The Case For Data Centre Hyperloops

Guillem López-Paradís*†[1] Isaac M. Hair‡[1] Sid Kannan ‡ Roman Rabbat ‡
Parker Murray ‡ Alex Lopes ‡ Rory Zahedi ‡ Winston Zuo ‡ Jonathan Balkind‡
*Barcelona Supercomputing Center †Universitat Politècnica de Catalunya ‡UC Santa Barbara

***Abstract*—Data movement is a hot-button topic today, with workloads like machine learning (ML) training, graph processing, and data analytics consuming datasets as large as 30PB. Such a dataset would take almost a week to transfer at 400gbps while consuming megajoules of energy just to operate the two endpoints' optical transceivers. All of this time and energy is seen as an unavoidable overhead on top of directly accessing the disks that store the data. In this paper, we re-evaluate the fundamental assumption of networked data copying and instead propose the adoption of embodied data movement. Our insight is that solid state disks (SSDs) have been rapidly growing in an under-exploited way: their data density, both in TB per unit volume and unit mass.**

With data centres reaching kilometres in length, we propose a new architecture featuring data centre *hyperloops*[2] (DHLs) where large datasets, stored on commodity SSDs, are moved via magnetic levitation in low-pressure tubes. By eliminating much of the potential friction inherent to embodied data movement, DHLs offer more efficient data movement, with SSDs potentially travelling at hundreds of metres per second. Consequently, a contemporary dataset can be moved through a DHL in seconds and then accessed with local latency and bandwidth well into the terabytes per second.

DHLs have the potential to massively reduce the network bandwidth and energy consumption associated with moving large datasets, but raise a variety of questions regarding the viability of their realisation and deployment. Through flexibility and creative engineering, we argue that many potential issues can be resolved. Further, we present models of DHLs and their application to workloads with growing data movement demands, such as training machine learning algorithms, large-scale physics experiments, and data centre backups. For a fixed data movement task, we obtain energy reductions of $1.6\times$ to $376.1\times$ and time speedups from $114.8\times$ to $646.4\times$ versus 400gbps optical networking. When modelling DHL in simulation, we obtain time speedups of between $5.7\times$ and $118\times$ (iso-power) and communication power reductions of between $6.4\times$ and $135\times$ (iso-time) to train an iteration of a representative DLRM workload. We provide a cost analysis, showing that DHLs are financially practical. With the scale of the improvements realisable through DHLs, we consider this paper a call to action for our community to grapple with the remaining architectural challenges.

I. Introduction

Architects today consistently cry foul on the cost of data movement. Despite specialised accelerators drastically cutting computation costs, Amdahl's Law tells us that we must now pick up the slack in terms of data movement costs, which have ballooned significantly. Accompanying this is a significant increase in the amount of data that we must process for applications including machine learning (ML), data analytics, genomics, and experimental physics [12], [31], [33], [42], [44], [52], [53], [60], [63], [68], [78], [93], [94], [98], [105]. The increasing amount of data generated per user per day is a problem growing at an alarming rate, already reaching petabytes (PB) per day for data centres, as new applications demand colossal amounts of data.

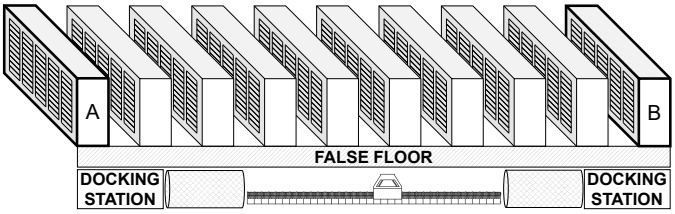


Fig. 1. A mockup of how a DHL would be located in a data centre.

Data centre networks are architected to satisfy the needs of varied, co-running applications. However, moving PB-scale datasets quickly creates bottlenecks, consuming a static portion of the data centre's total bandwidth which could be used by other, more dynamic applications. Even state-of-the-art networking and storage solutions may not be enough for the newest data-hungry applications' PB-scale datasets [77]. We motivate this emerging data movement bottleneck with a simple example. To move Meta's 29-PB ML dataset [107] from node A to B with 400Gb/s networking would take roughly 1 week. To achieve an optimised 1-hour transfer time through parallelisation of data transfer, we would need $161\times$ network speedup to greater than 64 Tbit/s, which exceeds today's top-of-rack switches. The energy of this data movement would be in the order of megajoules to hundreds of megajoules (We explain this example in more detail in Section II-C).

Data centre energy consumption has been rising and now represents a small but significant proportion of global energy consumption [26], [97]. Characterising the constituent sources [27], [39] shows that networking can become the second biggest source of data centre energy consumption [29], [79]. While copying data optically seems efficient at first, the transceiver, switching, and network interfacing energy costs are all fundamental overheads that make the process inefficient. These costs are growing as data centres become larger and larger; many data centres are already hundreds of metres long [3], [5]. In the case of ML, repeated data copies are common and necessary because datasets are distributed across the data centre, but, when it comes time to process the datasets, they must be collected onto compute nodes within the same rack or nearby racks. Data is aggregated in this way because repeated operations on a dataset are required to train a single neural network, and accessing data that is physically closer takes less time. We argue that rather than performing repeated data copying, we should change the data centre architecture

[1]These authors contributed equally to this work.

[2]HyperLoop™ is a term for high-speed transportation using magnetic levitation trains and low-pressure tubes; it does not imply a loop topology.

to move the data storage media in a form of *"embodied data movement"*. To outcompete optical networking for the movement of large datasets, we must be able to move our data for a lower cost. We do not question whether traditional networking is suitable for typical small transfers but rather focus on moving emerging PB-scale datasets.

In this paper, we explore restructuring the data centre architecture to add data centre hyperloops which physically transport SSDs containing large datasets. Hyperloops consist of a pair of rails from one endpoint to another. These rails use magnetic levitation to support and transport a payload, and they operate inside of a low-atmosphere chamber. We apply this paradigm at data centre scale to shuttle SSDs between compute nodes and cold storage.

The density of SSDs has been quietly skyrocketing, in terms of both data per volume and per mass. Large SSDs can store 10s to 100s of terabytes (TB) at a mass of hundreds of grams [16], [80]. We take advantage of the M.2 SSD form factor to pack data at exceptionally high density. Our DHL design enables high-speed, highly-efficient data transfer for very large quantities of data, eliminating data copying overheads. Using DHL, a data centre can save energy while also raising performance thanks to bulk network bandwidth being freed for other applications.

Our data centre hyperloop architecture as presented in Figure 1 is compatible with existing data centres and is feasible to implement, potentially by adapting existing, reliable small-scale hyperloop designs [64]. We evaluate our design with a high-level parameterised model and compare it against existing optical networking to demonstrate its superiority for bulk data transfers, especially over longer distances. **Note that we do not claim to offer solutions to every conceivable challenge that would come with our data propulsion approach. Rather, we make the case that the vast majority of these challenges are surmountable with a mix of flexibility and creative engineering.** With the scale of the improvements we realise through a DHL-based data centre architecture, we consider this paper a call to action for our community to grapple with the remaining architectural challenges. This work makes the following contributions:

- The observation that hyperloops can take advantage of SSD storage density, which has grown quietly but rapidly.
- Design of a DHL architecture to transport emerging petabyte-scale datasets in an energy-efficient manner.
- DHL obtains energy reductions up to $376.1\times$ over 400gbps optical fibre thanks to its improved embodied data transmission power efficiency of up to 73.3 GB/J.
- We improve data transmission rates in a variety of specific, practical use cases (vs optical networks) and pose interesting trade-offs in the architecture of data centres.
- We have modelled DHL inside the ASTRA-sim ML simulator. DHL obtains between $5.7\times$ and $118\times$ time speedups (iso-power) and between $6.4\times$ and $135\times$ power reductions (iso-time) to train an iteration of a representative DLRM workload, as compared to parallel optical links.

TABLE I
LARGE EMERGING DATASETS AND DATA CREATION RATES.

Name	Size	Type
LAION - 5B [9]	250 TB	Images
YouTube-8M [21], [25]	350k hours of video	Videos
Massive Text [82]	10.25 TB	NLP
Common Crawl [1], [19]	>9 PB	Web Crawl
MLMeta Datasets [107]	3/13/29 PB	ML
NIH Dataset [23], GSA [32], [38]	100k Genomes, 17 PB	Genomics
LHC CMS Detector [47]	150 TB/s	Physics
Meta New Daily Data [6]	4 PB/day	BigData
Youtube New Daily Videos [22], [93]	0.7-1.44 PB/day[1]	Videos

TABLE II
CURRENTLY AVAILABLE STORAGE SOLUTIONS.

Devices	Size (TB)	Package	Weight (g)	BW Seq (Rd/Wr) (MBps)
WD Gold [20]	24	3.5"	670	291
Nimbus ExaDrive [16]	100	3.5"	538	500/460
Sabrent Rocket 4 Plus [84]	8	M.2	5.67	7100/6000

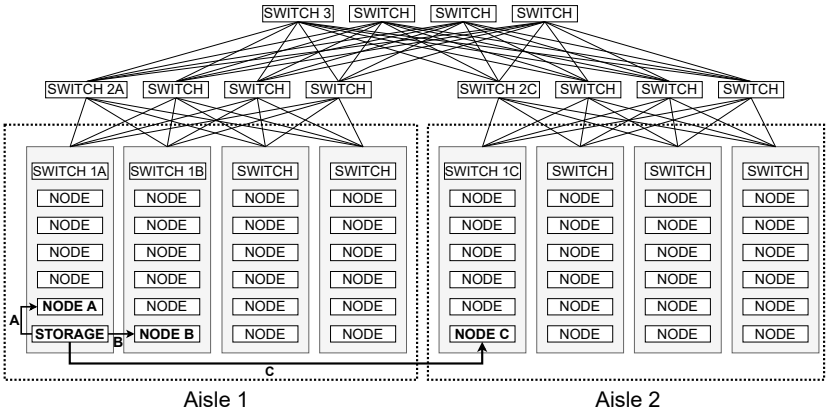
II. BACKGROUND AND MOTIVATION

For decades there has been exponential growth in data creation and dataset sizes, driven by new applications and computational capabilities. Table I summarises some of today's largest datasets and the data creation rates on several platforms. The first category shows text [45], [82], image [9], and video [21], [25] datasets used for ML training. These easily reach hundreds of TBs, with LAION - 5B consisting of 5.6 billion images (250 TB). The second category contains petabyte-scale datasets: ML training from Meta [107], massive web-crawling [19], and genomics archives [23], [32], [38]. Finally, the third category shows data creation rates from today's most popular platforms and largest scientific experiments, such as YouTube [22], [93], Meta [6], and the Large Hadron Collider (LHC) [47]. We emphasise that our proposed DHL use cases must involve large bulk transfers, as DHLs are not suited to transmitting a continuous stream of data.

A. Storage technologies

The growth of datasets has put stress on storage solutions. We observe that SSDs have been quietly growing in density without significant attention. 100TB SSDs, though expensive, beat the largest regular HDD in capacity by $5\times$. In addition, new form factors like U.2 and M.2 provide remarkably small and light packages. Table II shows three relevant examples of large HDD, SSD and M.2 SSD storage devices. Comparing the data storage per gram across form factors, the 8TB M.2 SSD is almost $100\times$ lighter than the 3.5" HDD for just $12.5\times$ less capacity. We select this form factor for the evaluation of DHL in this paper. Additionally, as storage density improves (we expect continued scaling for some time), DHLs will achieve higher embodied data transmission rates. In contrast to optical networking upgrades, we only need to upgrade the carts' SSDs and not the hyperloop itself.

[1]We have applied a conversion from 1 hour of video to 1GiB.



OPTION	ROUTE	ENERGY (MJ)
A0	STORAGE → TRAN → TRAN → NODE A	13.92
A1	STORAGE → NIC → NIC → NODE A	22.97
A2	STORAGE → NIC → SWITCH 1A → NIC → NODE A	50.05
B	STORAGE → NIC → SWITCHES 1A --- 2A --- 1B → NIC → NODE B	174.75
C	STORAGE → NIC → SWITCHES 1A -- 2A -- 3 -- 2C -- 1C → NIC → NODE B	299.45
DHL	LIBRARY → DHL Movement → STORAGE → NODE	Sec. V

Fig. 2. Left: a representative fat tree network topology with three paths between nodes, used to calculate the energy cost of transferring 29PB. Right: The five routes taken to traverse the network on the left and the corresponding energy consumption needed to transfer a 29PB dataset.

B. High-Performance Networks

Many applications demand high bandwidth networking to achieve the best performance. Consequently, data centres are usually designed with fibre optic networking. The current standards of GbE and Infiniband offer 400gbps, while 800 and 1600gbps are beginning to see adoption. One of the big challenges of optical networking is that while optical switching has proven promising [59], there has not been widespread adoption as there remains a drastic, inverse relationship between the per-port price of optical switches and the power penalty incurred by the switches [35]. Instead, we presently do electrical (digital) switching, incurring higher energy costs and delays.

Multiple studies suggest that networking is one of the main energy consumers in a data centre. Surprisingly, in some scenarios, the data centre network is underutilised [29], although it consumes up to 40% of total data centre energy. Much research has been done trying to improve the GB/J ratio, e.g., by changing the network topology, to trade off between performance and cost. Next, we characterise three possible networking setups in order to show the energy consumption of the network when moving a PB-scale dataset.

C. Case of transferring 29PB

Figure 2 shows a simple exercise to prove existing networking solutions are not suitable for transferring PB-scale datasets. We assume a storage node wants to transfer 29PB of data through the network to node A, B or C, and that node C is in a different aisle than nodes A and B. These connections are highly influenced by the network topology, which regularly trades off performance and cost. We choose a slightly simpler topology than existing cutting-edge topologies [96]. For each node, we show a plausible path, except for node A, to which we show three different connections: a direct minimal connection only accounting for the transceiver energy (A0); a direct, passive connection with regular NICs (A1); a passive connection through a switch (A2). The DHL path (last) shows that storage access is inevitable in all cases, and thus we do not model storage access in our evaluation.

TABLE III
CHARACTERIZATION OF MODERN NETWORKING POWER CONSUMPTION.

Component	Speed (Gbit/s)	Ports	Power (W)
Transceiver [2], [71]	**400**	**N/A**	**12**
NIC [8], [11]	100	N/A	15.8-22.5
NIC [15], [17]	**2x200**	**N/A**	**17-23.3**
Switch QM9700 [18]	**400 (per port)**	**32**	**747-1720**
Switch 9364D-GX2A [4]	400 (per port)	64	1324-3000

Table III shows the power for network cards and switches for different networking technologies. Note that switches' power per port depends on whether the cable is active or passive. Except A0, we assume that the network links from the nodes to the top-of-the-rack switch are passive and the rest are active. To maximise the available network performance, we use the table elements marked in bold. If we assume no interruption during the transfer, we obtain a total of 580k seconds (6.71 days) to transfer 29PB over the network at 400gbps. Of course, multiple optical links could be used in parallel, but this creates a significant burden on the data centre network. Next, we consider energy consumption.

Option *A0* is the most basic scenario only considering the power of the two directly connected transceivers with an estimated energy of 13.92MJ. Option *A1* and *A2* represent more realistic scenarios where the nodes are placed nearby, in this case, in the same rack. The estimated energy consumption of *A1* and *A2* are 22.97MJ and 50.05MJ, respectively. Option *B* and *C* represent scenarios where the nodes are in different racks, with a varying number of switches to transit. The estimated energy consumption of *B* and *C* are 174.75MJ and 299.45MJ respectively. *A more complex topology found in current data centres would further increase energy demands.*

With this simple exercise, we have shown that moving large quantities of data in a data centre can consume significant energy and several days. In a current system, this transfer would typically be distributed to several nodes and parallelized by adding more network connections to reduce the time, but this would also consume significant additional resources and not improve energy consumption. Additionally, any long

term data transfer means blocking a base amount of network bandwidth for the whole duration, which is undesirable. **This makes us question if, fundamentally, networking is not the most efficient way of moving PB-scale datasets**. A naive solution to this problem would be to even consider moving the disks by hand. Although this could be done relatively quickly, the fact that 29PB requires 1319 22TB HDDs or 290 100TB SSDs, makes the idea impractical without automation. Further, the energy and dollar cost of moving the disks by hand would likely eclipse that of optical networking.

D. Potential Applications for DHLs

We soon show the energy and cost savings of applying a DHL for large scale data movement, but we first motivate the existence of applications which would necessitate such large data transfers. We consider three settings, across experimental physics, data centre backups, and machine learning training. Each of these use cases could lead to a deployment of one or more DHLs which would be closely tailored for the specific setting. We expect the community will concoct many other potential use cases for DHLs with specialisations or generalisations of their own, especially as PB-scale datasets become more common.

1) Experimental Physics: As shown in Table I, an experiment like the CMS detector at the LHC can produce 150TB/s of data bandwidth [47]. Other high energy physics and astronomy experiments produce similarly large quantities of data which stress the available computing resources. At the LHC in particular, it is infeasible to capture all of the data from an experiment, forcing the adoption of aggressive filtering to make the problem more manageable. Not only does this filtering require custom chips, but those chips must also be radiation hardened against the experiment itself [51]. Performing machine learning directly on the unfiltered sensor data has been shown to have significant statistical power [49]. As a result, one use case we consider for DHL is to connect physics experiments to off-site processing, potentially in independent data centres, alleviating current bandwidth bottlenecks [47].

2) Data Centre Bulk Backups and Communication: To ensure redundancy and protect data in the event of a data centre failure, data centres are constantly undergoing backups. As the scale of emerging datasets grows, the size of bulk (not background trickling) backups will scale up proportionally, as will their power, bandwidth, and time costs. Bulk backups consume tremendous bandwidth and cause traffic spikes that lower the efficiency of networking in the data centre [102]. They can also reach several PB and can require significant time to realise. These bulk backups stress the network of a data centre, reducing the available bandwidth offered to other applications. Crucially, they are large in size, and occur in discrete chunks [69], which make them an ideal application for DHLs.

3) Machine Learning Applications: To build efficient machine learning systems, computer architects are constantly grappling with its prodigious computing requirements, significant parallelism, and high memory bandwidth needs. From the software perspective, there has been a trend of creating bigger and bigger models to improve accuracy in workloads like natural language processing (NLP), image and video recognition, and deep learning recommendation models (DLRM). Table IV summarises some recent large ML models in terms of their publicly shared sizes. Nowadays, it is common to see models with billions or trillions of parameters. All of the training data for these massive models must be ingested by compute nodes during the training process, creating I/O bottlenecks that are becoming more significant as accelerators' raw throughput increases [81]. Recently, there has been a rising recognition of and subsequent research into these bottlenecks in the architecture research community and beyond [99].

TABLE IV
ML MODELS WITH A SIGNIFICANT STORAGE FOOTPRINT.

Name	# Params	Size (Bytes)[2]	From	Year
GPT-3 [30]	175B	**700GB**	OpenAI	2020
Jurassic-1 [48]	178B	**712GB**	A21 labs	2021
Gopher [82]	280B	**1.12TB**	Google	2021
M6-10T [66]	10T	**40TB**	Alibaba	2021
Megatron-Turing NLG [88], [92]	1T	**4TB**	MSFT&NVDA	2022
DLRM 2022 [72], [107]	12T	**44TB**	Meta	2022

Major cloud providers now provide supercomputers in the data centre expressly for ML training, such as NVIDIA's new DGX GH200 [75]. If we consider that some of the biggest ML models require many hours of training, we can estimate the ML training energy bill at several million dollars [41]. Additionally, due to recent dataset growth, Meta has reported that the energy required for data ingestion and pre-processing can be larger than that of computation for model training [106]. This creates a strong argument for data centre architects to invest in special data centre-scale solutions to reduce the carbon footprint of training (both in terms of computation and data ingestion), potentially creating big savings in energy bills [14], [34], [54], [65], [72].

We observe that with the growth in demand for new foundation models, new models with their own independent architectures are regularly being trained on the same, large datasets. The adoption of DHLs enables the movement of these training datasets outside of the main data centre network and at significantly lower cost than if the regular network were used. We see potential for ongoing savings repeatedly and over the long term as these same datasets must be used again and again to train a variety of different models. **Given the importance of emerging machine learning models, we focus on this use case when analyzing the architecture and performance of the DHL for the rest of the paper.**

III. DHL ARCHITECTURE

DHL is a high-bandwidth data transfer mechanism that operates by shuttling SSDs at high speeds along a track, similar to a maglev train. The physical system itself could be hidden beneath the typical false floor of a data centre [100], whilst a software layer (administered via the existing optical network) abstracts away the management of data, SSDs, and

[2]Obtained from applying a common conversion of Param=32bits.

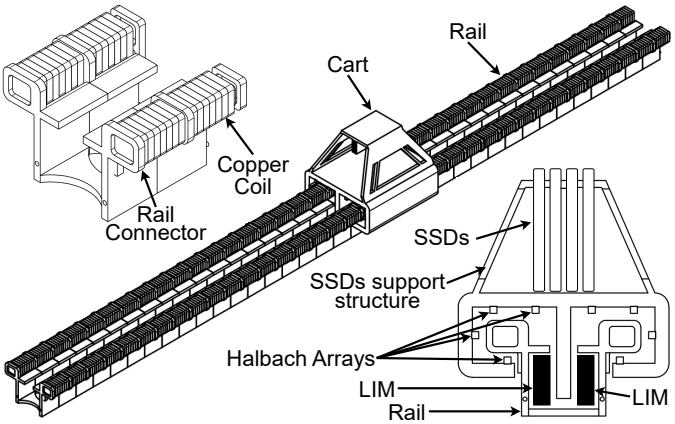


Fig. 3. An overview of the DHL cart and rail design showing the placement of their different components.

maglev carts. In this section, we explain the necessary details to realise a DHL inside a data centre.

A. Maglev Systems Foundations

We have opted for a vacuum-sealed maglev in DHL because maglev transport systems do not experience contact friction, making them long-lasting and energy efficient. Furthermore, maglev trains can operate at high speeds, enabling the system to transport SSDs hundreds of metres in just seconds. To suspend a payload over a rail, maglev typically uses a combination of permanent magnets and electromagnets. The suspension of the cart is achieved by integrating a Halbach array [58] of permanent magnets that generates a strong, directional magnetic field. When the cart is accelerated over a conductive rail [70], [73], the Halbach array generates a current in the rail, which produces a magnetic field that levitates the cart.

Maglev systems require active stabilisation to ensure the maglev train stays centred on the rails. It is only necessary to actively control the cart when it deviates from the equilibrium point. However, if the magnetic arrays are properly tuned, negligible force is required to hold the cart in place [73], minimising the costs of the stabilisation. Finally, accelerating the cart in a maglev system is typically performed by using one of the two main types of motors: linear induction motors (LIM) and linear synchronous motors (LSM). Both operate by exposing a metal fin (bottom of the cart) to a moving magnetic field generated in the rail, causing forward acceleration.

B. DHL Components

The DHL is comprised of six main components: 1) the cart, 2) the rail, 3) the accelerator, 4) the brake, 5) the docking station, and 6) the library. Below, we discuss the design of these components and their operation.

1) Cart: The cart (Figure 3) is the magnetically levitated vehicle that transports disks along the rail. The cart structure is modelled after existing maglev trains and can be constructed from polyacetal plastic, specially selected for its desirable machining characteristics, strength, and low density [61]. It contains a magnet array on the bottom to keep the cart

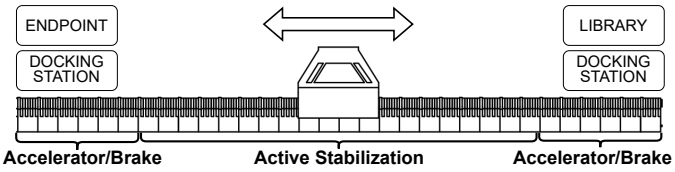


Fig. 4. Visualisation of a single DHL Rail segment.

levitating, while constraining unwanted horizontal or vertical motion. On the cart, we choose neodymium magnets because of their high magnetic flux per unit mass. The lightweight nature of current SSD technology (on the orders of grams per M.2) easily enables the design of a cart with a storage capacity of hundreds of TB. This capacity can be even higher, as SSD density continues to improve, requiring little to no engineering effort to update the cart. In addition, the number of SSDs per cart can scale depending on the needs or restrictions of a given system. *For our target use case, we assume that the SSDs never enter or leave the cart, and are instead fixed inside the cart, meaning that the cart docks with its SSDs as a single unit. SSDs may be written to/read from whenever a cart is connected to a server rack.*

2) Rail: The rail (Figure 3) is designed to be surrounded with a series of conductive rings made of aluminium or another cheap, conductive material to enable the levitation of the cart. This design is selected because it is easy to implement and it has a lift force to magnetic drag ratio exceeding 50 at speeds of greater than a few dozen metres per second (assuming copper coils) [73]. Thus, the design is highly efficient and allows us to ignore the effect of drag in our model. The rail itself is made of PVC (or another cheap plastic). The rail housing is entirely composed of individual DHL segments made of PVC which has the strength to withstand the internal vacuum. Each DHL segment contains 2 rails supporting the coils needed to induce magnetic levitation. There is an area between the 2 rails where the mechanism to accelerate or decelerate the cart is placed. It also contains the necessary electronics to achieve active stabilisation such as an array of sensors to measure the minimum deviations that the cart suffers. These segments of the rail are designed to be modular and easily connectable [89].

3) Accelerator: As noted, there are two main types of accelerators in maglev systems. We have selected LIMs for DHL due to the lower component complexity and cost associated with construction [56]. These electromagnetic accelerators are placed at the beginning of the rail (Figure 4) and move the cart at high speeds without making physical contact with it, which drastically increases the working lifetime of materials [56].

4) Brakes: Similar to how the carts are accelerated, they are decelerated using an LIM. When current is pushed in the opposite direction as during acceleration, a magnetic field is generated opposite the direction of motion, which induces a magnetic drag in the fin that will slow the cart down. The same LIM can be used to re-accelerate the cart to precisely control its location in the docking station (Figure 4).

5) Endpoints: Docking station (DS): The left of Figure 5 shows the solution proposed as the docking station for DHL. A docking station is a single location where a single cart is

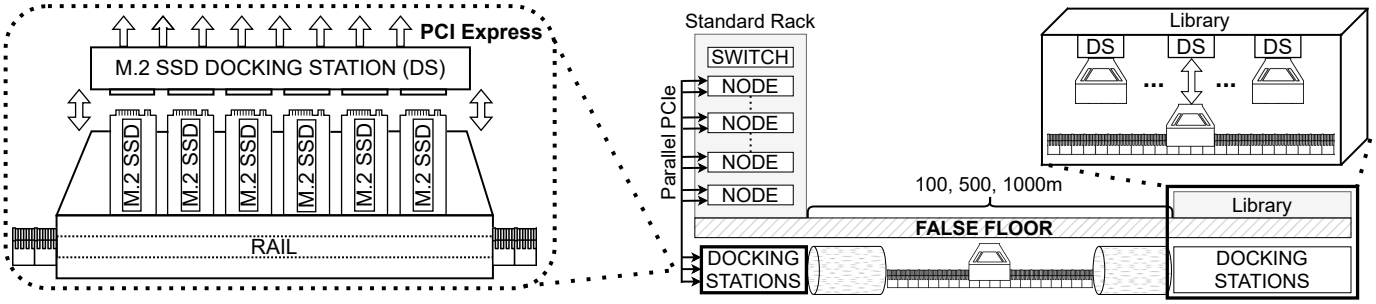


Fig. 5. DHL data centre integration: Docking Station (left); Rack connection and Library (right).

temporarily stored to interface with the adjacent rack's compute nodes. When the cart reaches an endpoint, electromagnets in the DS are briefly powered to pull the cart up off the track and engage a latch. In order to eject the cart back onto the track, the polarity of the electromagnets is reversed to release the latch, repel the cart, and place it back on the rail.

Once the cart has been connected to the docking station, by design, it has also coupled the SSDs with their PCI Express (PCIe) connectors. These give local PCIe bandwidth of many SSDs to the nodes on top of a given endpoint, which for example, in version 6 provides 3.8tbps for 64 lanes, corresponding with 1 lane per SSD in our evaluation's maximum cart configuration. We propose a solution with commonly available connectors, but these might wear out with excessive usage (Section VI).

Making the docking stations move the cart vertically enables having many docking stations at a given endpoint, located one after the other on the same rail. The only limitation is that during the cart docking process, it is not possible to shuttle another cart past the cart being docked (for a few seconds). Additionally, being able to have many docking stations in a single endpoint enables the pipelining of carts: while the endpoint is processing the first cart, the next cart can be shuttled. Each docking station can be connected to all nodes in the same rack using existing PCIe technology so each node can access many SSDs in parallel.

6) Endpoints: Library: The *Library* is a special node placed at the start/end of the DHL where SSD carts serve as cold storage. It stores SSD carts using its own internal docking stations to lift the carts from the main track. For our present design, these library docking stations are not connected to any servers. The library supports sending multiple carts in series, which enables pipelined block data transfers. Having the library enables easy expansion of the DHL's data capacity by extending the rail and adding more carts and docking stations. It also offers an easy solution to remove the carts for repair in the case of maintenance or failure.

C. Data Centre Adoption: Rack/Node Hardware Connection

Figure 5 shows DHL adoption in a data centre where one rack is connected. Carts are either stored in the library area or in one of several docking stations below the rack. For our chosen use case, we propose to have a straight DHL connecting an ML supercomputer (spanning one aisle) and the cart library. This is compatible with the typical grid design of existing data centres, and the DHL itself can be stored almost entirely underneath the "false floor" found in most data centres. The rack's docking stations each provide PCIe interfacing between the connected carts' SSDs and the compute nodes performing the ML training.

D. Data Centre Adoption: Software API

Adopting a DHL in a data centre also relies on management software to coordinate SSDs' movement. Software controls access through an API that is accessed through the standard network. It then schedules the shuttling of the carts between the library and the endpoints if the state of the system permits such an operation, e.g. the cart is not busy being read/written by a node. This can be incorporated into existing storage management APIs like *NVIDIA Magnum IO* [13], which provides a collection of tools to scale I/O for ML. DHL can be added to the storage layer to speed up bulk transactions: e.g. ML training data that can be distributed by multiple carts.

To ease adoption, DHL abstracts the scheduling of carts' launches, control of their location, and their data mapping. To avoid delays, the fact that a cart can only be in one place at a time needs to be considered. Scheduling must also account for the fact that data stored on a cart is inaccessible during transit. For our selected applications, this is not a concern. Relatedly, if an SSD fails in-flight, the endpoint's DHL API will report the error, and RAID and backups can ameliorate the issue. The API provides at least these four commands:

1) **Open**: The rack requests an SSD cart from the library. If the SSD cart is present, it is shuttled to the rack.
2) **Close**: The rack disconnects an SSD cart from a docking station. The cart is shuttled back to the library.
3) **Read**: Read data from a local SSD cart from an adjacent docking station.
4) **Write**: Writes data to an SSD cart at a specific docking station.

E. DHL Communication Architecture

Fundamentally, DHL opens the door for a different architecture contrary to existing norms, which reduces costly data movement. DHL can efficiently deliver a massive amount of data in a short period of time. Nodes at a DHL endpoint can then access this data with local latency and bandwidth, while also having random access to the many parallel SSDs in the cart. This removes several bottlenecks that current systems are forced to engineer around via techniques such as batching

and data distribution. However, DHL cannot trivially offer the same flexibility to connect thousands of nodes in the same fashion as existing networking as, for example, the SSDs cannot be accessed while in transit. Thus it is likely to replace only some uses of the data centre network. **The use cases presented in this paper are examples of where the DHL model of communication and computation could excel. Furthermore, at the current data creation rate, we envision more use cases to arise in the near future.**

Due to the complicated nature of data centre data management, we propose to use DHL as a standalone data transfer system, external to and not necessarily consistent with the overall data management system. In this form, DHL data can operate freely, accepting data from the outside system as necessary, without requiring costly global synchronisation. This model aligns with our proposed ML training use case where datasets are regularly reused (and mainly appended) to train a variety of new models.

IV. Methodology

This section describes the methodology employed to evaluate our DHL proposal. We first use an analytical model to understand the DHL parameter trade-offs and then compare it with state-of-the-art networking. Next, we perform a simulation-based comparison with state-of-the-art networking for a representative DLRM workload. Table V shows a summary of the model's parameters (the default being in bold). For the sake of analysis, we have presented a specific DHL design tailoring in this section. There exist many other, application-specific, effective design configurations/tailorings.

A. Cart Mass and Data Capacity

The components of the cart can be implemented as follows:

- **Magnets:** The mass of the cart is calculated assuming the density of neodymium magnets to be around 7.5g/cm^3. Each side of the cart consists of Halbach arrays with a few magnets each and some extra *correcting magnets* for dynamic stabilisation. With standard Halbach arrays and our track configuration, we only require 10% of the cart's mass to be comprised of magnets to achieve the necessary levitation force with an air gap of 10 mm, which is a standard levitation height for vehicles such as ours [73].
- **Central Fin:** The middle of the cart has a conductive aluminium *fin* for acceleration and braking. This fin only needs to constitute 15% of the cart's total mass in order for the LIM to produce the acceleration we require [90].
- **SSDs:** 8TB SSDs come in packages as light as 5.67 grams (in the M.2 form factor) [84]. Assuming 32 of these SSDs are loaded onto the cart, their packed size is approximately 60 mm by 60 mm by 80 mm and their mass is 180 grams. For 16 and 64 SSDs, we obtain a total mass of 91 and 363 grams, respectively.
- **Frame:** The mass of the cart frame is no greater than 30 grams per the diagram given. The frame can be light because it is made of a low-density plastic and is simply meant to stabilise the Halbach arrays and SSDs.

TABLE V
The lists represent the different DHL parameters later evaluated. The bolded entries are for our main setup.

Main Parameters	**Value**
Time to dock or undock (pessimistic)	3s
Mass of Cart	161, **282**, 524 g
Distance of DHL	100, **500**, 1000 m
Acceleration Rate	1000 m/s^2
Maximum Speed	100, **200**, 300 m/s
LIM efficiency	75%
LIM length	5, **20**, 45 m
Number of SSDs per cart	16, **32**, 64
Storage per cart	128, **256**, 512 TB

1) Acceleration: LIMs of the size used by the DHL are rated at high efficiencies ($> 75\%$) and, when implemented as in our design, can produce an acceleration of a few hundred metres per second [56]. Using a physics model, this efficiency can be used to calculate the size of the LIM and the energy required to accelerate to different speeds. The length of the LIM is proportional to the max speed of the cart, hence requiring an LIM of 5, 20 and 45 m for the different max speeds of the cart: 100, 200, and 300 *m/s*.

2) Active Stabilisation and Motion: We design a track and coils to have nearly constant lift and drag forces on the cart: the centre of the track has coils that are more spaced out, and either end has coils that are more tightly packed. This avoids issues with variable drag force on the cart, as drag force is a small constant proportion of the lift force [73]. Once the cart is coasting along the rail, the total energy loss due to drag can be calculated with the following equation:

$$L_d = (g + 2c_2)Mx/c_1$$

where g is gravitational acceleration, L_d is the energy lost to drag, M is the mass of the cart, x is the length of the rail, c_2 is the downward force generated by the bottom Halbach array, and c_1 is the lift-to-drag ratio. The lift-to-drag ratio is the only factor that is velocity dependent, and it becomes near constant at high speed [73]. Furthermore, c_2 may be driven to a small fraction of the normal force on the cart by simply having the cart ride low on the rail. Assuming a pessimistic $c_1 \approx 10$ [73], the energy lost at high velocities and small rail lengths (like 200 m/s and 500 metres or 1000 metres) is negligible. Therefore, the only power concern is from active stabilisation, which it is known to be conducted with minimal power usage [46].

3) Deceleration: We decelerate the cart using the same LIMs used for acceleration at either end of the track. Pessimistically, we assume that the cart requires as much energy to decelerate as it does to accelerate. In practice, the cart would take slightly less energy to decelerate than to accelerate because deceleration is aided by the inherent magnetic drag of the system, while the acceleration is hindered by this drag.

B. Vacuum conditions

We assume that DHL operates in a closed tube that is evacuated to a *rough vacuum* (for example, 1 millibar). This condition allows us to neglect air resistance and requires

TABLE VI
DHL DESIGN SPACE EXPLORATION FOR A SINGLE LAUNCH BETWEEN TWO ENDPOINTS (MIDDLE); ENERGY REDUCTION AND TIME SPEEDUP AS COMPARED TO DIFFERENT 400 GBPS NETWORK SCENARIOS MOVING 29PB (RIGHT).

Parameters			Metrics for a single launch					Metrics for moving 29PB					
Speed	Length	Cart Data	Energy	Efficiency	Time	BW	Peak Power	Time	Energy Reduction				
(m/s)	(m)	(TB)	(KJ)	(GB/J)	(s)	(TB/s)	(kW)	Speedup	A0	A1	A2	B	C
100	500	256	3.7	68	11	23	38	229.6x	16.3x	26.9x	58.7x	204.8x	350.9x
200	500	256	15	17	8.6	30	75	295.1x	4.1x	6.7x	14.7x	51.2x	87.7x
300	500	256	34	7.6	7.8	33	113	324.6x	1.8x	3.0x	6.5x	22.8x	39x
200	**100**	256	15	17	6.6	39	75	384.5x	4.1x	6.7x	14.7x	51.2x	87.7x
200	**500**	256	15	17	8.6	30	75	295.1x	4.1x	6.7x	14.7x	51.2x	87.7x
200	**1000**	256	15	17	11	23	75	228.6x	4.1x	6.7x	14.7x	51.2x	87.7x
200	500	**128**	8.6	15	8.6	15	43	147.5x	3.6x	5.9x	12.8x	44.8x	76.8x
200	500	**256**	15	17	8.6	30	75	295.1x	4.1x	6.7x	14.7x	51.2x	87.7x
200	500	**512**	28	18	8.6	60	140	587.5x	4.4x	7.2x	15.7x	54.9x	94.0x
100	500	**128**	2.1	60	11	12	22	114.8x	14.3x	23.6x	51.4x	179.4x	307.3x
100	500	**512**	7	73	11	46	70	457.3x	17.5x	28.8x	62.9x	219.5x	376.1x
300	500	**128**	19	6.6	7.8	16	64	162.3x	1.6x	2.6x	5.7x	19.9x	34.1x
300	500	**512**	63	8	7.8	66	210	646.4x	1.9x	3.2x	7.0x	24.4x	41.8x

minimal power to maintain [76]. It is reasonable to assume that such a vacuum can be created with minimal power usage because our *hyperloop* has a small cross-section area.

C. Library Insertion/Extraction

Using an LIM to decelerate the carts enables us to have precise control over where the carts stop. The current design of the *library* stores the carts with SSDs directly above the track. Then, it arrests the cart's motion when it is aligned with its slot in the library, and inserts it using auxiliary magnets. The docking/un-docking procedure can take less than 2 seconds using state-of-the-art maglev technology [73], [85], but we conservatively assume 3 seconds for the entire procedure.

D. Evaluation Metrics and Experiments

In order to evaluate DHL, we first explore the design space of the following parameters: the max speed of the cart, the length of the track, and the cart's data storage capacity. Second, we compare the energy and time required by DHL to move the 29PB dataset explained in Section II-C against a state-of-the-art fibre optic connection. We assume the whole dataset resides in the library. We characterise a single DHL motion between two endpoints using the following metrics:

- **Energy:** Necessary energy in *KJ*.
- **Time:** Total time in *seconds*.
- **Bandwidth:** Obtained bandwidth in *TB/s*.
- **Power:** Peak power needed in *KWatts*.
- **Efficiency:** Data moved per energy employed in *GB/J*.

E. ASTRA-sim Simulation

Next, we perform a simulation study using a state-of-the-art distributed ML simulator, ASTRA-sim [83], [101]. All experiments are based on training a DLRM ML model as used by Meta with their 29PB data set. We simulate the DHL as a high-bandwidth, high-latency network layer with the parameters from the previous design space exploration, and we compare its performance in terms of time and power required to perform a single iteration (gradient descent (GD) step) with respect to a regular optical network. We consider the basic case where DHL is a single rail connecting a library to a single server with different docking stations. Modelling the DHL as a network link does not capture the quantised nature of DHL data transfers. However, the latency and bandwidth are set such that the modelled DHLs account for this difference.

The time taken to transfer data over an optical link can be reduced by adding more links in parallel. Similarly, the time taken to transfer data over a DHL can be reduced by operating multiple DHL tracks in parallel. Both adjustments require increased power consumption. To account for this we fix a specific power budget and then include the maximum number of DHLs or network links in parallel that can operate at this power. For the sake of numerical stability, we linearly downscale the dataset size and the latency for DHL by a factor of 10^7, perform the simulation, and then upscale the resulting times by the same amount. We justified this by verifying that the time per GD iteration is in fact linear in the dataset size.

V. EVALUATION

In this section, we present the evaluation of DHL. First, we perform a design space exploration of the different DHL parameters. Second, we compare DHL against the state-of-the-art fibre optic with the exercise from section II-C. Finally, we present the estimated monetary cost of a DHL.

A. Design Space Exploration

In order to understand the implications of the DHL parameters, we explore different DHL lengths (100-500-1000 *m*), maximum speed (100-200-300 *m/s*), and cart storage capacity (128-256-512 *TB*). The first three columns of Table VI present the different configurations considered, and the rest of the columns are the resulting metrics obtained with every configuration explained and highlighted in bold next.

Energy (KJ) to both launch and decelerate a single cart between two endpoints. We do not have to consider any other

stages besides acceleration and deceleration because their energy consumption is negligible, and thus track length does not affect energy. However, the maximum speed and the cart size have a big impact on the energy consumed. We observe: *(a) max speed of 100 and 200 m/s have a huge advantage over higher speeds which may become prohibitive; (b) It costs a little less than double to transport twice the data (8.6-15-28KJ for 128-256-512TB carts), which makes it energy efficient to increase the cart's data storage.*

Time (s) to move the cart between two endpoints considering: un-dock, accelerate, motion at maximum speed, decelerate, and dock. The time has two components: the time spent during acceleration/motion/deceleration (depends on DHL parameters), and the time spent to dock and un-dock (pessimistic assumption of 3 seconds each). Consequently, we can improve time by reducing the docking/un-docking, increasing the max-speed of the cart or reducing, if possible, track length. We observe: *(a) The docking/un-docking time has a huge impact on the total time to move DHL; (b) maximum speed is the parameter that most reduces the time at the expense of energy, as seen previously.*

Efficiency (GB/J): The amount of data that DHL can move per unit of energy. We observe: *(a) maximum speed of 100 m/s obtains the highest efficiency of about 70 GB/J; (b) increasing the data storage per cart slightly improves efficiency (e.g. from 60-73GB/J for 100m/s 128-512TB cart).*

Bandwidth (TB/s): The *"embodied bandwidth"* of the DHL excluding time to load and unload data from SSDs, and without any pipelining to be conservative. We obtain from 15 to 60 *TB/s*, which is between 300× and 1200× faster than fibre optic network. We observe: *(a) the majority of the time is spent on the docking and un-docking, limiting bandwidth; (b) increasing data cart storage increases bandwidth, which will continue with technology scaling and further NAND stacking; (c) DHL obtains an outstanding embodied bandwidth compared to fibre optic, even without pipelining.*

Peak Power (kW) of a DHL launch operating at maximum capacity. Note: *We can reduce DHL's peak power by adjusting the acceleration rate and max speed, slightly increasing acceleration time but reducing power.*

B. Moving 29PB dataset

We now compare DHL to optical networking when transporting large datasets. Mimicking the previous exercise in Section II-C, we only focus on the "embodied bandwidth" and the energy of DHL to transport the dataset. We do not account for the time or energy of reading the data, which must be done in both the traditional and DHL settings. In this manner, we can compare both results. Table VI shows the energy and time reductions of DHL compared to the different data centre network scenarios. Depending on the storage capacity of each cart: 128, 256 or 512 *TB*, DHL needs 227, 114 or 57 trips to transport the 29PB dataset. Since the endpoint has a limited capacity to dock carts, it needs to return the carts to the library periodically. This limitation doubles the number of total trips required. We can remove this limitation if we account for the time to read the SSDs, where we can apply pipelining: while processing a cart, launch different ones. Additionally, with two unidirectional rails, we could avoid the return travel expense. We elaborate upon these ideas in Section VI.

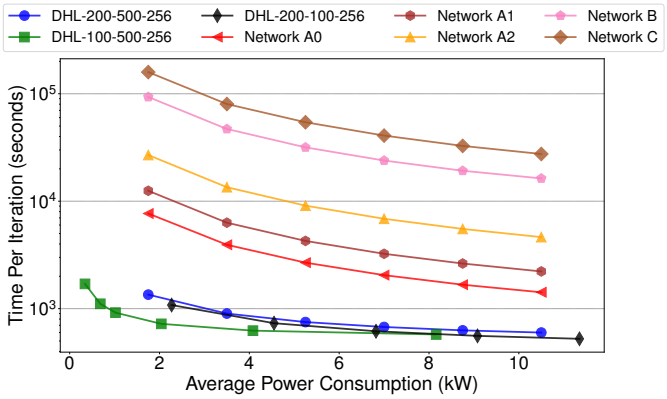


Fig. 6. Time required (Y-axis, log scale) to do an iteration of ML model training with respect to the communication power budget (X-axis) using a 29PB dataset. Modelling was conducted using ASTRA-Sim. DHL-X-Y-Z indicates a DHL operating at X m/s over a distance of Y metres with cart capacity Z terabytes. Networks A, B and C are the same as in Section II-C

DHL obtains energy reductions from 1.6× up to 376.1× depending on the baseline and DHL configuration. For the default design (maximum speed of 200 m/s, length of 500m, and cart capacity of 256 TB), DHL consumes from 4.1× up to 87.7× less energy than the optical network. Across all configurations, DHL outperforms the unrealistic direct connection scenario (Option A0 from Section II-C) in which we only account for the transceivers' energy cost. Specifically, we achieve an improvement of between 1.6× and 17.5×. If we look at the time to move the whole dataset, DHL clearly outperforms the optical fibre by a factor of between 114.8× and 646.4×. We can conclude that DHL is more energy efficient and faster than optical fibre when transferring large datasets like those used for ML training, our target use case.

C. ASTRA-sim Simulation

We use ASTRA-sim to model the time taken for a ML training iteration (including time to fetch training data and to perform the computations) using different DHL and network schemes. Figure 6 plots the time per iteration as a function of the average power budget allotted for the network. Each DHL datapoint represents the performance achieved with a discrete number of DHLs operating in parallel (the leftmost point in each curve represents the performance of a single DHL). Similarly, each network datapoint represents the performance obtained using as many links as the power budget allows (assuming a continuous, not quantised number of links for simplicity). We observe that for a fixed power budget, DHL consistently outperforms the different network scenarios.

Table VII compares the average power consumption and time taken to perform one training iteration for different communication schemes. We compare against a DHL implementation with a single 500 metre track operating at 200 m/s with 256 TB per cart. Table VII (a) gives a comparison where all networks are allotted a fixed power budget of 1.75 kW

TABLE VII
RELATIVE PERFORMANCE OF NETWORK SCHEMES VERSUS DHL

Scheme	Avg Power (kW)	Time/Iter (s)	Slowdown w.r.t. DHL
(a) Time Comparison with Fixed Average Power			
DHL	1.75	1350	1x
A0	1.75	7680	5.7x
A1	1.75	12500	9.3x
A2	1.75	26900	19.9x
B	1.75	93300	69.1x
C	1.75	159000	118x
(b) Communication Power Comparison with Fixed Iteration Time			
Scheme	**Avg Power (kW)**	**Time/Iter (s)**	**Power Increase w.r.t. DHL**
DHL	1.75	1350	1x
A0	11.2	1350	6.4x
A1	18.3	1350	10.5x
A2	39.9	1350	22.8x
B	139	1350	79.4x
C	237	1350	135x

(this is the average power consumed by our example DHL implementation). Table VII (b) gives an analogous comparison with a fixed time quota. The tables show that DHL outperforms all optical network schemes both in terms of time and energy consumption.

D. Cost

To estimate the materials cost of the DHL, we look at the commodity cost of each of the components. We do not analyse the construction cost, as this is highly variable and application-specific. The largest consideration is the components that scale with distance. The rail, made of PVC, is surrounded by aluminium rings, each of which is designed to be around 3.62 grams. The rail and carts are enclosed in a PVC tube. These costs are summarised in Table VIII (a). The acceleration and deceleration portions of DHL are achieved using an LIM that is placed at each endpoint and has 3 main components: a PVC stator, current-carrying copper coils, and a variable frequency drive (for control). The costs per LIM are summarised in Table VIII (b). Finally, the total DHL costs for a given configuration are summarised in Table VIII (c). DHL costs roughly twenty thousand dollars, which is a typical price for a large 400gbps switch of the type used in our evaluation.

E. Minimum Specifications for DHL to Outperform Optical

For the majority of this paper, we analyse the case of a 29 PB dataset transmitted over large distances. DHLs can also outperform optical networking at much smaller scales. The main trade-off involves the time required to dock and undock carts. This 6 second overhead is unavoidable, even for very small transfers and very short distances. DHLs do have one advantage here: we can afford to launch carts at much lower speeds. For a DHL with 360 GB carts, 10 m/s top speed, and 10m distance, each one-way data transfer requires 7.2 seconds and a minuscule amount of energy. A single optical link under the simplest scenario (A0) would be able to transfer the same amount of data in 7.2 seconds, but at the cost of 144 J. Therefore, DHL is desirable when transferring datasets of size at least 360 GB over at least 10 metres.

TABLE VIII
COMMODITY COST OF THE DHL MATERIALS TAKEN ON MAY 2023.

	Cost	Distance (m)		
(a) Total Rail Cost				
	(USD/kg)	**100**	**500**	**1000**
Aluminium	2.35	$117	$585	$1,170
PVC (rail)	1.20	$116	$580	$1,160
PVC (vacuum tube)	1.20	$500	$2,500	$5,000
Total	-	$733	$3,665	$7,330
(b) Total Accelerator/Decelerator Cost				
	Cost	**Top Speed (m/s)**		
	(USD/kg)	**100**	**200**	**300**
Copper Wire	8.58	$792	$2,904	$6,512
VFD	-	$8,000	$8,000	$8,000
Total	-	$8,792	$10,904	$14,512
(c) Overall Total Cost				
		Top Speed (m/s)		
		100	**200**	**300**
Distance (m)	**100**	$9,525	$11,637	$15,245
	500	$12,457	$14,569	$18,177
	1000	$16,122	$18,234	$21,842

VI. DISCUSSION

Alternative Track Designs For our primary use case, we have considered using a single DHL between endpoints with LIMs at each endpoint to both accelerate and brake. However, we can also consider a dual DHL design, with one outbound and the other inbound. This enables two notable improvements: First, it eases pipelining, as carts could shuttle back and forth simultaneously. Second, it enables the use of passive brakes that do not require external power e.g. an eddy current brake, which is a set of permanent magnets that induce magnetic drag in the fin as it passes through. This would eliminate the power cost of using an LIM for braking, essentially halving DHL's power consumption.

Regenerative braking: It is possible to recover some energy when braking the cart [40] as is done for other electric vehicles, which would increase DHL's efficiency. Regenerative braking implementations' efficiency range from 16%-70%. We emphasise that *even without regenerative braking, DHL surpasses the speed and power consumption of standard network transfer potentially by orders of magnitude.*

Multi-stops: For our target use case, the proposed design only includes two endpoints, however a multi-stop DHL is also possible. This would entail a slightly more complicated mechanism to stop carts at any desired location as well as management of carts at different velocities. Our proposed system is designed to extend to this use case without significant modifications. Multi-stop would motivate higher speeds to ameliorate potential contention from different users.

Repairs We propose placing DHLs below the *false-floor*. This makes it possible to do repairs with reasonable access.

Heat Sinks An M.2 SSD can consume up to 10W under load, hence using many at the same time can potentially create

a heat dissipation problem. It can be solved by placing heat sinks between M.2 connectors to conductively cool them.

SSD Failures Raised and oriented vertically, SSDs are kept out of the stronger magnetic field to avoid errors due to induced currents.

Increasing Connector Longevity USB-C connectors (which can physically carry PCIe) are designed for 10K-20k plug/unplug cycles, making them a good choice for repeated docking and undocking, compared to M.2's 100s of cycles.

Safety Considerations DHL presents only a minor safety concern due to carts' speed since carts' mass would be in the hundreds of grams, keeping their embodied energy small. Placing the DHL beneath the false-floor would minimise the risk of equipment damage. Additional measures can be as simple and cheap as placing sandbags at rails' ends.

System-level Concerns Our present API is very simple in correspondence with the primary intended application having limited requirements. However, in a real system, concerns such as filesystem organisation, global consistency, reliability, etc would come to the fore. In some ways, DHL looks like a more limited traditional network link (with e.g. high latency), like those used to connect poorly-connected communities. This area of work may serve as useful inspiration for future work on more general purpose system use cases of DHL.

VII. Related work

A. Near-Data Computing and In-Memory Processing

Memory bottlenecks are exacerbated when PB-scale datasets must be distributed. Caribou [57] is a distributed storage layer that offloads portions of database operations to the storage nodes themselves for efficiency and performance. Cho et al. [36] propose Near Data Accelerators (NDA) to optimise the storage locality of memory, improving power and performance. Mutlu et al. [74] seek to eliminate memory bottlenecks by performing Processing in Memory (PIM).

B. Moving large quantities of data

Moving large datasets reduces the bandwidth available for other applications while consuming significant power. 60GHz wireless networks could offer a dedicated, power-efficient data centre network [95]. Physically moving data is often branded *Sneakernet [50]*, which originated decades ago. One incarnation sent drives overnight through postal service. Cloud providers offer a similar product: AWS Snowmobile [10] physically ships customer data on hard drive storage via a *"45-foot-long ruggedised shipping container pulled by a semi-trailer truck"*, shipping over 100 PB of data in only up to a few weeks' time. Additionally, there are solutions to physically move 4xSSDs inside a removable 5.25" drive [7], easing smaller scale data transport. All of these methods limit energy savings due to friction-limited movement.

C. Energy reductions in data centres

Energy consumption is a primary concern in data centres and many proposals orthogonal to DHL have been made. VMT [91] is a thermal aware job placement technique that reduces peak cooling load up to 12.8%. Popoola et al. find the most energy-optimal switch-centric DCN topologies, preferring fat trees [79]. Fuchs et al. [43] propose an accelerator layer and specialised storage layer to eliminate the occurrence of re-computations, showing savings of 50% in energy and 68% in EDP.

D. Network improvements

Optical networks create overhead when routing due to the lack of optical switching. There are proposals to address this: A low power, error-free 100-Gb/s optical packet switching method [28]; Hybrid switches with TCP congestion control [35]; New architectures for fast optical flow control [35]; Optical interconnects for rack-scale computing [103]. Additionally, energy consumption could be reduced by turning on/off network links [55]; switching on/off individual optical fibres [24]; using existing Ethernet mechanisms to reduce the speed when possible [87] and a policy to control them [86]; proposing better network topologies [37]. DHL sidesteps many bandwidth and power consumption issues, demonstrating significant performance improvements in terms of GB/Joule.

E. Optimisation of Distributed ML Models

Different proposals have been made to reduce and optimize for the demands of modern machine learning applications: Deep Gradient Compression (DGC) [67] reduces communication bandwidth in large-scale distributed training. Zhang et al. [104] introduce an efficient communication architecture, Poseidon, that exploits layered model structures by overlapping communication and computation to reduce bursty network communication. Khani et al. [62] propose custom optical network interconnects to build high-bandwidth ML training clusters that improve training time up to $9.1\times$.

VIII. Acknowledgements

We thank the anonymous reviewers for their useful feedback. We thank Sebastian Gulz-Haake, Enrique Vallejo, Jeremy Lau, and Santiago Marco for providing useful assistance. G. López-Paradís has been supported by the Generalitat de Catalunya through a FI fellowship 2021FI-B00994 and by the BSC Mobility Call. This work has been partially supported by the *Ministerio de Ciencia Innovacion y Universidades* Grant PID2019-107255GB-C21 funded by MCIU/AEI/10.13039/501100011033, and Grant TED2021-132634A-I00 funded by MCIN/AEI/10.13039/501100011033, the European Union NextGenerationEU/PRTR, and by Arm through the Arm-BSC Center of Excellence.

References

[1] "Blog – Common Crawl." [Online]. Available: https://commoncrawl.org/blog/

[2] "Broadcom transceiver AFCT-91DRDHZ." [Online]. Available: https://www.broadcom.com/products/fiber-optic-modules-components/networking/optical-transceivers/qsfp-dd/afct-91drdhz

[3] "Chicago ord10." [Online]. Available: https://www.digitalrealty.com/data-centers/americas/chicago/ord10

[4] "Cisco Nexus 9300-GX2 Series Fixed Switches Data Sheet." [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-743854.html

[5] "Equinix ny4." [Online]. Available: https://www.gbig.org/activities/energystar-3130544

[6] "Facebook's Top Open Data Problems - Meta Research." [Online]. Available: https://research.facebook.com/blog/2014/10/facebook-s-top-open-data-problems/

[7] "ICY Dock-CP132." [Online]. Available: https://global.icydock.com/product_356.html

[8] "Intel® Ethernet Network Adapter E810-CQDA1 for OCP 3.0 - Product Specifications." [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/184816/intel-ethernet-network-adapter-e810cqda1-for-ocp-3-0/specifications.html

[9] "LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS | LAION." [Online]. Available: https://laion.ai/blog/laion-5b

[10] "Massive Data Transfers – AWS Snowmobile – Amazon Web Services." [Online]. Available: https://aws.amazon.com/snowmobile/

[11] "N1100G - 1 x 100GbE OCP 3.0 Adapter." [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/network-adapters/n1100g

[12] "NVIDIA Deep Learning Accelerator." [Online]. Available: http://nvdla.org/

[13] "NVIDIA Magnum IO." [Online]. Available: https://www.nvidia.com/en-us/data-center/magnum-io/

[14] "NVIDIA Teams With Microsoft to Build Massive Cloud AI Computer." [Online]. Available: http://nvidianews.nvidia.com/news/nvidia-microsoft-accelerate-cloud-enterprise-ai

[15] "P2200G - 2 x 200GbE PCIe NIC." [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/network-adapters/p2200g

[16] "Specifications." [Online]. Available: https://nimbusdata.com/products/exadrive/specifications/

[17] "Specifications - ConnectX-6 InfiniBand/VPI OCP 3.0 - NVIDIA Networking Docs." [Online]. Available: https://docs.nvidia.com/networking/display/ConnectX6VPIOCP3/Specifications

[18] "Specifications - NVIDIA QM97X0 NDR SWITCH SYSTEMS USER MANUAL - NVIDIA Networking Docs." [Online]. Available: https://docs.nvidia.com/networking/display/QM97X0PUB/Specifications

[19] "Want to use our data? – Common Crawl." [Online]. Available: https://commoncrawl.org/the-data/

[20] "WD Gold Enterprise Class SATA HDD." [Online]. Available: https://www.westerndigital.com/products/internal-drives/wd-gold-sata-hdd?sku=WD1005FBYZ

[21] "YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research." [Online]. Available: https://research.google.com/youtube8m/index.html

[22] "YouTube for Press." [Online]. Available: https://blog.youtube/press/

[23] "NIH's All of Us Research Program Releases First Genomic Dataset of Nearly 100,000 Whole Genome Sequences," Mar. 2022. [Online]. Available: https://www.nih.gov/news-events/news-releases/nih-s-all-us-research-program-releases-first-genomic-dataset-nearly-100000-whole-genome-sequences

[24] D. Abts, M. Marty, P. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," in *Proceedings of the International Symposium on Computer Architecture*, 2010, pp. 338–347, iSCA'10 June 19–23,2010, Saint-Malo, France.

[25] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016. [Online]. Available: http://arxiv.org/abs/1609.08675

[26] A. S. G. Andrae and T. Edler, "On Global Electricity Usage of Communication Technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, Jun. 2015, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-1547/6/1/117

[27] J. A. Aroca, A. Chatzipapas, A. F. Anta, and V. Mancuso, "A Measurement-Based Characterization of the Energy Consumption in Data Center Servers," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 12, pp. 2863–2877, Dec. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7274318/

[28] H. Ballani, P. Costa, R. Behrendt, D. Cletheroe, I. Haller, K. Jozwik, F. Karinou, S. Lange, K. Shi, B. Thomsen, and H. Williams, "Sirius: A flat datacenter network with nanosecond optical switching," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 782–797. [Online]. Available: https://doi.org/10.1145/3387514.3406221

[29] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 92–99, Jan. 2010. [Online]. Available: http://doi.org/10.1145/1672308.1672325

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[31] D. S. Cali, K. Kanellopoulos, J. Lindegger, Z. Bingöl, G. S. Kalsi, Z. Zuo, C. Firtina, M. B. Cavlak, J. Kim, N. M. Ghiasi, G. Singh, J. Gómez-Luna, N. A. Alserr, M. Alser, S. Subramoney, C. Alkan, S. Ghose, and O. Mutlu, "SeGraM: a universal hardware accelerator for genomic sequence-to-graph and sequence-to-sequence mapping," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ser. ISCA '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 638–655. [Online]. Available: http://doi.org/10.1145/3470496.3527436

[32] T. Chen, X. Chen, S. Zhang, J. Zhu, B. Tang, A. Wang, L. Dong, Z. Zhang, C. Yu, Y. Sun, L. Chi, H. Chen, S. Zhai, Y. Sun, L. Lan, X. Zhang, J. Xiao, Y. Bao, Y. Wang, Z. Zhang, and W. Zhao, "The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types," *Genomics, Proteomics & Bioinformatics*, vol. 19, no. 4, pp. 578–583, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1672022921001637

[33] X. Chen, T. Huang, S. Xu, T. Bourgeat, C. Chung, and A. Arvind, "FlexMiner: A Pattern-Aware Accelerator for Graph Pattern Mining," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2021, pp. 581–594, iSSN: 2575-713X.

[34] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, "Dadiannao: A machine-learning supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014, pp. 609–622.

[35] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov 2018. [Online]. Available: https://opg.optica.org/optica/abstract.cfm?URI=optica-5-11-1354

[36] B. Y. Cho, Y. Kwon, S. Lym, and M. Erez, "Near data acceleration with concurrent host access," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, pp. 818–831.

[37] N. Chrysos, C. Minkenberg, M. Rudquist, C. Basso, and B. Vanderpool, "Scoc: High-radix switches made of bufferless clos networks," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 2015, pp. 402–414.

[38] CNCB-NGDC Members and Partners, Y. Xue, Y. Bao, Z. Zhang, W. Zhao, J. Xiao, S. He, G. Zhang, Y. Li, G. Zhao, R. Chen, J. Zeng, Y. Zhang, Y. Shang, J. Mai, S. Shi, M. Lu, C. Bu, Z. Zhang, Z. Du, J. Xiao, Y. Wang, H. Kang, T. Xu, L. Hao, Y. Bao, P. Jia, S. Jiang, Q. Qian, T. Zhu, Y. Shang, W. Zong, T. Jin, Y. Zhang, D. Zou, Y. Bao, J. Xiao, Z. Zhang, S. Jiang, Q. Du, C. Feng, L. Ma, S. Zhang, A. Wang, L. Dong, Y. Wang, D. Zou, Z. Zhang, W. Liu, X. Yan, Y. Ling, G. Zhao, Z. Zhou, G. Zhang, W. Kang, T. Jin, T. Zhang, S. Ma, H. Yan, Z. Liu, Z. Ji, Y. Cai, S. Wang, M. Song, J. Ren, Q. Zhou, J. Qu, W. Zhang, Y. Bao, G. Liu, X. Chen, T. Chen, S. Zhang, Y. Sun, C. Yu, B. Tang, J. Zhu, L. Dong, S. Zhai, Y. Sun, Q. Chen, X. Yang, X. Zhang, Z. Sang, Y. Wang, Y. Zhao, H. Chen, L. Lan, Y. Wang, W. Zhao, Y. Ma, Y. Jia, X. Zheng, M. Chen, Y. Zhang, D. Zou, T. Zhu, T. Xu, M. Chen, G. Niu, W. Zong, R. Pan, W. Jing, J. Sang, C. Liu, Y. Xiong, Y. Sun, S. Zhai, H. Chen, W. Zhao, J. Xiao, Y. Bao, L. Hao, M. Zhang, G. Wang, D. Zou, L. Yi, W. Zhao, W. Zong, S. Wu, Z. Xiong, R. Li, W. Zong, H. Kang, Z. Xiong, Y. Ma, T. Jin, Z. Gong, L. Yi, M. Zhang, S. Wu, G. Wang, R. Li, L. Liu, Z. Li, C. Liu, D. Zou, Q. Li, C. Feng, W. Jing, S. Luo, L. Ma, J. Wang, Y. Shi, H. Zhou, P. Zhang, T. Song, Y. Li, S. He, Z. Xiong, F. Yang, M. Li, W. Zhao, G. Wang, Z. Li, Y. Ma, D. Zou, W. Zong, H. Kang, Y. Jia, X. Zheng, R. Li, D. Tian, X. Liu, C. Li,

X. Teng, S. Song, L. Liu, Y. Zhang, G. Niu, Q. Li, Z. Li, T. Zhu, C. Feng, X. Liu, Y. Zhang, T. Xu, R. Chen, X. Teng, R. Zhang, D. Zou, L. Ma, F. Xu, Y. Wang, Y. Ling, C. Zhou, H. Wang, A. E. Teschendorff, Y. He, G. Zhang, Z. Yang, S. Song, L. Ma, D. Zou, D. Tian, C. Li, J. Zhu, L. Li, N. Li, Z. Gong, M. Chen, A. Wang, Y. Ma, X. Teng, Y. Cui, G. Duan, M. Zhang, T. Jin, G. Wu, T. Huang, E. Jin, W. Zhao, H. Kang, Z. Wang, Z. Du, Y. Zhang, R. Li, J. Zeng, L. Hao, S. Jiang, H. Chen, M. Li, J. Xiao, Z. Zhang, W. Zhao, Y. Xue, Y. Bao, W. Ning, Y. Xue, B. Tang, Y. Liu, Y. Sun, G. Duan, Y. Cui, Q. Zhou, L. Dong, E. Jin, X. Liu, L. Zhang, B. Mao, S. Zhang, Y. Zhang, G. Wang, W. Zhao, Z. Wang, Q. Zhu, X. Li, J. Zhu, D. Tian, H. Kang, C. Li, S. Zhang, S. Song, M. Li, W. Zhao, Y. Liu, Z. Wang, H. Luo, J. Zhu, X. Wu, D. Tian, C. Li, W. Zhao, H. Jing, J. Zhu, B. Tang, D. Zou, L. Liu, Y. Pan, C. Liu, M. Chen, X. Liu, Y. Zhang, Z. Li, C. Feng, Q. Du, R. Chen, T. Zhu, L. Ma, D. Zou, S. Jiang, Z. Zhang, Z. Gong, J. Zhu, C. Li, S. Jiang, L. Ma, B. Tang, D. Zou, M. Chen, Y. Sun, L. Shi, S. Song, Z. Zhang, M. Li, J. Xiao, Y. Xue, Y. Bao, Z. Du, W. Zhao, Z. Li, Q. Du, S. Jiang, L. Ma, Z. Zhang, Z. Xiong, M. Li, D. Zou, W. Zong, R. Li, M. Chen, Z. Du, W. Zhao, Y. Bao, Y. Ma, X. Zhang, L. Lan, Y. Xue, Y. Bao, S. Jiang, C. Feng, W. Zhao, J. Xiao, Y. Bao, Z. Zhang, Z. Zuo, J. Ren, X. Zhang, Y. Xiao, X. Li, X. Zhang, Y. Xiao, X. Li, D. Liu, C. Zhang, Y. Xue, Z. Zhao, T. Jiang, W. Wu, F. Zhao, X. Meng, M. Chen, D. Peng, Y. Xue, H. Luo, F. Gao, W. Ning, Y. Xue, S. Lin, Y. Xue, C. Liu, A. Guo, H. Yuan, T. Su, Y. E. Zhang, Y. Zhou, M. Chen, G. Guo, S. Fu, X. Tan, Y. Xue, W. Zhang, Y. Xue, M. Luo, A. Guo, Y. Xie, J. Ren, Y. Zhou, M. Chen, G. Guo, C. Wang, Y. Xue, X. Liao, X. Gao, J. Wang, G. Xie, A. Guo, C. Yuan, M. Chen, F. Tian, D. Yang, G. Gao, D. Tang, Y. Xue, W. Wu, M. Chen, Y. Gou, C. Han, Y. Xue, Q. Cui, X. Li, C.-Y. Li, X. Luo, J. Ren, X. Zhang, Y. Xiao, and X. Li, "Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022," *Nucleic Acids Research*, vol. 50, no. D1, pp. D27–D38, Jan. 2022. [Online]. Available: https://academic.oup.com/nar/article/50/D1/D27/6413834

[39] M. Dayarathna, Y. Wen, and R. Fan, "Data Center Energy Consumption Modeling: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2016, conference Name: IEEE Communications Surveys & Tutorials.

[40] R. A. H. de Oliveira, L. S. Mattos, A. C. Ferreira, and R. M. Stephan, "Regenerative braking of a linear induction motor used for the traction of a MagLev vehicle," in *2013 Brazilian Power Electronics Conference*, Oct. 2013, pp. 950–956, iSSN: 2165-0454.

[41] B. Dickson, "The GPT-3 economy," Sep. 2020. [Online]. Available: https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/

[42] D. R. Ditzel and t. E. team, "Accelerating ML Recommendation With Over 1,000 RISC-V/Tensor Processors on Esperanto's ET-SoC-1 Chip," *IEEE Micro*, vol. 42, no. 3, pp. 31–38, May 2022, conference Name: IEEE Micro.

[43] A. Fuchs and D. Wentzlaff, "Scaling Datacenter Accelerators with Compute-Reuse Architectures," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2018, pp. 353–366, iSSN: 2575-713X.

[44] D. Fujiki, A. Subramaniyan, T. Zhang, Y. Zeng, R. Das, D. Blaauw, and S. Narayanasamy, "GenAx: A Genome Sequencing Accelerator," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2018, pp. 69–82, iSSN: 2575-713X.

[45] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," Dec. 2020, arXiv:2101.00027 [cs]. [Online]. Available: http://arxiv.org/abs/2101.00027

[46] T. Gao, J. Yang, L. Jia, Y. Deng, W. Zhang, and Z. Zhang, "Design of New Energy-Efficient Permanent Magnetic Maglev Vehicle Suspension System," *IEEE Access*, vol. 7, pp. 135 917–135 932, 2019, conference Name: IEEE Access.

[47] V. V. Gligorov, "Real-time data analysis at the lhc: present and future," 2015. [Online]. Available: https://arxiv.org/abs/1509.06173

[48] A. Gopani, "Jurassic-1 vs GPT-3 vs Everyone Else," Aug. 2021. [Online]. Available: https://analyticsindiamag.com/jurassic-1-vs-gpt-3-vs-everyone-else/

[49] E. Govorkova, E. Puljak, T. Aarrestad, T. James, V. Loncar, M. Pierini, A. A. Pol, N. Ghielmetti, M. Graczyk, S. Summers, J. Ngadiuba, T. Q. Nguyen, J. Duarte, and Z. Wu, "Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the large hadron collider," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 154–161, feb 2022. [Online]. Available: https://doi.org/10.1038%2Fs42256-022-00441-3

[50] J. Gray, W. Chong, T. Barclay, A. Szalay, and J. vandenBerg, "Terascale sneakernet: Using inexpensive disks for backup, archiving, and data exchange," 2002.

[51] G. D. Guglielmo, F. Fahim, C. Herwig, M. B. Valentin, J. Duarte, C. Gingu, P. Harris, J. Hirschauer, M. Kwok, V. Loncar, Y. Luo, L. Miranda, J. Ngadiuba, D. Noonan, S. Ogrenci-Memik, M. Pierini, S. Summers, and N. Tran, "A reconfigurable neural network ASIC for detector front-end data compression at the HL-LHC," *IEEE Transactions on Nuclear Science*, vol. 68, no. 8, pp. 2179–2186, aug 2021. [Online]. Available: https://doi.org/10.1109/tns.2021.3087100

[52] T. J. Ham, D. Bruns-Smith, B. Sweeney, Y. Lee, S. H. Seo, U. G. Song, Y. H. Oh, K. Asanovic, J. W. Lee, and L. W. Wills, "Genesis: A Hardware Acceleration Framework for Genomic Data Analysis," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, May 2020, pp. 254–267.

[53] T. J. Ham, L. Wu, N. Sundaram, N. Satish, and M. Martonosi, "Graphicionado: A high-performance and energy-efficient accelerator for graph analytics," in *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Oct. 2016, pp. 1–13.

[54] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang, "Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Feb. 2018, pp. 620–629, iSSN: 2378-203X.

[55] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving energy in data center networks," in *7th USENIX Symposium on Networked Systems Design and Implementation (NSDI 10)*. San Jose, CA: USENIX Association, Apr. 2010. [Online]. Available: https://www.usenix.org/conference/nsdi10-0/elastictree-saving-energy-data-center-networks

[56] T. Higuchi, S. Nonaka, and M. Ando, "On the design of high-efficiency linear induction motors for linear metro," *Electrical Engineering in Japan*, vol. 137, no. 2, pp. 36–43, 2001, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eej.1086. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/eej.1086

[57] Z. István, D. Sidler, and G. Alonso, "Caribou: Intelligent distributed storage," *Proc. VLDB Endow.*, vol. 10, no. 11, p. 1202–1213, aug 2017. [Online]. Available: http://doi.org/10.14778/3137628.3137632

[58] S.-M. Jang, S.-S. Jeong, and S.-D. Cha, "The application of linear halbach array to eddy current rail brake system," *IEEE Transactions on Magnetics*, vol. 37, no. 4, pp. 2627–2629, 2001.

[59] M. R. Jokar, J. Qiu, F. T. Chong, L. L. Goddard, J. M. Dallesasse, M. Feng, and Y. Li, "Baldur: A power-efficient and scalable network using all-optical switches," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2020, pp. 153–166.

[60] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, "In-Datacenter Performance Analysis of a Tensor Processing Unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ser. ISCA '17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 1–12. [Online]. Available: http://doi.org/10.1145/3079856.3080246

[61] O. Kashihara, "Polyacetal resin composition," *Patent NumberEP 0449605*, 1991.

[62] M. Khani, M. Ghobadi, M. Alizadeh, Z. Zhu, M. Glick, K. Bergman, A. Vahdat, B. Klenk, and E. Ebrahimi, "SiP-ML: high-bandwidth optical network interconnects for machine learning training," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*,

ser. SIGCOMM '21. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 657–675. [Online]. Available: http://doi.org/10.1145/3452296.3472900

[63] S. Knowles, "Graphcore," in *2021 IEEE Hot Chips 33 Symposium (HCS)*, Aug. 2021, pp. 1–25, iSSN: 2573-2048.

[64] J. Lee, W. You, J. Lim, K.-S. Lee, and J.-Y. Lim, "Development of the reduced-scale vehicle model for the dynamic characteristic analysis of the hyperloop," *Energies*, vol. 14, no. 13, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/13/3883

[65] S. Lie, "Multi-Million Core, Multi-Wafer AI Cluster," in *2021 IEEE Hot Chips 33 Symposium (HCS)*, Aug. 2021, pp. 1–41, iSSN: 2573-2048.

[66] J. Lin, A. Yang, J. Bai, C. Zhou, L. Jiang, X. Jia, A. Wang, J. Zhang, Y. Li, W. Lin, J. Zhou, and H. Yang, "M6-10T: A Sharing-Delinking Paradigm for Efficient Multi-Trillion Parameter Pretraining," Oct. 2021, arXiv:2110.03888 [cs]. [Online]. Available: http://arxiv.org/abs/2110.03888

[67] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," Jun. 2020, arXiv:1712.01887 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1712.01887

[68] A. Lines, P. Joshi, R. Liu, S. McCoy, J. Tse, Y.-H. Weng, and M. Davies, "Loihi asynchronous neuromorphic research chip," in *2018 24th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, 2018, pp. 32–33.

[69] P. Lu, L. Zhang, X. Liu, J. Yao, and Z. Zhu, "Highly efficient data migration and backup for big data applications in elastic optical inter-data-center networks," *IEEE Network*, vol. 29, no. 5, pp. 36–42, 2015.

[70] C. Luo, K. Zhang, J. Duan, and Y. Jing, "Study of Permanent Magnet Electrodynamic Suspension System with a Novel Halbach Array," *Journal of Electrical Engineering & Technology*, vol. 15, no. 2, pp. 969–977, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s42835-019-00342-3

[71] Mia, "Optical Transceiver Technology Trends of Data Center in 2022," Mar. 2022. [Online]. Available: https://www.fibermall.com/blog/optical-transceiver-technology-trends-of-data-center-in-2022.htm

[72] D. Mudigere, Y. Hao, J. Huang, Z. Jia, A. Tulloch, S. Sridharan, X. Liu, M. Ozdal, J. Nie, J. Park, L. Luo, J. A. Yang, L. Gao, D. Ivchenko, A. Basant, Y. Hu, J. Yang, E. K. Ardestani, X. Wang, R. Komuravelli, C.-H. Chu, S. Yilmaz, H. Li, J. Qian, Z. Feng, Y. Ma, J. Yang, E. Wen, H. Li, L. Yang, C. Sun, W. Zhao, D. Melts, K. Dhulipala, K. Kishore, T. Graf, A. Eisenman, K. K. Matam, A. Gangidi, G. J. Chen, M. Krishnan, A. Nayak, K. Nair, B. Muthiah, M. khorashadi, P. Bhattacharya, P. Lapukhov, M. Naumov, A. Mathews, L. Qiao, M. Smelyanskiy, B. Jia, and V. Rao, "Software-hardware co-design for fast and scalable training of deep learning recommendation models," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ser. ISCA '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 993–1011. [Online]. Available: http://doi.org/10.1145/3470496.3533727

[73] T. Murai and H. Hasegawa, "Electromagnetic analysis of inductrack magnetic levitation," *Electrical Engineering in Japan*, vol. 142, no. 1, pp. 67–74, 2003, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eej.10061. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/eej.10061

[74] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, "Processing data where it makes sense: Enabling in-memory computation," *Microprocessors and Microsystems*, vol. 67, pp. 28–41, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141933118302291

[75] NVIDIA, "Nvidia dgx gh200," 2023. [Online]. Available: www.nvidia.com/en-us/data-center/dgx-gh200/

[76] J. K. Nøland, "Evolving toward a scalable hyperloop technology: Vacuum transport as a clean alternative to short-haul flights," *IEEE Electrification Magazine*, vol. 9, no. 4, pp. 55–66, 2021.

[77] L. Papageorgiou, P. Eleni, S. Raftopoulou, M. Mantaiou, V. Megalooikonomou, and D. Vlachakis, "Genomic big data hitting the storage bottleneck," *EMBnet.journal*, vol. 24, p. e910, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958914/

[78] A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, "SCNN: An accelerator for compressed-sparse convolutional neural networks," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2017, pp. 27–40.

[79] O. Popoola and B. Pranggono, "On energy consumption of switch-centric data center networks," *The Journal of Supercomputing*, vol. 74, no. 1, pp. 334–369, Jan. 2018. [Online]. Available: https://doi.org/10.1007/s11227-017-2132-5

[80] F. P. published, "Solidigm's 61 TB SSD Hopes to Vanquish HDDs," Nov. 2022. [Online]. Available: https://www.tomshardware.com/news/solidigm_61tb_ssd

[81] S. Pumma, M. Si, W.-C. Feng, and P. Balaji, "Scalable deep learning via i/o analysis and optimization," *ACM Transactions on Parallel Computing*, vol. 6, no. 2, p. 1–34, Jun 2019.

[82] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. v. d. Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. d. M. d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. d. L. Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving, "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," Jan. 2022, arXiv:2112.11446 [cs]. [Online]. Available: http://arxiv.org/abs/2112.11446

[83] S. Rashidi, S. Sridharan, S. Srinivasan, and T. Krishna, "ASTRA-SIM: Enabling sw/hw co-design exploration for distributed dl training platforms," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2020.

[84] Sabrent, "Rocket 4 Plus SSD." [Online]. Available: https://sabrent.com/products/sb-rkt4p-8tb

[85] Y. Sanagawa, H. Ueda, M. Tsuda, A. Ishiyama, S. Kohayashi, and S. Haseyama, "Characteristics of lift and restoring force in hts bulk-application to two-dimensional maglev transporter," *IEEE Transactions on Applied Superconductivity*, vol. 11, no. 1, pp. 1797–1800, 2001.

[86] K. P. Saravanan and P. M. Carpenter, "Perfbound: Conserving energy with bounded overheads in on/off-based hpc interconnects," *IEEE Transactions on Computers*, vol. 67, no. 7, pp. 960–974, 2018.

[87] K. P. Saravanan, P. M. Carpenter, and A. Ramirez, "Power/performance evaluation of energy efficient ethernet (eee) for high performance computing," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2013, pp. 205–214.

[88] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," Mar. 2020, arXiv:1909.08053 [cs]. [Online]. Available: http://arxiv.org/abs/1909.08053

[89] M. D. Simon, L. O. Heflinger, and A. K. Geim, "Diamagnetically stabilized magnet levitation," *American Journal of Physics*, vol. 69, no. 6, pp. 702–713, Jun. 2001, publisher: American Association of Physics Teachers. [Online]. Available: https://aapt.scitation.org/doi/10.1119/1.1375157

[90] M. D. Simon, L. O. Heflinger, and A. K. Geim, "Diamagnetically stabilized magnet levitation," *American Journal of Physics*, vol. 69, no. 6, pp. 702–713, 2001. [Online]. Available: https://doi.org/10.1119/1.1375157

[91] M. Skach, M. Arora, D. Tullsen, L. Tang, and J. Mars, "Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change Materials," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, Jun. 2018, pp. 15–28, iSSN: 2575-713X.

[92] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model," Feb. 2022, arXiv:2201.11990 [cs]. [Online]. Available: http://arxiv.org/abs/2201.11990

[93] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big Data: Astronomical or Genomical?" *PLOS Biology*, vol. 13, no. 7, p. e1002195, Jul. 2015, publisher: Public Library of Science.

[Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

[94] Y. Turakhia, G. Bejerano, and W. J. Dally, "Darwin: A Genomics Co-processor Provides up to 15,000X Acceleration on Long Read Assembly," in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '18. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 199–213. [Online]. Available: http://doi.org/10.1145/3173162.3173193

[95] S. G. Umamaheswaran, S. A. Mamun, A. Ganguly, M. Kwon, and A. Kwasinski, "Reducing Power Consumption of Datacenter Networks with 60GHz Wireless Server-to-Server Links," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec. 2017, pp. 1–7.

[96] I. User, "Introducing data center fabric, the next-generation Facebook data center network," Nov. 2014. [Online]. Available: https://engineering.fb.com/2014/11/14/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

[97] W. Van Heddeghem, S. Lambert, B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, "Trends in worldwide ICT electricity consumption from 2007 to 2012," *Computer Communications*, vol. 50, pp. 64–76, Sep. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366414000619

[98] S. Venkataramani, V. Srinivasan, W. Wang, S. Sen, J. Zhang, A. Agrawal, M. Kar, S. Jain, A. Mannari, H. Tran, Y. Li, E. Ogawa, K. Ishizaki, H. Inoue, M. Schaal, M. Serrano, J. Choi, X. Sun, N. Wang, C.-Y. Chen, A. Allain, J. Bonano, N. Cao, R. Casatuta, M. Cohen, B. Fleischer, M. Guillorn, H. Haynie, J. Jung, M. Kang, K.-h. Kim, S. Koswatta, S. Lee, M. Lutz, S. Mueller, J. Oh, A. Ranjan, Z. Ren, S. Rider, K. Schelm, M. Scheuermann, J. Silberman, J. Yang, V. Zalani, X. Zhang, C. Zhou, M. Ziegler, V. Shah, M. Ohara, P.-F. Lu, B. Curran, S. Shukla, L. Chang, and K. Gopalakrishnan, "RaPiD: AI accelerator for ultra-low precision training and inference," in *Proceedings of the 48th Annual International Symposium on Computer Architecture*, ser. ISCA '21. Virtual Event, Spain: IEEE Press, Nov. 2021, pp. 153–166. [Online]. Available: http://doi.org/10.1109/ISCA52012.2021.00021

[99] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A Survey on Distributed Machine Learning," *ACM Computing Surveys*, vol. 53, no. 2, pp. 30:1–30:33, Mar. 2020. [Online]. Available: http://doi.org/10.1145/3377454

[100] J. Wan, X. Gui, S. Kasahara, Y. Zhang, and R. Zhang, "Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey," *IEEE Access*, vol. 6, pp. 48 867–48 901, 2018.

[101] W. Won, T. Heo, S. Rashidi, S. Sridharan, S. Srinivasan, and T. Krishna, "ASTRA-sim2.0: Modeling hierarchical networks and disaggregated systems for large-model training at scale," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2023.

[102] H. Yang, J. Zhang, Y. Zhao, Y. Ji, J. Han, Y. Lin, S. Qiu, and Y. Lee, "Experimental demonstration of time-aware software defined networking for openflow-based intra-datacenter optical interconnection networks," *Optical Fiber Technology*, vol. 20, no. 3, pp. 169–176, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1068520013001776

[103] P. Yang, Z. Wang, Z. Wang, J. Xu, Y.-S. Chang, X. Chen, R. K. V. Maeda, and J. Feng, "Multidomain Inter/Intrachip Silicon Photonic Networks for Energy-Efficient Rack-Scale Computing Systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 3, pp. 626–639, Mar. 2020, conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[104] H. Zhang, Z. Zheng, S. Xu, W. Dai, Q. Ho, X. Liang, Z. Hu, J. Wei, P. Xie, and E. P. Xing, "Poseidon: an efficient communication architecture for distributed deep learning on GPU clusters," in *Proceedings of the 2017 USENIX Conference on Usenix Annual Technical Conference*, ser. USENIX ATC '17. USA: USENIX Association, Jul. 2017, pp. 181–193.

[105] J. Zhao, Y. Yang, Y. Zhang, X. Liao, L. Gu, L. He, B. He, H. Jin, H. Liu, X. Jiang, and H. Yu, "TDGraph: a topology-driven accelerator for high-performance streaming graph processing," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ser. ISCA '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 116–129. [Online]. Available: http://doi.org/10.1145/3470496.3527409

[106] M. Zhao, N. Agarwal, A. Basant, B. Gedik, S. Pan, M. Ozdal, R. Komuravelli, J. Pan, T. Bao, H. Lu, S. Narayanan, J. Langman, K. Wilfong, H. Rastogi, C.-J. Wu, C. Kozyrakis, and P. Pol, "Understanding data storage and ingestion for large-scale deep recommendation model training," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*. ACM, jun 2022. [Online]. Available: https://doi.org/10.1145%2F3470496.3533044

[107] M. Zhao, N. Agarwal, A. Basant, B. Gedik, S. Pan, M. Ozdal, R. Komuravelli, J. Pan, T. Bao, H. Lu, S. Narayanan, J. Langman, K. Wilfong, H. Rastogi, C.-J. Wu, C. Kozyrakis, and P. Pol, "Understanding data storage and ingestion for large-scale deep recommendation model training: industrial product," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ser. ISCA '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 1042–1057. [Online]. Available: http://doi.org/10.1145/3470496.3533044