tape to measure the performance of operations and optimize them.

Some initiatives have focused on tape storage, such as The Data Carousel Project [6] at CERN. Its objective is to stream tape data to disks so that they can be processed in a pipeline without waiting for all data to be transferred. However, the solutions are very specific to the data access pattern, and do not focus on improving the tape system performance itself. IO-SEA [7] was a EuroHPC funded project that also focuses on HSM paradigm. It aimed at building an I/O software stack for Exascale. IO-SEA intensively made use of tapes, interfacing it with classical disk filesystem. IO-SEA focused on the software stack itself, by introducing new paradigms such as ephemeral services, and made use of object store. IO-SEA used tapes as a brick of the hierarchy to derive a complete IO stack, but did not provide performance enhancement for tapes themself. VHS focuses on a smarter use of tapes, making it possible to introduce new directions for optimizing data accesses. Thus, IO-SEA would benefit from those optimizations by making a better use of capacitive but slow devices such as tapes.

Our ambition in the VHS project is to focus on the interactions between the operating system and the tape storage firmware itself, in order to be able to perform more efficient operations on tapes. To the best of our knowledge, this approach is novel, yet complementary to the aforementioned initiatives.

Magnetic tape is a very sophisticated storage, with a long-term roadmap [8], that did not receive the same attention as disk storage in the computing community.

Position in relation to national/European projects

As presented in Table 1, VHS relates to several French and European research projects.

The PEPR NumPEx is a French national initiative on Exascale computing, with strong constraints on the sustainability of the systems. The NumPEx project focuses on developing a full software stack to deploy supercomputers (from application to system software components). This project does not yet consider tape storage, and focuses on classical storage mediums that are used by applications (DRAM, disks, etc). The VHS project adopts a complementary positioning by focusing on tape storage, as disks cannot store all data on Exascale machines. We plan to use the application demonstrators of NumPEx in VHS as validation codes for our contributions.

The PEPR Cloud is an initiative that is part of the "National Cloud Acceleration Strategy", which wants to address the scientific and technological challenges in Cloud computing related to performance, security, and economic aspects, with a strong emphasis on digital sovereignty. Tape storage clearly falls in these aspects, with its operational properties : low energy consumption, longevity and offline storage. The STEEL project in the PEPR Cloud is dedicated to designing secure and efficient data storage and processing in Cloud infrastructure. The VHS project lies within the scope, and we expect to collaborate during the VHS project with members of this project.

IO-SEA [9] was a EuroHPC-funded project which also focused on HSM paradigm. It aimed at building an I/O software stack for Exascale machines. We discuss in details in the Section "Positioning" the relation with our project. VHS focuses on a smarter use of tapes, making it possible to introduce new tracks for optimizing data accesses. Thus, IO-SEA would benefit from those optimizations by making a better use of capacitive but slow devices such as tapes.

"HyperScaleway" is a joint project between Inria and Scaleway that intends to construct a federated software stack to perform distributed computation in a cloud environment. As a world-class cloud provider, Scaleway uses tapes for middle- to long-term storage. More specifically, the project focuses on data management using scheduling and replication for storage on tape, while the VHS project is interested in software interactions between the operating system and the tape drive.

Overall, there have been many academic efforts dedicated to mass storage for computing facilities. However, none of this project have focused on optimizing the tape storage level, and derive new software interaction at the tape drive level. We intend in the VHS project to fulfill this gap.

Project Name	Type	Description
PEPR NumPEx	ANR 2023-2029	The PEPR NumPEx is a project funded by the ANR, which focuses on developing a software stack to deploy supercomputers from applications to system components.
PEPR Cloud	ANR 2023-2029	The PEPR Cloud is an initiative which addresses the scientific and technological challenges in Cloud computing related to performance, security, and economic aspects, with a strong emphasis on digital sovereignty.
IO-SEA	Euro-HPC 2021-2024	IO-SEA [7] was a EuroHPC project that aimed at building an I/O software stack for exascale computing, based on hierarchical storage management and on-demand provisoning of storage services.
HyperScaleway	BPI Scaleway/Inria (to be started)	HyperScaleway is a joint project between Inria and Scaleway. This project intends to construct a federated software stack to perform distributed computation in a cloud environment

Table 1: Related French and European projects

Objectives

The objective of the VHS project is to increase IO throughput at the tape level, so that performance of the global Hierarchical Storage Management system is improved.

The general idea of the VHS project is to propose new interactions between resource management and tape systems. Using filesystem on tapes such as LTFS [10], we plan to design novel data placement strategies that will propose efficient data accesses by optimizing its operational cost. Such strategies will allow the operating system to decide **when to make use of tapes** and **at what cost**, using **efficient operations** on tapes, thus enhancing the overall performance of the tape storage. In this objective, we propose a bi-directional yet complementary approach. The former investigates performance improvements at the tape level itself, the second at the level of the interactions between the tape system and the HSM.

1. Improving operations on tape based on firmware operations The first direction of the VHS project is to focus on the tape component itself, to find lever in order to optimize data access operations. Contrary to disks, very few works in the literature have studied the impact of the mechanical processes on the performance of the read/write operations. We plan to investigate if optimizations on the internal processes of tape can be done to improve performance of data access on tapes, using the functionalities of the firmware. The literature lacks understanding of internal tape behavior, hence limiting the possibilities to use the firmware functionalities to perform local optimizations for data accesses. To achieve this step, we plan to perform extensive benchmarking on high performance tape drives, associated with fine-grain performance evaluation based on extracted performance metrics and firmware interactions. Partners of the consortium managing large tape storage systems will be an asset to achieve these objectives. One of the main scientific challenges in this direction is to evaluate the flexibility we can obtain at the firmware level. We must evaluate the action levers with combining and/or reorder mechanical processes in the tape system in order to optimize data accesses on tapes. We present in the task description section below our roadmap to achieve the objectives of this axis.

2. Improving tape integration in HSM software While our first direction primary focus is tape internal behavior, we want here to study the piece of software that connects tape drive's firmware and the operating system. This software is the central piece in our objective of improving I/O throughput of the overall HSM. We want to obtain permissiveness with the tape level of the HSM, by leveraging knowledge of both I/O requirements from the operating system. We envision in that direction to derive **efficient requests to the type** from the software connecting it to the HSM. For instance, prioritizing operations on a tape that is unwound at a given point should perform better than unwinding the tape for each of those operations. For writing data on tapes, the problem of determining how to pack data on tapes based on properties (user, project, etc) in order to improve future reading operations is also challenging. Solutions to those aspects require both characterization of efficient operations on tape, but also a view on the I/O requests in it, given by the operating system. This direction of the project will have strong interactions with the resource management system.

Methodology and risks management

Methodology

In the VHS project, we target fully open-source solutions. As we expect to impact not only academic actors but also other tape users, open-source solutions are at the core of the proposal of this project. The Commissariat à l'Energie Atomique et aux Energies Alternatives has dedicated many efforts in the development of open-source software dedicated to tape storage systems. We plan to rely on Phobos [11], a tool designed to manage large volumes of data for various kinds of storage technologies from disks to tapes. Phobos uses the LTFS filesystem [10] to interact with the tapes. We plan to integrate our software solutions with Phobos as it is already capable of performing operations both on disks and tapes. The expertise of the partners from the CEA in the development of Phobos will be a strong asset to achieve this objective.

Performance evaluation on a production tape storage system is challenging, due to security restrictions and limited action range. We intend to first develop and benchmark solutions on a local setup including hardware for which we ask funding in this project. Even though the setup does not benefit from the scale of a production system, the hardware and software environment will be very similar to the one we could find in the installations of partners such as CEA. Hence, it will be sufficient to achieve the objectives of this project and demonstrate the practicality of our solutions.

Risk management

The project has two main risks, inherent to any exploratory project such as the VHS project.

The first one concerns the understanding and performance assessment of the tape driver. We do not know in advance what will be the latitude given by the tape firmware to perform fine-grained performance analysis of tape operations. Identifying the levers for action at the controller level will be a crucial step in the design of a performance evaluation toolkit for tape. To mitigate this risk, we will rely on a software environment mastered by our partner from the CEA, so that we limit the risk of blockage in the setup of the experimental environment, and maximize the number of options we will explore in order to derive an effective performance evaluation. To the best of our knowledge, no such studies have been performed in the literature, mostly because tape storage systems host very sensitive data. Thus, our contributions in this direction will be of interest for all the community, and all the data will be made publicly available.

The second risk is that our strategy to better integrate the tape layer in the HSM depends on the results of the performance evaluation process, which deliverable will be more new firmware APIs for more efficient operations on tape. We plan to use these operations to design a new piece of software interfacing the operating system and the tape drive, which would be able to make a more efficient use of tapes by leveraging knowledge on both the state of the overall HSM and the state of the tape drive. We will mitigate this dependency by not relying only on theses APIs in the design of the piece software. We plan to identify key characteristics of data management in the HSM for which the tape layer could enhance the data throughput. The firmware APIs will be one of the lever of action at the tape level itself, but not a restricted to this one only.

Overview

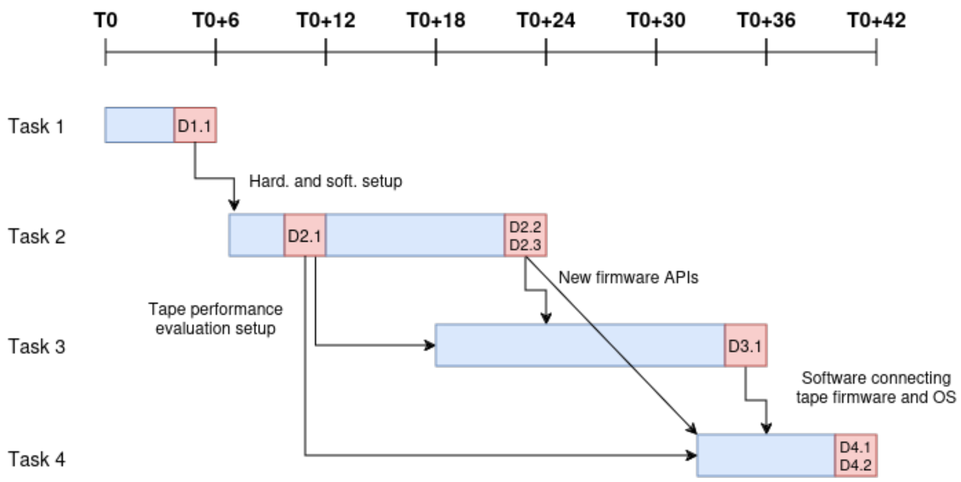

Figure 1: Gantt chart of the project

As described in the Gantt chart in Figure 1, the VHS project is structured in 4 main tasks:

- **Task 1** will setup the hardware for which we ask a funding in the project (an high-performance tape drive with tapes). In this task, we will plug the driver to a work station, inside our local cluster, and setup the associated software layer with the support of Philippe DENIEL, our collaborator from the CEA.

- **Task 2** will use the setup and environment of Task 1 to perform a performance analysis of the various tape operations using the tape drive. We will use metrics extracted from the controller and firmware. We will thus analyze the extracted data from these measurements in order to extract takeaways and insights on the internal tape behavior. Using these results, we will then explore how tape operations can be optimized using the tape firmware (operation reordering for instance). The objective is to design a new API that would optimize tape operations at the drive level.

- **Task 3** will leverage the results of Task 2, this task will design new software interactions between the operating system and the tape drive. The objective is to take into account the I/O requests in the HSM software to optimize the requests on the tape storage system, using the APIs from Task 2. This will be done by defining optimal data structures for tape storage, and reordering operations on tape. We plan to modify the software connecting the tape firmware and the operating system. At this level, we will be able to take into account both the state of the tape storage system and the data pressure on the HSM.

- **Task 4** aims at evaluating the software prototypes and data structures developed in Tasks 3, using the experimental setup of Task 2. We also plan to rely on real use-cases that we plan to get from collaborations (CEA or private partners) or related projects (PEPR NumPEx, Exascale Computing Project). We will make all the software and experiments reproducible and fully open-source to facilitate the dissemination of the results.

Description of the Tasks

Task 1: Software and hardware setup

Leader: Valentin Honoré		
Involved partners	Valentin Honoré	Intern
Involvement	2	3
Input: Hardware funding of the VHS project		
Output: A stable installation of the tape drive in the local cluster of the BENAGIL team at Télécom SudParis		
Deadline	**Deliverable**	
T0+3	D1.1: Hardware acquisition and pre-deployment for software installation	
T0+6	D1.2: Integration of the tape drive in the local cluster at Télécom SudParis	

Goal: To setup all the hardware and software environment of the VHS project.

Detailed work program: In the VHS project, we aim to perform a fine-grain study of tape storage. In this objective, we need to acquire an adapted hardware, and integrate it in our local cluster in Télécom SudParis. We ask in the project the funding for an high-performance tape drive with compatible tapes. We cannot access the infrastructures of our partner from CEA due to security restrictions, hence the need to acquire our own hardware.

The goal of Task 1 will be to order and setup this hardware to offer the proper environment for the remaining tasks. We will choose a tape drive, and install the related software in consultation with our partner from CEA, Philippe Deniel, who is an expert in mass storage system for HPC.

We will first place the order for the hardware, in close collaboration with our partner from CEA, Philippe Deniel. We already estimated the cost of this hardware to be around 33 000€. While waiting for the hardware to be available, we will prepare the software installation and integration of the tape drive with a workstation inside the local cluster of Télécom SudParis. The cost of the workstation is included in the hardware budget.

Once the setup installed, we will design some preliminary tests using simple benchmarking to validate the installation and get some basic reference values (performances etc). We have experience in maintaining and upgrading the cluster, that we manage internally in the team. With the experience of Philippe Deniel, we do not expect major problems in the software and hardware installation of the funded material.

Delivrables:

T0+3 [D1.1]: Hardware purchase and pre-deployment for software installation

T0+6 [D1.2]: A complete integration of the hardware in our cluster, with a proper software layer.

Task 2: Performance evaluation of tape storage system

Leader: Valentin HONORÉ				
Involved partners	Valentin HONORÉ	François TRAHAY	Ph.D. student	Intern
Involvement	12	1	15	3
Input: Environment setup from Task 1				
Output: Performance evaluation environment for tape storage, APIs of the tape firmware for efficient operations				
Start: T_0+6		**End:** T_0+24		

Deadline	**Deliverable**
T0+12	D2.1: Performance evaluation setup of tape firmware: datasets and data analysis.
T0+24	D2.2: Firmware APIs for efficient tape operations
T0+24	D2.3: A mid-term report on the progress of the project, with an expense summary

Goal: To evaluate the performance of tape operations to propose new APIs for more efficient operations

Detailed work program: One of the main scientific objective of the VHS project is to better understand the low-level interaction in-between the tape drive, the tape cartridge and the connected device. The approach that we propose in this task is to measure the operational cost of tape operations at a fine-grain level. To the best of our knowledge, no such previous approach has been publicly made available, contrary to disks on which many studies have been made for mechanical and SSD hard drives for several decades already [12, 13].

We will determine the most impacting features regarding operation latency and global throughput, and then leverage this knowledge to propose a new API for efficient tape operations at the firmware level.

Task 2.1: Performance evaluation of tape operations

Goal: To perform fine-grain performance evaluation of tape operations

Detailed work program: In order to enhance the performance of tape operations, we first need to be able to determine which features impact the performance of these operations.

This task is dedicated to the extraction of metrics and performance data resulting of the different operations done within the tape drive. We are interested here in correlating the mechanic behavior involving the tape cartridge (reading head acceleration, throughput of reading/writing etc...) while performing tape operations.

Our roadmap is as follows :

* Introduce performance counters and metrics inside the tape firmware and the related software controlling the tape drive. This could be done by taking an approach similar to implementing SMART [14] counters in HDDs.
* Run the appropriate benchmarking set of operations to generate exhaustive datasets of performance metrics. This benchmark will probably be involving synthetic data in order to cover a wide range of operation parameters (short or long read/write etc...).
* Perform a fine-grain analysis of the extracted performance datasets by determining the main features to be considered for the performances of the different operations on tape. We will publicly release in a mid-term project report (Deliverable 2.3) the benchmark toolkit, including full sources in a open-source repository.

Delivrables:

T0+12 [D2.1]: A public report of the performance evaluation of tape operation, with takeaways and guidelines for future levers of action.

Task 2.2 : Designing tape drive firmware APIs

Goal: To propose an API for more efficient tape operations

Detailed work program: The performance analysis from Task 2.1 will determine which feature and criteria impact the performance of the different tape operations.

In this task, we intend to define new APIs of tape operations at various level, involving the embedded software in the tape drive and the fiber channel adapter, connecting the drive to the work station, as well as the lowest level of the operating system. Those API will leverage knowledge of the mechanical processes involved in tape operations in order to provide more efficient tape operations.

Designing such APIs will require to determine the flexibility offered by the tape firmware. For instance, we may reorder operations, or substitute an operation by a smarter one. We will then evaluate the gain provided by those new APIs, by reusing the performance evaluation toolkit developed in Task 2.1.

In order to achieve the roadmap for Task 2.1 and 2.2, we will collaborate with Bertrand Simon, an expert in algorithm design and data structures. Bertrand Simon will help us in the analysis of the data, and design the data structure that will be used as the basic brick for the firmware API of Task 2.2. We will also deeply involve Philippe Deniel, who will guide us for identifying the obstacles on the road.

Delivrables:

T0+24 [D2.2]: Firmware APIs for efficient operations on tape, available on an open-source repository with the associated documentation.

T0+24 [D2.3]: A mid-term report on the progress of the project, with an expense summary

Risks and backup solutions: There are two main risks associated to this task.

The first risk addresses the performance metrics at the firmware level. We estimate that the risk is low as we have in the consortium an expert in operating tape storage system, Philippe Deniel. Even though we may have to move up to an upper software layer to make measurements, we should be able to obtain enough data to extract insights and takeaways on the performance of tape operations. Even though the firmware are black-boxes, we have no access to the sources. Their behavior is to be analyzed and understood, during the processing of all tasks.

The second risk concerns the flexibility that will be offered by the firmware to optimize the operations on our features of interest. If we do not obtain enough permissiveness, we plan here again to change the software layer, and act at the level of the software connecting with the firmware.

Task 3: Improvement of tape storage in HSM

Leader: Valentin HONORÉ			
Involved partners	Valentin HONORÉ	François TRAHAY	Ph.D. student
Involvement	11	1	16
Input: The firmware API (from Task 2)			
Output: A piece of software for efficient tape integration in the operating system			
Start: T_0+18		**End:** T_0+36	

Deadline	Deliverable
T0+36	D3.1: A piece of software smartly connecting tape drive's firmware and operating system

Goal: A new piece of software connecting the tape firmware and the operating system.

Detailed work program: In this task, we improve the integration of the tape storage layer in the Hierarchical Storage Management system (HSM). In this scope, we will combine and improve both operating system's and tape device's operations in order to optimize the overall data throughput.

This task designs a new piece of software able to proactively optimize and reorder the requests to the tape device. This piece of software will leverage information from the upper application layers as well as input from the operating system. We envision different approaches for this piece of software: packing data on an application-based, pre-fetching of data, data-buffering and write-behind etc.

As part of this task, we will design new data structures to optimize the storage of I/O data on tapes, as well as identifying points of interest to be extracted from the operating system. Task 4 will validate the outcomes of this piece of software on user applications. Task 3 will focus on synthetic data with an approach similar to Task 2.1.

This task involves two partners, Bertrand SIMON and Philippe DENIEL. Bertrand SIMON will bring his expertise on I/O scheduling and data structure to extract important features for I/O performance at the system level. Philippe DENIEL will guide us in the design of the software interactions between the tape system and the operating system.

Delivrables:

T0+30 [D3.1]: A piece of software smartly interfacing tape drive's firmware and the operating system with concrete optimizations, available on an open-source repository with documentation.

Risks and backup solutions: The main risk of this task lies in the design of the piece of software. The previously designed software may address only some use-cases and not others. In this situation, redesign operations should be considered, as well as data-specific management strategies. Keeping the proposed solutions as generic as possible is critical. In case the software only works for specific patterns, we will investigate the key reasons affecting the performances, then propose new solutions.

Task 4: Experimental evaluation and dissemination

Leader: Valentin HONORÉ		
Involved partners	Valentin HONORÉ	Ph.D. student
Involvement	3	5
Input: Performance evaluation environment (from Task 2), piece of software between the operating system and the tape drive (from Task 3)		
Output: A stable release of the piece of software from Task 3		
Start: T_0	**End:** T_0+42	

Deadline	**Deliverable**
T0+42	D4.1: Evaluation of Task 3 on real data and setup
T0+42	D4.2: A website with a report on all the work of the project : software production, extensive documentation, reproducibility artifacts, public datasets of tape metrics
T0+42	D4.3: A final report on the project

Task4.1: Large-scale experimental evaluation

Goal: To test the results of Task 3 on actual applications.

Detailed work program: This task aims at evaluating the developed piece of software on real-case applications, running on our local cluster. As part of the development of the performance analysis setup in Task 2.1, the firmware API in Task 2.2 and the piece of software in Task 3, we will evaluate our implementation on dedicated datasets and applications. We will experiment with proxy applications for Exascale computers. For instance, we plan to rely on the developed demonstrators as part of the NumPEx project, or as part of the Exascale Computing Project.

In this Task, we plan to evaluate our implementation on real applications running on larger-scale machines. This evaluation aims at assessing the scalability of our software when the I/O pressure increases in charge: is the data throughput enhanced in the global HSM? Is the performance of the tape storage layer improved with the new API?

Delivrables:

T0+42 [D4.1]: Evaluation of the piece of software of Task 3 on real data and applications.

Task4.2: Project valorization

Goal: To finalize the software and reproducibility artifacts so that they can be disseminated

Detailed work program: In this task, we will set up facilities so that the software developed in VHS can be used by the research community. An effort will be put in stabilizing the source code of the produced software alongside its dependencies, that we will fully document and make reproducible.

We will also create a repository containing all the datasets and performance metrics related to tape operations, gathered during the VHS project. To the best of our knowledge, such datasets do not exist in the public domain. The objective of this repository is to allow other researchers to use these datasets, either for theoretical studies (scheduling problems related to tape), or to propose new optimization strategies regarding the tape operations.

Delivrables:

T0+42 [D4.2]: A website with all software production, documentation, performance datasets and reproducibility artifacts.

T0+42 [D4.3]: A final report on the project.

Risks and backup solutions: There is a risk that we cannot experiment our software on actual applications due to either hardware or software limitations in our cluster. In this case, we will use mini-apps developed for evaluating the I/O subsystem such as the one from the Virtual Institute for I/O [1]. We could also use the demonstrators that will be made available by the PEPR NumPEx. Those codes will be recent and supported, hence a good candidate for a real-application. Finally, we also plan to rely on the expertise of our partner from CEA to guide us through the choice of the right demonstrator code to be chosen.

Task schedule, deliverable, and milestones

We sum up the Task schedule graphically in Table 2. A synthetic view of the deliverables is presented in Table 3.

		Year 1				Year 2				Year 3				Year 4	
		3	6	9	12	15	18	21	24	27	30	33	36	39	42
Task 1			✪												
Task 2	Subtask 2.1						✍								
	Subtask 2.2								✪						
Task 3													✪		
Task 4	Subtask 4.1														✍
	Subtask 4.2														✪
Progress report/Expenses									☆						☆

Legend:
✪ Deliverable is a software ✍ Deliverable is a report ☆ Progress report & expense summary

Table 2: Task schedule for VHS

Task	Title and substance of the deliverable and milestone	Delivery
	T1: Software and hardware setup	
D1.1	Hardware acquisition and pre-deployment for software installation	T0+6
D1.2	Integration of the tape drive in the local cluster at Télécom SudParis	T0+6
	T2: Performance evaluation of tape storage system	
T2.1: Designing tape drive firmware APIs		
D2.1	A public report of the performance evaluation of tape operations	T0+18
T2.2: Designing a firmware API		
D2.2	A new firmware API that offers more efficient operations on tape	T0+24
D2.3	A mid-term report on the progress of the project, with an expense summary	T0+24
	T3: Improvement of tape storage in HSM	
D3.1	Piece of software connecting tape drive's firmware and operating system	T0+36
	T4: Experimental evaluation and dissemination	
T4.1: Large-scale experimental evaluation		
D4.1	Evaluation of Tasks 2 and 3 on real application (report)	T0+42
T4.2: Project valorization		
D4.2	A website with all the software production,reproducibility artifacts and public datasets of tape metrics	T0+42
D4.3	A final report on the project	T0+42

Table 3: Deliverables and Milestones for VHS

[1] https://www.vi4io.org/

Requested resources

Cost supported by ANR proposal

Staff One intern (master-level, 6 months) will be hired for the project at T0+3. This internship will focus on setting up the software environment for the tape drive, and on the development of the performance evaluation environment of the tape drive.

We will also hire a Ph.D student (36 months) that will start at T0+6, as a continuation of the work of the internship. The Ph.D student will thus work on the design and implementation of the performance analysis suite, firmware APIS as well as on the design of the new piece of software that targets a better integration of tape storage in the HSM.

Both the intern and the Ph.D student will be closely supervised by the permanent researchers involved in the project (see "Permanent staff"). In addition, we will closely collaborate with Philippe Deniel, an expert in mass storage systems from CEA and Bertrand Simon from CC-IN2P3, an expert in algorithm sand data structures.

Mission, traveling, and workshops The traveling costs are evaluated as follows:

- Attending one international conference for one member of the project during 3 years. *Cost: 3 x 3 000€= 9 000€*
- Attending one European conference for one member of the project during 3 years. *Cost: 3 x 2 000€= 6 000€*

Finally, we plan to organize one-day workshops at different milestones of the project, at T0+18 and T0+36. These workshops will contribute to the dissemination of the results in the VHS project, and gather researchers and industrial partners operating tape systems in order to trigger collaborations. The cost (food and breaks) will be partly supported by ANR : 3000€.

Hardware & Inward billing Two laptops will be bought for members of the project for a cost of *2 x 1 000€*. In order to run the performance experiments on a real tape drive and develop software related to tape storage, we will buy a high-performance tape drive with tapes. According to a quotation, such hardware would cost approximately 33 000€ with a PC station.

Cost not supported by ANR proposal

Permanent staff Valentin Honoré, the principal investor of VHS, will spend 66 % of his time (28 p.m) working on the project. In addition to supervising the Ph.D student and the intern, he will participate in the design and development of the different software contributions of the project.

Prof. François Trahay will spend 2 p.m working on the project. He will help supervising the Ph.D student, and will guide the design of the software to be developed in the project.

Summary of the project cost

	Cost	Requested funding
Permanent staff / *Frais de personnel permanents*	219 900€	0
Staff / *Frais de personnel*	146 000€	146 000€
Hardware / *Coûts des instruments et du matériel*	35 000€	35 000€
Travel / *Missions*	18 000€	18 000€
Administrative fees / *Frais d'environnement*	26 865€	26 865€
Sub-total / *Sous-total*	**418 900€**	**225 865€**
Requested funding / *Aide demandée*	-	**225 865€**

Cost difference with pre-demand

In the pre-demand of the VHS project, we had a total cost of 216 785€ including a Ph.D student and budgets from travelling and hardware.
Since the submission in October 2024, the cost of a Ph.D student in Télécom SudParis increased from 138 000€ to 142 000€. Thus, we have an overhead of 4 000€ on the cost of the project's Ph.D. funding. In addition, we did not count the cost of the 6 month-internship, that sums up to 4 000€.
In total, we added 8 000€ to the VHS financial support in this second proposal. With the overhead of 13.5%, we obtain an additional amount of 9 080€.
Thus, the requested financial support for the VHS project sums up to 216785€+9080€= 225865€. The difference between the initial and updated budget is a 4% increase.

Consortium around VHS

Scientific leader

The Principal Investigator of the VHS project is Valentin Honoré. He will devote sixty-six percent of his research time to the project. Valentin Honoré is an associate professor at École Nationale Supérieure d'Informatique pour l'Industrie et les Entreprises (ensIIE) since 2022. He is a member of the BENAGIL team of the SAMOVAR laboratory and Inria. He received his Ph.D. degree in computer science from the University of Bordeaux in 2020. From November 2020 to June 2022, he was a post-doctoral fellow at the IN2P3 Computing Center in Villeurbanne (CNRS). His research interests focused on scheduling algorithms and performance models for distributed applications. Since his arrival in the BENAGIL team, his research activities now mostly focus on performance analysis for HPC applications and systems, and on tape storage. He has been working on tape storage since 2021, focusing on algorithmic foundations of scheduling related to tapes [3]. This project intends to move one step further in his research activities dedicating to the optimization of tape storage, by providing new software interactions between the tape drive and the operating system. He is the co-author of 8 research papers in top international conferences and journals such as IEEE TPDS, IPDPS or ICAPS.

Scientific environment

In addition to the principal investigator, the project will involve other members of the BENAGIL team. BENAGIL is composed of experts in runtime systems for high-performance computing (Dr. Élisabeth Brunet, Pr. François Trahay), and in operating systems (Prof. Gaël Thomas, Dr. Mathieu Bacou). Pr. François Trahay will help supervising the Ph.D student hired for this project. He is an expert in performance evaluation for distributed system. He is the leader of the eztrace software [15], and is already working with the principal investigator on the Pallas trace format [16].

VHS will also involve Dr. Philippe Deniel from the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA). He is an expert in the management and deployment of large-scale tape storage system. He was the head of the Storage Team of the CEA-DAM from 2008 to 2015. He was then the head of the laboratory "Stockage HPC et stations scientifiques" in the same place. Since 2021, he is an "Fellow Expert" for CEA, in charge of steering the research activities on storage and quantum computing, and the integration of quantum machines in the infrastructure of the TGCC computing center. Dr. Philippe Deniel was the scientific leader of the IO-SEA project (cf "Related Projects). IO-SEA intensively made use of tapes, interfacing it with classical disk filesystem. The expertise of Dr. Philippe Deniel on the management of mass storage systems will be a central asset for the VHS project. He will collaborate with us at every step of the project.
Finally, we also plan to collaborate with Dr. Bertrand Simon from the IN2P3 Computing Center (CC-IN2P3, CNRS). He will bring his expertise on algorithm design and data structures, that will be useful to complete Tasks 2 and 3 of the VHS project.

Impact of the proposal

In the short term, we expect that the VHS project leads to a better usage of tape storage by refining the interactions between Hierarchical Storage Management layers and tape system. Tape technologies and some disk technologies tend to converge in some ways. Servo-tracks on IBM Jaguar Tapes and later in LTO is an example [17]. Optimisations rely on a fine knowledge about disk behavior from the firmware, the operating system and some applications. We believe a similar approach is to be used with tapes, with benefits for all communities using this technology. With the important pressure on HSM, improvements at one level of the hierarchy is beneficial to all levels of the HSM.

In the long term, we expect that the results of the VHS project will infuse to other domains such as Cloud computing. Indeed, storage on tapes is the de-facto standard for such industries, that have to handle gigantic amounts of data. Some other projects such as HyperScaleway are considering data management for long-term storage on tapes. We hope that our VHS project will propose solutions that could leverage the optimization of the tape storage layer, not only restricting to long-term storage.

We also expect that, in the long term, the VHS project will allow new functionalities in terms of software interactions between the operating system and the tape controller, so that data usually handled on disks could be used on tape at a much cheaper cost. One example of possible interesting use case is Key-Value Store (KVS). KVS is a type of NoSQL database that stores data as key-value pairs, where each unique key is associated with a specific value. This model offers high performance, scalability, and flexibility, making it widely used in distributed storage for real applications. Modern key-value stores leverage in-memory processing, using advanced indexing techniques to reduce latency and increase throughput. Recent works enable distributed KVS that work efficiently across cloud and edge environments. The current key-value stores are designed for today's typical hardware environment with a memory hierarchy of slow, but rather capacitive hard disks, and in-memory technologies, using dedicated optimized data structures such as Log Structured Merge (LSM) trees [18,19]. Currently, no key-value store can leverage the high capacity of tape, even though such large volume of data would be very suited for tape storage, which tends to favor large batches of operations. Using the better tape support we plan to provide with the VHS project, we could study how those disk-specific approaches in KVS storage could be adapted on tape. KVS would then be able to benefit from the large capacity of tape at a much lower cost than a similar approach using disks.

Positionning with respect to the call

The VHS project addresses the "Axe E.03 (CE25) : Sciences et génie du logiciel - Réseaux de communication multi-usages, infrastructures numériques" challenge. Indeed, we intend to increase the performance of tape storage system that are widely used in computation-intensive facilities. Such objectives will address the need of sustainability of hardware resources and a dynamic management of data at scale. In addition, the VHS project also addresses several "*Technical Research Priorities*" from the 2022 ETP4HPC Strategic Research Agenda [20]. In particular, VHS lies in the priority #3 *System Software and Management*, and priority #5 *IO and Storage* [21] where tape storage is clearly stated as an unavoidable storage technology associated with many challenges.

Recently, large-scale initiatives on Exascale Computing such as the PEPR NumPEx have been initiated in Europe, with strong constraints on the sustainability of such systems. The NumPEx project focuses on the full software stack to deploy supercomputers. This project does not yet consider tapes, and focuses on classical storage medium that are used by applications (DRAM, disks etc). We believe that the VHS project is complementary to this initiative, as disks are too costly to store all data from Exascale machines. Thus, we hope this project will trigger or enforce collaborations with the computing communities to provide more efficient storage systems with tapes.

Valorization strategies, dissemination, and exploitation of results

Software The investigator of the project will continue his commitment to open-source as all the performance evaluation tool and software developed during the project will be freely available as open-source software. Besides, the different contributions will be maintained beyond the end of the project, and efforts will be made to bootstrap a visible and active community of users and contributors, through appropriate channels and organizations. In this objective, we plan to organize workshops and meetings during the project to gather people working on tape storage management and related software development such as CEA, CERN or private actors (storage and computing providers for instance). These workshops may be hold in the framework of the PER3S workshop, an annual meeting of the french community of storage with both academic and private actors.
In addition to software, we also plan to make our research results (datasets of metrics, related analysis etc) freely available (see Section "Reproducibility").

Collaboration We expect that the outcomes of the project will trigger collaborations or enforce existing ones. In particular, VHS is expected to have direct exchanges with the storage team of the CEA throughout the collaboration with Dr. Philippe DENIEL. The expertise of the CEA will bring a strong expertise in tape storage management, that will help the project to fulfill its objectives.

Moreover, the BENAGIL team of the project investigator is a part of a joint project between Inria and Scaleway, named "HyperScaleway". This project intends to construct a federated software stack to perform distributed computation in a cloud environment. As a world-class cloud provider, Scaleway uses tapes for middle- to long-term storage. The BENAGIL team is associated with this project through the funding of a Ph.D thesis, focusing on scheduling and replication for storage on tape. We envision this collaboration as a great opportunity to study data access patterns on tapes alongside associated system logs and user's production code. Hence, we expect this project to trigger collaborations with the VHS project, and generate strong interactions with Scaleway with complementary contributions.

Publication During the project, we intend to publish the scientific results. These results will be published in major international conferences (such as IPDPS, EuroPar, ICPP, ASPLOS, SOSP, etc.) and journals (TPDS, IJPP, etc.) We also intend to disseminate scientific results to the french community as part of the Compas conference.

Reproducibility

The ability to reproduce research results is a growing concern. VHS will contribute to this effort in several ways:

- All the software developed as part of this ANR project will be freely available and open-source, with their associated documentation ;
- We will create a data repository where experiment results will be publicly available (information on the software dependencies, experimental setup etc...). We will also release all the generated datasets concerning performance metrics of the tape drive.

We believe that such data and source repositories will help the research community reproduce research results, and enforce the confidence in the results that will be produced in the scope of the VHS project.

References

[1] Germán Cancio Meliá, "LTO performance - Make Tape Reading Great Again," CERN IT/ST, Tech. Rep., 2018, [Online]. Available: Presentation slides.

[2] Irwin, J. W. and Cassie, J. V. and Oppeboen, H. C., "The IBM 3803/3420 Magnetic Tape Subsystem," *IBM Journal of Research and Development*, vol. 15, no. 5, pp. 391–400, 1971. [Online]. Available: https://doi.org/10.1147/rd.155.0391

[3] V. Honoré, B. Simon, and F. Suter, "An exact algorithm for the linear tape scheduling problem," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 32, 2022, pp. 151–159. [Online]. Available: https://doi.org/10.1609/icaps.v32i1.19796

[4] C. Cardonha and L. V. Real, "Online algorithms for the linear tape scheduling problem," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 26, 2016, pp. 70–78. [Online]. Available: https://doi.org/10.1609/icaps.v26i1.13743

[5] C. H. Cardonha, A. A. Cire, and L. C. V. Real, "Design and analysis of efficient sequencing policies for linear cold storage devices," *PRODUCTION AND OPERATIONS MANAGEMENT*, 2025.

[6] "Data Carousel Project," https://npps.bnl.gov/software/data-carousel.html, Accessed: 2023-10-12.

[7] D. Medeiros, E. B. Gregory, P. Couvee, J. Hawkes, S. Gougeaud, M. Gilliot, O. Bressand, Y. Valeri, J. Jaeger, D. Chapon, F. Bournaud, L. Strafella, D. Caviedes-Voullième, G. Tashakor, J. Zjupa, M. Holicki, T. Ridley, Y. Müller, F. S. M. Guimarães, W. Frings, J.-O. Mirus, I. Zhukov, E. R. Borba, N. Moti, R. Salkhordeh, N. Derbey, S. Mimouni, S. Derr, B. B. Gursoy, J. Grogan, R. Furmánek, M. Golasowski, K. Slaninová, J. Martinovič, J. Faltýnek, J. Wong, M. Cakircali, T. Quintino, S. Smart, O. Iffrig, S. Narasimhamurthy, S. Happ, M. Rauh, S. Krempel, M. Wiggins, J. Nováček, A. Brinkmann, S. Markidis, and P. Deniel, "IO-SEA: Storage I/O and Data Management for Exascale Architectures," in *Proceedings of the 21st ACM International Conference on Computing Frontiers: Workshops and Special Sessions*, ser. CF '24 Companion. New York, NY, USA: Association for Computing Machinery, 2024, p. 94–100. [Online]. Available: https://doi.org/10.1145/3637543.3654620

[8] "Linear Tape Open (LTO) Ultrium Roadmap," https://www.lto.org/roadmap/, Accessed: 2023-10-12.

[9] "The IO-SEA Project," https://iosea-project.eu/, Accessed: 2023-10-12.

[10] T. Chakravarty, "LTFS Transforms LTO Tape into Nearline Storage: Accelerating 4K Media Workflows," in *SMPTE 2014 Annual Technical Conference & Exhibition*, 2014, pp. 1–7. [Online]. Available: https://doi.org/10.5594/M001577

[11] "Parallel Heterogeneous OBject Store," https://github.com/cea-hpc/phobos, Accessed: 2023-10-12.

[12] M. Seltzer, P. Chen, and J. Ousterhout, "Disk scheduling revisited," in *Proceedings of the winter 1990 USENIX technical conference*. Washington, DC, 1990, pp. 313–323.

[13] T. J. Teorey and T. B. Pinkerton, "A comparative analysis of disk scheduling policies," *Commun. ACM*, vol. 15, no. 3, p. 177–184, Mar. 1972. [Online]. Available: https://doi.org/10.1145/361268.361278

[14] T. Pitakrat, A. Van Hoorn, and L. Grunske, "A comparison of machine learning algorithms for proactive hard disk drive failure detection," in *Proceedings of the 4th international ACM Sigsoft symposium on Architecting critical systems*, 2013, pp. 1–10.

[15] F. Trahay, F. Rue, M. Faverge, Y. Ishikawa, R. Namyst, and J. Dongarra, "Eztrace: a generic framework for performance analysis," in *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 2011, pp. 618–619.

[16] C. Guelque, V. Honoré, P. Swartvagher, G. Thomas, and F. Trahay, "Pallas: a generic trace format for large hpc trace analysis," in *IPDPS 2025-39th IEEE International Parallel & Distributed Processing Symposium*, 2025.

[17] B. Haeusser, E. Winters, and E. Zwemmer, *IBM System Storage Tape Library Guide for Open Systems*. IBM, International Technical Support Organization, 2006. [Online]. Available: https://www.redbooks.ibm.com/redbooks/pdfs/sg245946.pdf

[18] P. O'Neil, E. Cheng, D. Gawlick, and E. O'Neil, "The log-structured merge-tree (lsm-tree)," *Acta Informatica*, vol. 33, pp. 351–385, 1996.

[19] Y. Dai, Y. Xu, A. Ganesan, R. Alagappan, B. Kroth, A. Arpaci-Dusseau, and R. Arpaci-Dusseau, "From {WiscKey} to bourbon: A learned index for {Log-Structured} merge trees," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, 2020, pp. 155–171.

[20] S. Narasimhamurthy, N. Mittenzwey, F. Magugliani, M. Duranton, C. Prunty, P. Rossé-Laurent, M. Marazakis, P. Carpenter, G. Antoniu, S. Neuwirth, P. Deniel, D. Pleiter, U.-U. Haus, E. Laure, A. Wierse, T. Becker, R. Haas, M. Malms, H.-C. Hoppe, V. Bartsch, S. Dolas, O. Vysocky, M. Perez, A. Forrester, K. Michielsen, and E. Suarez, *ETP4HPC's SRA 6 - Strategic Research Agenda for High Performance Computing in Europe*, ser. ETP4HPC Strategic Research Agenda. Zenodo, 2024. [Online]. Available: https://hal.science/hal-04846393

[21] S. Neuwirth, P. Deniel, J.-T. Acquaviva, M. Golasowski, M. Hennecke, A. Jackson, T. Leibovici, J. Luettgau, and R. Nou, "ETP4HPC SRA 6 White Paper - I/O and Storage," Jan. 2025, this is a white paper released as part of ETP4HPC's Strategic Research Agenda 6. [Online]. Available: https://hal.science/hal-04905038