Use Only What You Need: Judicious Parallelism For File Transfers in High Performance Networks

Md Arifuzzaman
University of Nevada, Reno
Reno, Nevada, USA
marifuzzaman@unr.edu

Engin Arslan
University of Nevada, Reno
Reno, Nevada, USA
earslan@unr.edu

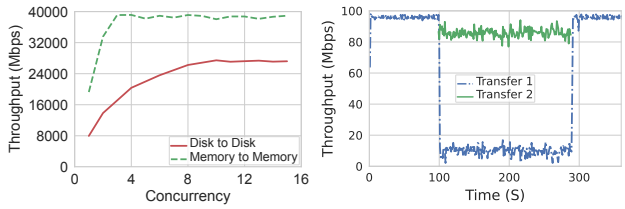
ABSTRACT

Parallelism is key to efficiently utilizing high-speed research networks when transferring large volumes of data. However, the monolithic design of existing transfer applications requires the same level of parallelism to be used for read, write, and network operations for file transfers. This in turn overburdens system resources since setting the parallelism level for the slowest component results in unnecessarily high parallelism for other components. Using more than necessary parallelism lead to high overhead on system resources and unfair resource allocation among competing transfers. In this paper, we introduce modular file transfer architecture, `Marlin`, to separate I/O and network operations for file transfers such that parallelism can be adjusted for each component independently. `Marlin` adopts online gradient descent algorithm to swiftly search the solution space and find the optimal level of parallelism for read, transfer, and write operations. Experimental results collected under various network settings show that `Marlin` minimizes system overhead significantly by identifying a minimum parallelism level for each component. We also show that it ensures fairness among competing transfers despite requiring a different level of I/O parallelism. Finally, separating network transfers from write operations allows `Marlin` to outperform the state-of-the-art solutions by more than $2x$ when transferring small datasets.

ACM Reference Format:
Md Arifuzzaman and Engin Arslan. 2024. Use Only What You Need: Judicious Parallelism For File Transfers in High Performance Networks. In *Proceedings of ACM Conference (ICS'23).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ICS'23, June 2023, Orlando., FL, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Impedance Mismatch (b) Unfair Resource Sharing

Figure 1: While transfer parallelism (aka concurrency) is necessary to achieve high transfer throughput, its optimal level is not the same for all components of file transfers (a). Setting concurrency level based on slowest operation leads to unfair resource sharing between transfers (b)

1 INTRODUCTION

Distributed science projects such as Large Hadron Collider [4] and Vera Rubin Observatory [2] require high-performance data transfers in the orders to tens of gigabits-per-second to move data between geographically distant locations in timely manner. Research networks (e.g., Internet-2 and ESnet) provide high-speed connectivity between research and education institutions with up to $100Gbps$ bandwidth to separate scientific data transfers from internet traffic thereby facilitating large-scale data movements. However, legacy file transfer applications (e.g., scp and FTP) fail to reach high utilization in these networks mainly because they adopt one file at-a-time approach which limits their I/O and network throughput significantly. Similar to compute jobs, file transfers also require I/O and network parallelism to reach high speeds since single file read/write performance as well as single network connection performance is well below available I/O and network bandwidth in research networks.

A standard approach to overcome the limitations of legacy transfer applications is transferring multiple files simultaneously (henceforth concurrency) as it can improve aggregate I/O throughput by reading and writing multiple files and overall network throughput by creating multiple network connections [24]. Hence, most previous work in this area focused on finding the optimal level of concurrency that can maximize transfer throughput [13, 14, 24, 27, 31]. However,

the monolithic architecture of existing file transfer applications requires the same level of concurrency to be used for I/O and network operations, incurring unnecessary overhead to some system resources. Figure 1(a) presents the throughput of disk-to-disk (D2D) and memory-to-memory (M2M) transfers in a network with 40Gbps bandwidth and 1 ms delay. D2D transfer moves $1000 \times 1GB$ files from the disk of the source node to the disk of destination node. Both servers are equipped with four NVMe SSD disks configured in a RAID-0 storage. M2M transfer, on the other hand, transfers dummy data from the memory of source node (`/dev/zero`) to the memory of the destination node (`/dev/null`). When concurrency is not used (i.e., concurrency level of one), M2M transfer obtains around 18Gbps whereas D2D transfer attains 8Gbps. When concurrency is set to three, the throughput of the M2M transfer reaches to maximum possible performance in this network, 40 Gbps. On the other hand, D2D transfer yields at most 26 Gbps when the concurrency is set to $9-10$.

Despite the speed mismatch between read, write, and network operations, the current implementation of concurrency in file transfers results in the same level of parallelism in all components of the transfers. In the D2D example, setting the concurrency value to 10 will create 10 transfer threads/processes in the source node to read 10 separate files from the file system and transfer them using 10 separate network connections. Similarly, there will be 10 threads/processes on the destination end to receive data packets from the network and write to the file system. Although using a high level of concurrency is harmless to transfer performance, it can have adverse impacts on resource usage. For example, previous studies show that high transfer concurrency causes up to 50% increase in energy consumption of data transfer nodes due to an increase in CPU utilization [7, 8].

The monolithic design of existing transfer applications also leads to unfair resource sharing when concurrency is used. Figure 1(b) shows the throughput of two transfers between two separate server pairs. The transfers share a network link with a capacity of 100 Mbps. We throttled the read I/O throughput of each thread to 10Mbps for Transfer-2 to simulate increased I/O performance by means of multithreading similar to parallel file systems. Transfer-1, on the other hand, does not have any I/O limitations and can attain close to 1Gbps read/write I/O throughput using a single transfer thread. Assume that users are aware of existing bottlenecks and set the concurrency level to optimal values to maximize transfer throughput, which is 1 for Transfer-1 (because single I/O and network thread is sufficient to attain reach 100Mbps throughput) and 10 for Transfer-2 (since we need 10 threads to increase read I/O throughput to 100Mbps). When the first transfer starts, it obtains 100Mbps throughput by transferring one file at-a-time (concurrency=1). When Transfer-2 joins, it sets its concurrency level to 10 to overcome the I/O limitation and attain maximum throughput. However, a concurrency value of 10 requires 10 connections to be created due to the monolithic architecture of existing file transfer applications. This in turn causes unfair bandwidth allocation since Transfer-2 creates more network connections and thus yields nearly 90% of the available bandwidth. *Therefore, while concurrency is necessary for speeding up the file transfers in high-speed networks, its current implementation results in increased resource usage and unfair resource sharing between competing transfers.*

Although researchers proposed solutions to separate I/O and network operations for file transfers (e.g. mdtmFTP [33] and FDT [1]) to overcome the limitations of monolithic designs, these solutions require manual tuning for concurrency to perform well. Moreover, they solely focus on increasing the throughput of transfers without considering system overhead (e.g., memory footprint and network congestion), hence they fall short to offer fully automated, low-overhead alternatives to existing solutions. Thus, in this work, we introduce a modular file transfer architecture, `Marlin`, to tune concurrency for read, transfer, and write operations independently. `Marlin` utilizes a game-theory-inspired utility function with online optimization algorithm to discover the optimal concurrency level for each component. The utility function is used to evaluate the fitness of different concurrency values in terms of increasing throughput and decreasing resource consumption (i.e., minimal number of threads/processes and low network packet loss). Since it is important to swiftly scan the solutions space for concurrency levels for I/O and read operations in real-time, we implemented Gradient Descent and Bayesian Optimization algorithms that can converge to the optimal in $10-15$ search intervals. Our extensive evaluations using both isolated and production systems show that `Marlin` mostly obtains similar throughput compared to the state-of-the-art solutions that use the same level of concurrency for both I/O and network operations despite minimizing system overhead significantly. We also show that it addresses the fairness issue between competing transfers with different I/O performance behavior. Finally, we show that `Marlin` can speed up the network performance of small transfers by almost $2x$ when write I/O is the bottleneck as it can take advantage of high network performance to transfer files to the cache space (e.g., main memory or NVMe buffer) on the receiver end. In summary, the contributions of this paper are as follows:

- We introduce a modular file transfer framework, `Marlin`, to separate read, transfer, and write operations of file transfers to overcome the limitations of the legacy monolithic design of state-of-the-art solutions. We show that the modular architecture allows transfers

to use "just enough" concurrency to keep the system overhead low and ensure fairness while still achieving high transfer throughput.
- We develop a game theory-inspired utility function to evaluate the performance of different concurrency values for read, transfer, and write operations. We further implement an online gradient descent algorithm to quickly and accurately discover component-specific concurrency values in real time.
- We conduct extensive experiments to demonstrate the performance of `Marlin` in both emulated and real-world production networks. In particular, we show that `Marlin` can automatically discover the slowest operation of file transfers and tunes its concurrency in real-time to maximize the performance while keeping the concurrency for other operations at a minimum. We also show that `Marlin` provides fair resource sharing between competing transfers, which is critical for production systems to ensure shared network resources are allocated to users/jobs in a fair manner.
- Finally, we show that `Marlin` can speed up the throughput of small transfers (i.e., less than 100GiB) by more than $2x$ when transfers are write-limited by quickly transferring data to cache space on receiver ends. Since more than 80% of transfer jobs in research networks transfer less than 100GiB, optimizing them is key to offering performance enhancements to most transfers.

2 RELATED WORK

Most of the existing work on the optimization of wide-area data transfers focused on designing new congestion control algorithms such as TCP BBR [17] and Vivace [18]. TCP BBR achieves higher performance than legacy TCP variants such as TCP Cubic in the presence of random packet losses. However, since file transfers in high-speed networks often face I/O performance limitations, improving the performance of congestion control algorithms is not sufficient to overcome the performance issues in today's high-performance networks.

Researchers also proposed application-layer transfer optimization solutions such as pipelining transfer commands [16], creating parallel network connections [19], transferring multiple files concurrently [24], setting TCP buffer size [22], tuning I/O block size [28], and distributing transfer load to multiple DTNs [9] to address the performance problems of file transfers. However, finding the optimal transfer setting in a timely manner has risen as a challenging problem due to having a large search space. Previous work proposed heuristic [10], historical analysis [13, 23, 27], and real-time optimization [11, 25, 31, 32] approaches to discover the optimal configuration for some of the transfer settings. As an example, Ito et al. [21] proposed Golden Section Search to automatically adjust the number of parallel TCP connections for the GridFTP transfers. Prasanna et al. [15] proposed direct search optimization to dynamically tune transfer parameters on the fly based on measured throughput for each transferred chunk.

Globus [3] is a widely-adopted data transfer service used to schedule, maintain, and optimize large data transfers in high-speed networks. It either relies on system administrators to configure transfer settings or relies on a heuristic model to estimate the optimal transfer settings for some of the application-layer transfer parameters such as pipelining, parallelism, and concurrency. To avoid overwhelming end system and network resources, it typically underestimates the value of some critical settings such as the number of concurrent transfers, and thus fails to achieve high performance in most networks. Yun et al. proposed ProbData [32] to tune the number of parallel streams and buffer size for memory-to-memory TCP transfers using stochastic approximation. ProbData can identify near-optimal configurations, but it takes several hours to find a solution, which makes it impractical to use as most transfers in high-speed networks only runs for a few minutes [26].

Yildirim et al. proposed PCP [30] to tune the values of command pipelining, concurrent file transfers, and concurrent network connections. It uses a simple hill-climbing method to scan the optimal solution for each parameter in a sequential way, thus it is neither fast nor precise. Arslan et al. proposed heuristic [14] and historical data-based (HARP [13]) models to determine the transfer settings for file transfers that can maximize the throughput. While heuristic models fail to guarantee high performance, the performance of historical data-based solutions is bound to the availability of large-scale, up-to-date historical data collected under various background loads, datasets, and transfer settings. However, collecting such datasets in a periodic manner is a daunting task for isolated networks and nearly impossible for production systems. Arifuzzaman et al. developed an online learning model, Falcon, to discover the optimal concurrency for file transfers that can maximize the transfer throughput while ensuring fairness among competing flows [11]. While Falcon addresses fairness issues when competing transfers have similar file system configurations (i.e., the same level of I/O parallelism is needed for competing transfers), it fails to do so when transfers have different I/O characteristics. Furthermore, Falcon adopts a monolithic transfer application structure, thus it fails to address the wasteful use of concurrency. Fast Data Transfer (FDT) [1] and Multicore-aware Data Transfer Middleware (mdtmFTP) implemented the idea of separating I/O and network operations, however, they both require users to tune transfer settings such as the

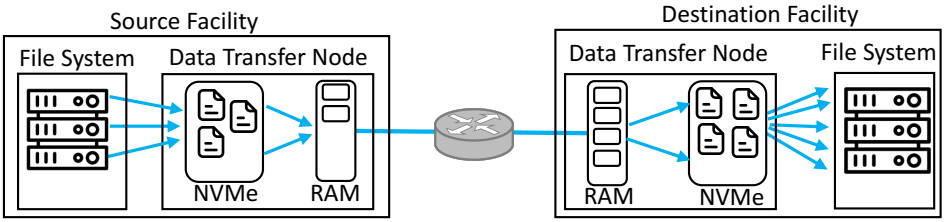


Figure 2: `Marlin` introduces modular file transfer architecture to enable fine-granular, adaptive parallelism for end-to-end transfers.

number of concurrent I/O and network threads and memory size. This is rather a challenging task even for domain experts as the optimal settings change over time.

3 MARLIN: MODULAR FILE TRANSFER APPLICATION

To scale file transfers to high speeds while avoiding to overload system resources, we build a modular file transfer application, `Marlin`, as illustrated in Figure 2. Specifically, `Marlin` separates I/O and network operations at source and destination servers to be able to tune their parallelism independently. This allows it to take advantage of I/O and network parallelism when needed to increase transfer throughput while avoiding unnecessary parallelism on well-performing components.

When selecting a concurrency level for read, transfer, and write operations, `Marlin` uses two criteria as low-overhead and high performance. That is, it searches for "just enough" concurrency values for each operation that leads to close-to-maximum performance using as minimal concurrency as possible. Hence, we adopted *utility function* proposed in [11] that rewards high throughput and penalizes high concurrency. While [11] searches for one concurrency value for read, transfer, and write operations due to using a monolithic transfer architecture, `Marlin` tunes each component separately. Thus, `Marlin` extends the utility function to meet the unique design of the proposed modular transfer architecture. Since identifying the optimal concurrency level for read, transfer, and write operations quickly is key to reaching maximum and stable transfer speed, `Marlin` uses online gradient descent algorithm as it can converge to the optimal solution with only a few sample transfers. We next discuss the details of the utility function and online optimization algorithms.

3.1 Utility Function

Utility functions are used to quantify the fitness of a configuration in terms of maximizing the benefit and minimizing the cost. In the context of file transfers, we aim to maximize the throughput and minimize the number of concurrent read/write threads and network connections. Including a punishment term into a utility function does not only help to lower the resource overhead, but also play an important role to converge to a fair and optimal solution in the presence of competition [20, 29, 34]. Hence, we adopted the following utility function as proposed in [11]

$$u(n_i, t_i, L_i) = \frac{n_i t_i}{K^{n_i}} \tag{1}$$

where n_i is the number of concurrent files to transfer (i.e., concurrency) and t_i is the average throughput of each file transfer, and K is a constant coefficient that is used to determine the severity of punishment for the concurrency level. Although previous studies show that the utility functions that incorporate monotonically increasing penalty terms in linear form guarantee high performance for single transfer and optimal and fair convergence for competing transfers (i.e., Nash Equilibrium) [20, 34], it is challenging to achieve both high-performance and fair and optimal convergence when the penalty for concurrency is incorporated in a linear form as shown in [11]. Thus, we adopted a nonlinear form for concurrency penalty that experimentally satisfies both higher performance and fairness between competing agents. As the throughput improvement ratio is not directly proportional to increased concurrency (i.e., the ratio of gain starts to lower at higher concurrency values), the value of K can be tuned to require small but non-negligible gain (e.g., 1%) for increasing concurrency values. By doing so, we ensure that the utility will increase as long as a non-negligible amount of throughput gain is observed and decrease upon exceeding the optimal concurrency value.

While the utility function given in Equation 1 is sufficient to lower resource overhead for read and write I/O operations, it is not sufficient for network transfers since it does not capture the impact on the network adequately. More specifically, one can attain higher network throughput with increased concurrency at the expense of causing or exacerbating congestion in the network. Hence, we incorporated the packet loss ratio as an additional cost for the transfer operations to keep the network congestion at a minimum. Please note that while congestion control algorithms already take packet loss into account while determining a sending rate (i.e., congestion window) for transfers, they do it on

individual connection levels. This in turn does not capture the full impact of concurrent file transfers (that are part of the same transfer job) on the network. As an example, a file transfer that uses a concurrency level of 10 for network connections can lead to a high (e.g., 1%) packet loss rate despite individual network connections experiencing a relatively small packet loss rate (e.g., 0.1%). Consequently, we calculate a total packet loss rate for all network connections and add it to the utility function as a penalty term to lower the severity of network congestion as

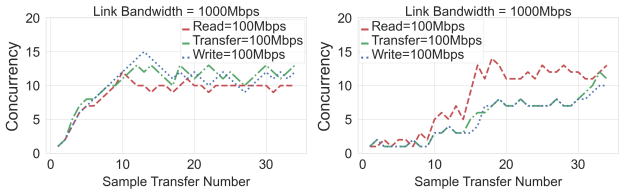
$$u(n_i, t_i, L_i) = \frac{n_i t_i}{K^{n_i}} - n_i t_i L_i \times B \tag{2}$$

where B is a constant coefficient that is used to determine the severity of punishment for packet loss penalty. We observe that $B = 10$ works well for the most commonly used TCP variants (i.e., TCP Cubic and Reno, and HSTCP) by keeping packet loss rate below $1 - 2\%$ while achieving over 95% network utilization. As a result, the utility function in the form of Equation 2 can be used to prevent high packet losses caused by suboptimal concurrency settings.

As `Marlin` is designed to tune the concurrency level for read, transfer, and network operations to different values, we have two main options in designing a search algorithm. In the first approach, we can come up with a single utility function that combines the performance of each operation to produce a single value and utilize multi-parameter optimization algorithms (e.g., conjugate gradient descent and Bayesian optimization) to search for the optimal concurrency for all operations simultaneously. In the second approach, we can utilize a separate utility function and search algorithm to tune the concurrency level of each stage of file transfers independently. For the first approach, the utility function needs to reward increased throughput for read throughput, network throughput, and write throughput while penalizing increased concurrency level for each operation as well as increased packet loss rate. Hence, we can calculate the utility of each operation using Equation 1 and 2 as

$$u(n_i, t_i, L_i) = \frac{t_r}{K^{n_r}} + \frac{t_n}{K^{n_n}} + \frac{t_w}{K^{n_w}} - t_n L \times B \tag{3}$$

where t_r, t_n, t_w are the throughput of read, transfer, and write operations; n_r, n_n, n_w are the concurrency level of read, transfer, and write operations, and L_{n_n} is the packet loss rate. We show in the evaluations that the convergence rate of optimization algorithms is extremely slow when using a combined utility function. This is mainly because of the dependence between the concurrency levels which prevents the evaluation of some settings. As an example, if the network speed is faster than the read I/O throughput, it may not be possible to evaluate a setting with a large network concurrency level and a small read I/O concurrency level due to a lack of data in the stage-in area. Hence, `Marlin` adopted a separate optimization approach to tune the concurrency level for each operation independently. Please note that optimization algorithms can be executed concurrently to minimize the solution time.



(a) Univariate Gradient Descent (b) Conjugate Gradient Descent

Figure 3: Convergence comparison of univariate and conjugate gradient descent algorithms in terms of convergence speed. While conjugate gradient can find the optimal concurrency solution for read, transfer, and write operations altogether, it takes almost three times longer to find the solution compared to using separate but concurrent univariate gradient descent algorithms.

Utility functions in the form of Equation 1 and Equation 2 converge to a fair and optimal state in the presence of multiple competing transfers due to being in the concave form. The term $1 - L_i \times B$ in Equation 2 follows a monotonically decreasing pattern for the increasing number of concurrent transfers since packet loss either stays the same or increases as the number of concurrent connection increase. Thus, both Equations 1 and 2 are guaranteed to be concave as long as $\frac{n_i t_i}{K^{n_i}}$ is concave. It is proved in [11] that for a value of $K = 0.02$, $\frac{n_i t_i}{K^{n_i}}$ is guaranteed to be strictly concave as long as n_i less than 100, which we find to be true in all production networks.

3.2 Online Search Algorithm

Naive algorithms such as brute force search may be feasible for scanning small search spaces or when the cost of evaluating a setting is minimal. However, neither of these conditions is true for file transfers as the search space is very large and it takes several seconds to accurately test a concurrency setting. As an example, there are $8,000$ possible combinations of the concurrency values for read, transfer, and write operations even when restricting concurrency values to 20 ($20 \times 20 \times 20$). Hence, it is important to devise a search algorithm that can quickly converge to the optimal solution.

Online Gradient Descent (OGD) is known to accelerate optimization processing significantly with the help of adaptive step size. OGD works by testing two close settings and calculating the gradient of the utility of these settings. As an example, assume that we test the concurrency values of 2 and 3 in two consecutive intervals. We then calculate the utility value for these two concurrency levels, say u_1 and

u_2 using the utility function (Equation 1 for I/O operations and Equation 1 for network transfer), and calculate the gradient using $\frac{u_2-u_1}{2-1}$. The gradient value is then used to decide which direction to continue the search as well as how big of a jump to make. As an example, if the gradient value is a large positive value in the above example, OGD can jump to a concurrency value of 10 instead of testing 4 or 5. By doing this, it can take large steps when current evaluated values are far from optimal to converge to the optimal solution quickly.

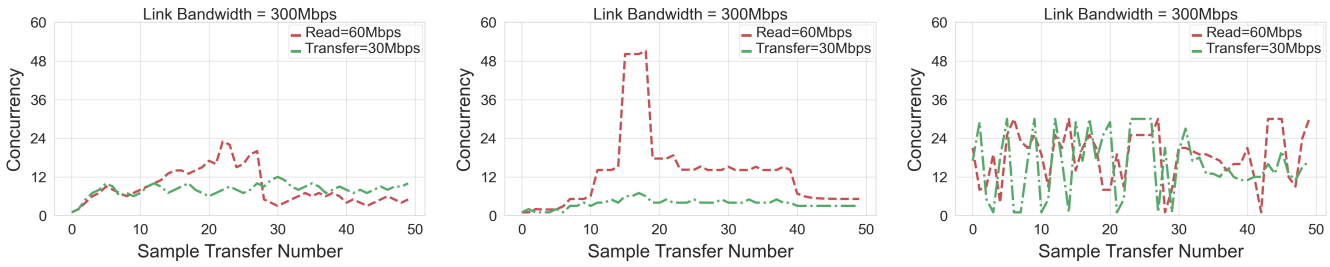
Moreover, OGD can be easily extended to keep searching continuously in case of the optimal solution changes over time. This is especially relevant for long-running transfers since network and I/O congestion may change over time so does the optimal solution. Since OGD does not have memory, it can avoid being stuck in earlier solutions and respond to changing conditions by converging to the new optimal, a key requirement to ensure fairness and high performance in shared environments. Yet, we observe that OGD can still be stuck in suboptimal regions due to a lack of differentiable differences when comparing concurrency values in suboptimal regions. For example, if the optimal concurrency is 12 but OGD ended up testing concurrency values around 20 (say it is testing 19 and 20 to calculate the gradient) and then it may not be able to learn that a lower concurrency value is better because both evaluated concurrency values return similar utility value. To overcome this limitation, we extended the base OGD implementation to keep track of the optimal concurrency setting with the highest utility value. By doing so, the OGD can avoid being stuck in the suboptimal region and continue the search around the optimal. For example, if we observed the maximum utility value at a concurrency value of 10 in the last 20 intervals (interval duration is equal to the duration of testing a setting, which is three seconds, by default.), but OGD is currently stuck at around 20 in the last few intervals, then it will come back to 10 as its utility value is highest among all concurrency values it has tested in last 20 intervals.

The conjugate gradient is commonly used to find a solution for multiple variables. Hence, we implemented both conjugate gradient descent and independent univariate gradient descent algorithm. The conjugate gradient uses Equation 3 as a utility function to evaluate the fitness of concurrency combination n_r, n_n, and n_w for read, transfer, and network operations, respectively. The univariate gradient descent, on the other hand, uses a separate OGD to identify the optimal concurrency value for each operation independently. Instead of running the three OGDs sequentially, we execute them in parallel to lower solution time as well as to capture the interplay between read, transfer, and write operations. Figure 3 presents the convergence time for univariate gradient and conjugate gradient algorithms. We set up the experiment in a way that the optimal concurrency is 10 for all three operations. It takes nearly 35 sample transfers (each sample transfer lasts for three seconds) for the conjugate gradient to find the optimal solution for each operation while it takes around 15 sample transfers when using a univariate gradient descent algorithm. Even worse, the conjugate gradient can be stuck in suboptimal regions for a long period of time before converging to the optimal. Section 4.1 presents further evaluations for the performance of `Marlin` when using different online search algorithms including univariate gradient descent, conjugate gradient descent, and bayesian optimization. Despite offering fast alternatives for multivariate parameter optimization, we find that both conjugate gradient and bayesian optimization require extensive tuning to perform well, thus we settled on univariate gradient descent as a search algorithm for `Marlin`.

Another challenge for a search algorithm is the dependence between I/O and network operations, which makes it difficult to test any random setting. As we use memory space as a stage-in area for read and transfer operations as well as transfer and write operations and memory capacity is limited, it may not be possible to evaluate a concurrency setting properly if the stage-in area is full or there is not enough data in the stage-in area to move. As an example, if read I/O performance is faster than network transfer performance (either due to lack of network parallelism or because of hardware performance differences), then the stage-in area can reach the limit. This in turn will prevent testing higher I/O concurrency values accurately because I/O threads will not be able to operate at the full speed. In the opposite scenario, if the read throughput is slower than the network throughput and there is little to no data in the stage-in area, it is not possible to test higher concurrency values for the transfer operation if there is not enough data to transfer. A similar relationship exists between the transfer and write operations at the receiver end. Under such scenarios, we ignore the outcome of sample transfers due to misleading results. As an example, if we try to test a concurrency value of 5 by opening five connections to transfer files over the network but only three connections are able to work at full speed due to lack of enough data in the staging area, we cannot use the measured network throughput as "true" performance of concurrency value 5.

4 EVALUATION

We assess the performance of `Marlin` in four networks as listed in Table 1 out of which Expanse [6] and Bridges-2 [5] are production HPC clusters, HPCLab is an isolated lab cluster, and Emulab is an emulated network testbed. Except for Emulab, all clusters have parallel file systems or RAID arrays as storage due to which the use of concurrency is required to



(a) Univariate Gradient Descent (b) Conjugate Gradient Descent (c) Bayesian Optimization

Figure 4: Comparisons of different search algorithms. While conjugate gradient descent and bayesian optimization algorithms can tune multiple parameters at the same time, their long convergence time as well as poor prediction accuracy make them hard to adapt. Univariate gradient descent, on the other hand, works well as it can converge to optimal quickly and does not require extensive tuning.

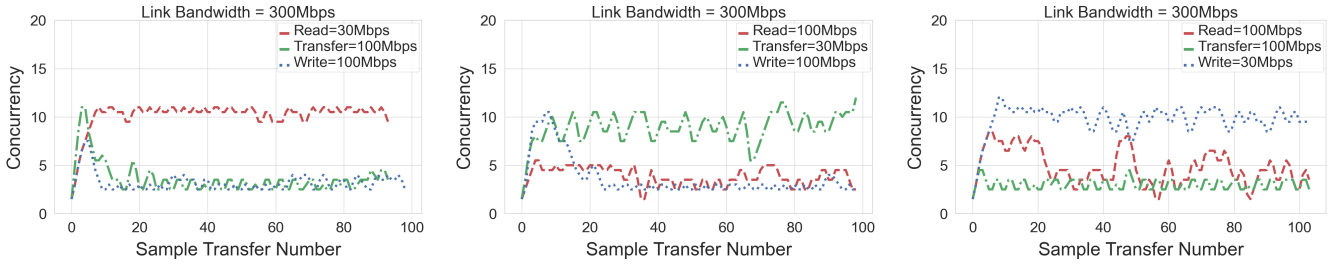
Source	Destination	Storage	Bandwidth	RTT
Emulab	Emulab	RAID-0 SSD	1G	2ms
HPCLab	HPCLab	RAID-0 SSD	20G	0.1ms
HPCLab	Expanse (SDSC)	Lustre	10G	15ms
Bridges2 (PSC)	Expanse (SDSC)	Lustre	10G	58ms

Table 1: Specifications of experimental networks. Emulab is an emulated network testbed and HPCLab is an isolated lab cluster. Expanse [6] and Bridges-2 [5] are production supercomputers that are connected via high-speed research networks.

maximize I/O performance. Since Emulab nodes have direct-attached single disk storage volumes, we throttle per process disk read throughput to necessitate concurrent I/O accesses to reach maximum performance, similar to parallel file systems. We use Bridges-2 and Expanse clusters for real-world wide-area experiments. HPCLab servers are located in the same local-area network, thus delay between the hosts is less than a millisecond. Unless otherwise states, we used datasets containing multiple 1 GiB files. The number of files is adjusted based on achievable throughput in each network. We compare `Marlin` against the state-of-the-art monolithic transfer application Falcon [11]. Similar to `Marlin`, Falcon uses the online gradient descent algorithm to search for the optimal transfer concurrency. Falcon is shown to outperform other file transfer applications (HARP [12] and Globus [10]) by up to 2×, thus we believe that comparison to Falcon would be sufficient to evaluate the efficiency of `Marlin`.

4.1 Evaluation of Optimization Algorithms

We created a testbed with 300Mbps link bandwidth, 60Mbps read I/O limitation per thread, and 30Mbps network limitations for per network connection to compare the performance of different optimization algorithms. Hence, the optimal concurrency is 5 for the read operation, 10 for the network operation, and 1 for the write operation. Figure 4 presents convergence behavior when using univariate gradient descent, conjugate gradient descent, and bayesian optimization. Gradient Descent quickly discovers that the optimal concurrency for the network is around 10. However, it keeps increasing read concurrency until it hits the memory limit because increased read concurrency leads to an increase in read throughput until the staging area becomes full. Then, it reduces the number of read threads to around 5 to match the speed of the network transfer. Conjugate gradient descent, on the other hand, chooses a particular search direction and keeps exploring that direction until it converges. This behavior leads to undesired behavior when increasing read concurrency results in a temporary increase in read throughput despite slower network speed due to having a high-performance staging area. As can be seen in Figure 4(b), the conjugate gradient descent increases read concurrency to almost 50 because of misleading information collected in the first 20 intervals during which increasing read concurrency resulted in an almost proportional increase in the read throughput. It eventually lowers the number of read threads but never falls short to find the optimal concurrency for the network thread. Bayesian optimization, similar to conjugate gradient, can tune all three concurrency values simultaneously using Equation 3 as a utility function. It starts with a few random concurrency combinations to start building a Gaussian surrogate model. It updates the model after each new observation (i.e., new sample transfer with a different concurrency setting), and predicts new values to test in the next interval. The performance of the bayesian optimization is highly dependent on the accuracy of observations. Throughput fluctuations and a temporary increase in read throughput in turn cause the bayesian optimization to build an incorrect model, which then affects its ability



(a) Read I/O Bottleneck (b) Network Bottleneck (c) Write I/O Bottleneck

Figure 5: `Marlin` can identify the bottleneck operation in file transfers and tune the concurrency accordingly to maximize transfer performance while keeping the overhead on other components at a minimum.

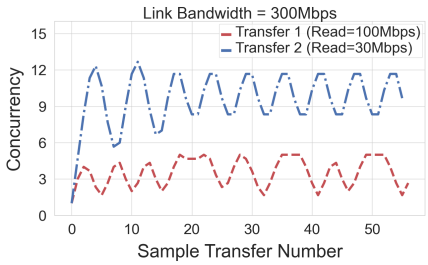


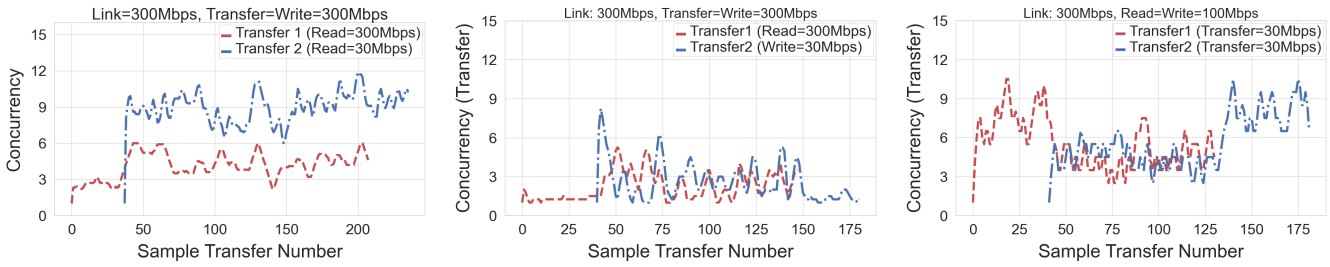
Figure 6: Due to its monolithic design, Falcon chooses the same concurrency level for read, write, and transfer operations based on the lowest-performing component.

to make accurate predictions. The simplicity of the univariate gradient descent algorithm along with the absence of memory makes it the perfect choice for `Marlin`. Please note however that it can be possible to tame conjugate gradient descent and bayesian optimization algorithms using a different utility function and/or restricting their memory, but we leave it as a future work and utilize the univariate gradient descent in the rest of the paper.

4.2 Fine-Tuning Concurrency

We next evaluate the performance of `Marlin` in terms of its ability to detect the bottleneck operation in a transfer and find the optimal concurrency to reach maximum utilization. Figure 5 presents the `Marlin`'s performance when concurrency is needed only for one of the read, transfer, and write operations. We used Emulab to manually restrict the throughput for I/O and network operations per thread and connection. For example, in Figure 5(a), read threads are limited to 30Mbps, network connections are limited to 100Mbps, and write threads are throttled to 100Mbps. We also limited bandwidth from the source node to the destination node to 300Mbps. Since the maximum total I/O speed is 1Gbps, the transfer task can attain 300Mbps at most (limited by network capacity) if the right concurrency values are configured. In the read bottleneck scenario (Figure 5(a)), the optimal concurrency levels are 10, 3, and 3, for read, transfer, and write operations, respectively. We observe that although `Marlin` initially increases the concurrency for all three operations, it lowers them for transfer and write operations after a few iterations while keeping it between $9-11$ for the read operation.

In the case of network bottleneck (Figure 5(b)), the optimal concurrency for read and write operations is 3 and transfer is 10. We can see that the concurrency level for both read and write operations settle at around $3-4$ while transfer concurrency changes around $8-11$. The main reason for the fluctuations in concurrency value is the continuous search functionality of the OGD. While it is possible to run OGD once and keep using the selected values, it is not desired in shared environments as the optimal concurrency is dependent on the level of congestion. Hence, OGD keeps searching around higher and lower values even after finding the optimal to be able to react to changing conditions quickly. Finally, `Marlin` is again performs well in the write bottleneck scenario(Figure 5(c)) by increasing write concurrency to around 10 while keeping the network and transfer concurrency at around $2-4$. We also executed Falcon in read I/O bottlenecks settings (Figure 6) to find what concurrency values it picks. In the first transfer (Transfer 1), three read threads (i.e., concurrency) are needed to reach maximum transfer throughput while a single network connection and write thread is sufficient. In the second transfer (Transfer-2), a concurrency level of 10 is needed for the read operation whereas a concurrency level of 1 is sufficient for transfer and write operations. It is clear that Falcon is able to find the optimal concurrency level quickly for both scenarios, however, it creates the same number of read, transfer, and write threads, unnecessarily overloading the network and receiver node.



(a) `Falcon` (Different Read I/O Performance) (b) `Marlin` (Different Read I/O Performance) (c) `Marlin` (Similar Read I/O Performance)

Figure 7: Fairness analysis for Falcon and `Marlin`. Falcon causes unfair network bandwidth allocation between competing transfers due to using the same concurrency for network and I/O operations. On the other hand, `Marlin` is able to identify I/O bottlenecks and tune the concurrency for transfer and I/O operations separately, which helps it to ensure fairness between transfers regardless of I/O characteristics.

4.3 Fairness Analysis

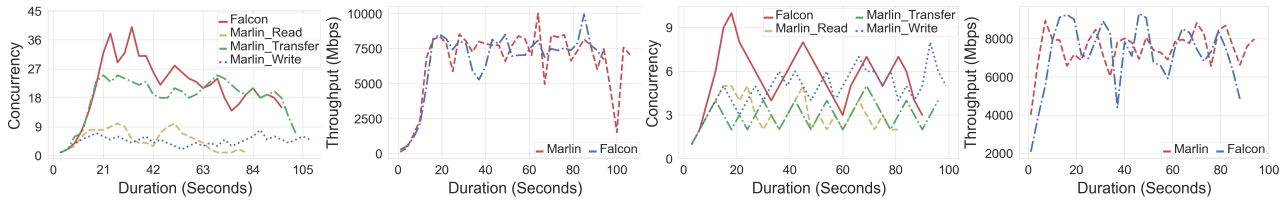
We next compare the performance of independent transfers competing for the same bottleneck link in Emulab as shown in Figure 7. We run two transfers between two different source-destination pairs with different read concurrency requirements. Specifically, read I/O throughput per thread is limited to 300Mbps for the first (Transfer-1), whereas it is limited to 30Mbps for the second transfer (Transfer-2). The transfers share a network link whose capacity is limited to 300Mbps. Thus, single write and transfer threads are sufficient for both transfers to attain the maximum possible throughput (i.e., 300Mbps) in this network as no limitations are injected for write and transfer operations. On the other hand, Transfer-2 requires 10 concurrent read threads to attain 300Mbps read I/O throughput.

Figure 7(a) presents the results for two competing `Falcon` transfers. When Transfer-1 starts, it settles on a concurrency value between $1-3$ as small concurrency is sufficient for it to reach 300Mbps throughput. When Transfer 2 joins, it tests higher concurrency values than 3 and observes a considerable increase in the utility function due to an increase in read I/O throughput. As the same concurrency level is used for read, transfer, and write operations in Falcon, this in turn causes Transfer-2 to attain a higher share in the network. That is, the capacity of the bottleneck link is shared between Transfer 1 and Transfer 2 in proportion to the number of concurrent network connections they create. Since Transfer-2 chooses a higher concurrency value than Transfer-1, it initially obtains almost three times higher network throughput than Transfer-1. Although Transfer-1 responds to this by increasing its concurrency value to around 5, it does not observe a sufficient increase in utility to increase it even further. As a result, bottleneck link capacity is shared in a 2-1 ratio between Transfer-2 and Transfer-1.

Figure 7(b) shows results for `Marlin` for the same configuration as Falcon transfers are executed. Unlike Falcon, both transfers choose to create $1-3$ network connections (on average 2.25 for Transfer 1 and 2.53 for Transfer-2) and share the network bandwidth almost equally (47% to 53%) despite having different read I/O concurrency requirement. Figure 7(c) shows two competing `Marlin` transfers with network limitations of 30Mbps, thus 10 concurrent transfer threads are needed to fully utilize the network capacity of 300Mbps. Transfer-1 converges to concurrency level 10 for the transfer operation when it is the only transfer in the network. When Transfer-2 joins, Transfer-1 lowers its network concurrency as it realizes that concurrency value $5-6$ is sufficient to attain its fair share in the network (i.e., around 150Mbps) while minimizing the packet loss rate. Transfer-2 also converges to a concurrency level of $5-6$ and obtains its fair share. When Transfer-1 completes, Transfer-2 is able to claim the free network bandwidth by increasing its network concurrency with the help of the continuous search functionality of the OGD.

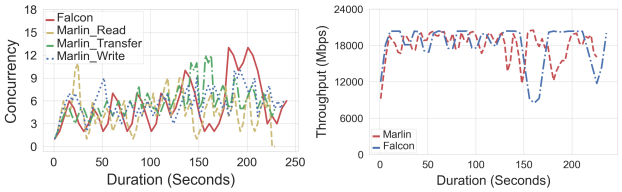
4.4 Evaluations in Production Systems

Figure 8 compares the performance comparison of Falcon and `Marlin` in production high-performance networks. While both Expanse and Bridges-2 supercomputers are equipped with high-performance Lustre file systems, have high-speed connectivity to research networks, and utilize dedicated data transfer nodes, transfers need parallelism to unleash the available capacity. In particular, Bandwidth Delay Product is 69 MiB (10Gbps×58ms) for Bridges-2 to Expanse communication, thus, TCP requires nearly 70MiB buffer space to reach 10Gbps throughput using a single connection. However, the maximum TCP buffer size is limited to 5.8 MiB in Bridges-2



(a) Bridges2-Expanse (Concurrency) (b) Bridges2-Expanse (Throughput) (c) HPCLab-Expanse (Concurrency) (d) HPCLab-Expanse (Throughput)

Figure 8: Performance comparison for `Falcon` and `Marlin` in real-world networks. `Marlin` attains competitive results in transfer throughput while lowering the concurrency value significantly.



(a) HPCLab (Concurrency) (b) HPCLab (Throughput)

Figure 9: Performance comparison for `Falcon` and `Marlin` in HPCLab network with 20Gbps bandwidth.

nodes, which is nearly twelve times smaller than the requirement. Since end users cannot change the TCP buffer size, the use of multiple concurrency network connections is the only way to mitigate TCP buffer size limitation as each connection can attain a separate TCP buffer space equal to a maximum value (i.e., 5.8MiB). Figure 8(a) and 8(b) show the concurrency and throughput values using Falcon and `Marlin`. As TCP buffer size is the main limitation for the performance, both Falcon and `Marlin` chooses a high concurrency value (around 20) for the network. On the other hand, `Marlin` realizes $3-5$ read and write threads are sufficient to read and write files at the Figure 8 compares the performance comparison of Falcon and `Marlin` in production high-performance networks. While both Expanse and Bridges-2 supercomputers are equipped with high-performance Lustre file systems, have high-speed connectivity to research networks, and utilize dedicated data transfer nodes, transfers need parallelism to unleash the available capacity. In particular, Bandwidth Delay Product (BDP) is 69 MiB (10Gbps×58ms) for Bridges-2 to Expanse communication, thus, TCP requires nearly 70MiB buffer space to reach 10Gbps throughput using a single connection. However, the maximum TCP buffer size is limited to 5.8 MiB in Bridges-2 nodes, which is nearly twelve times smaller than the requirement. Since end users cannot change the TCP buffer size, the use of multiple concurrency network connections is the only way to mitigate TCP buffer size limitation as each connection can attain a separate TCP buffer space equal to a maximum value (i.e., 5.8MiB). Figure 8(a) shows the concurrency values using Falcon and `Marlin`. As TCP buffer size is the main limitation for the performance, both Falcon and `Marlin` chooses a high concurrency value (around 20) for the network. On the other hand, `Marlin` realizes that $4-6$ read and write threads are sufficient to increase read and write throughput to maximize transfer throughput. Figure 8(b) shows that both Falcon and `Marlin`are able to obtain 8Gbps throughput. As a result, while achieving similar performance to Falcon, `Marlin` is able to lower the number of I/O processes used to read and write files. More specifically, Falcon creates $20-30$ read and write threads on end servers while `Marlin` only creates $4-6$ threads.

TCP buffer size limitation is still an issue in HPCLab-Expanse transfers (Figure 8(c) and 8(d)) since Expanse nodes are configured with 8MiB maximum TCP buffer size while BDP is around 17MiB ((10Gbps×15ms). In addition to TCP buffer size limitation, both read and write I/O operations require parallelism to overcome the I/O limitations. Specifically, a concurrency value of 6 is needed to reach close to 8Gbps write I/O throughput on the receiver end, Expanse. While read operation also needs parallelism, it can reach 8Gbps throughput with a slightly smaller concurrency value. Similar to Bridges2-Expanse transfers, `Marlin` attains comparable throughput to Falcon despite using a lower read and transfer threads. Finally, Figure 9 presents `Marlin`'s performance in a local-area network with 20Gbps bandwidth. The optimal concurrency level for read, transfer, and write operations is almost the same, around 5. Hence, both Falcon and `Marlin` performs similarly both in terms of concurrency values and throughput.

4.5 Performance Enhancements for Short Transfers

A previous analysis of the data transfers in the research networks showed that the median dataset size is 10GiB and more than 80% all transfers move less than 128GiB data. Thus, it is important to improve the performance of small transfers that last a few seconds to minutes. In most cases, I/O performance is the bottleneck for small transfers. Hence, optimizing I/O performance is critical to enhancing the throughput of short transfers. One possible solution is placing high-performance

Data Size	Falcon (Sec)	Marlin (Sec)	Improvements (%)
10 GB	9.8	7.4	24.5
25 GB	21.7	14.1	35.1
50 GB	43.8	24.3	44.5
75 GB	65.6	34.2	47.8
100 GB	86.9	42.1	51.6

Table 2: Performance comparisons of Falcon and Marlin for the transfer of small datasets when write I/O is the bottleneck of the transfers. Since Marlin is able to cache files on staging area (RAM), it can attain higher read and network throughput and move entire dataset to the destination node quicker.

storage caches (e.g., NVMe SSD, nonvolatile memory) on data transfer nodes such that files can be staged at a faster speed than directly reading/writing from parallel file systems (as illustrated in Figure 2) similar to burst buffers in HPC clusters. It is important to note small transfers are more likely to observe a significant gain through this approach because large transfers are likely to hit the capacity limit of the staging area and lower their speed to the speed of the file system.

Marlin lends itself to this idea as its modular architecture allows it to stage in files to temporary space before transferring to the network and writing to file systems. Although we utilized volatile main memory as a staging area between read and transfer operations and between transfer and write operations, it can be replaced with NVMe SSDs or nonvolatile memory units to ensure that data will not be lost in the event of a power outage. Table 2 presents the transfer duration for Falcon and Marlin when transferring small datasets in HPCLab. To simulate a scenario in which the write speed of staging space for Marlin is significantly (more than 2×) higher than the write speed of a file system, we limited the write speed to 10 Gbps while the read I/O speed is 30 Gbps, network bandwidth is 20Gbps. Thus, it is possible to transfer files to the staging area at around 20 Gbps speed. On the other hand, only 10Gbps throughput can be attained if a monolithic file transfer application is used to write data directly to the file system. Clearly, Marlin can move the files to the staging area of the destination node by more than $2x$, reducing the transfer time by up to 51.6%.

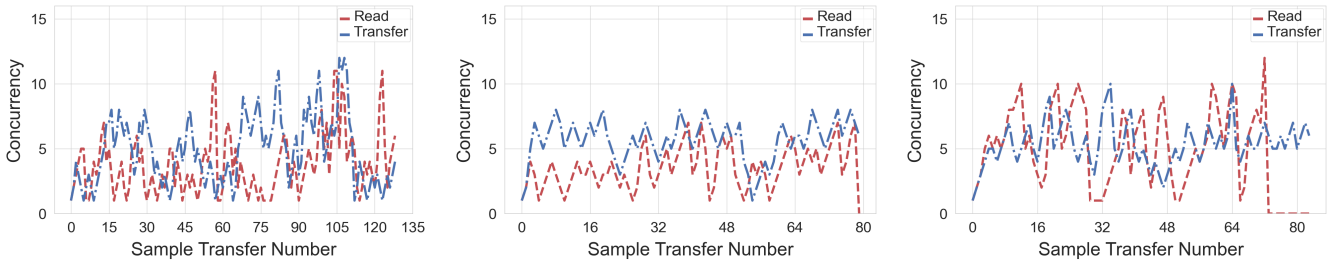
4.6 Impact of Memory Limit

Marlin uses main memory (tmpfs) on sender and receiver nodes as a staging area between network and I/O operations. Hence, it is important to limit memory usage to avoid saturating the whole memory space for a single transfer application. We, therefore, define a hard limit for memory usage on both the sender and receiver sides. By default, we set the limit to be 30% of free memory space. Figure 10 evaluates the impact of memory limit for HPCLab transfers. We varied the buffer limit on the source node between 3GiB and 100GiB. Since the speed of transfers is around 20Gbps, 3GiB allows Marlin to store around 0.6 seconds worth of data on memory before hitting the limit. This value becomes 4 seconds when using the buffer size limit of 10GiB and 40 seconds when setting the memory limit to 100GiB. The figures show that using a very small memory limit causes Marlin to experience significant fluctuations in the concurrency value of read and transfer operations. The number of read threads often hit 1 since it cannot create and test multiple read threads accurately due to lack of memory space. As a result, the transfers take 60% longer compared to using 10GiB or 100GiB memory space as shown in Figure 11. Therefore, Marlin necessitates a memory space that is at least as big as to hold a couple of seconds' worth of data. We believe this is a reasonable expectation as the memory capacity of data transfer nodes in production systems is high enough to accommodate this. As an example, the Expanse transfer node has a 64GiB memory and the Bridges-2 transfer node has a 128GiB memory.

5 CONCLUSION

Similar to compute jobs, data transfers in wide-area high-performance networks require parallelism to overcome I/O and network limitations. However, the implementation of transfer parallelism (aka concurrency) in existing transfer applications has two main issues. First, it creates the same level of parallelism for read, transfer, and write operations when transferring files. This in turn overburdens some system resources since not all operations require the same level of parallelism to achieve a similar throughput. Second, it causes unfair resource allocation when multiple transfers with different I/O characteristics share the bottleneck network link. Instead of trying to overcome these problems by manipulating existing monolithic transfer applications, we propose a modular file transfer application, Marlin, to separate read, transfer and write operations. Marlin combines game theory-inspired utility functions with online gradient descent algorithm to swiftly discover the fair and optimal parallelism levels for each operation.

We evaluated the performance of Marlin both in emulated and real-world networks to show that it is able to identify the minimum concurrency level for read, transfer, and write operations to maximize transfer throughput while minimizing system overhead and ensuring fairness among competing transfers. We also show that the modular architecture of Marlin lends itself to the implementation of burst buffer design in data transfer nodes to expedite the transfer of small datasets by caching data on high-performance storage



(a) 3GB Buffer (b) 10GB Buffer (c) 100GB Buffer

Figure 10: Impact of memory space on the performance of `Marlin`. Clearly, 10GiB is sufficient for `Marlin`to perform normally and attain high performance.

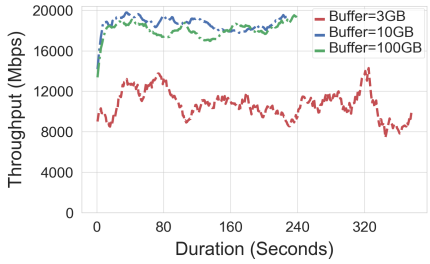


Figure 11: Throughput of `Marlin` with different memory limits. It requires at least 10GiB memory space to achieve 20 Gbps throughput in HPCLab network.

spaces (e.g., NVMe SSD and PMEM). Specifically, we find that `Marlin` can speed up the transfer performance of short transfers by more than 2×.

REFERENCES

[1] 2015. Fast Data Transfer. http://monalisa.cern.ch/FDT/.

[2] 2018. Lighting up the LSST Fiber Optic Network: From Summit to Base to Archive. lsst.org/news/lighting-lsst-fiber-optic-network-summit-base-archive.

[3] 2021. Globus. https://www.globus.org.

[4] 2021. The network challenge. https://home.cern/science/computing/network.

[5] 2023. Bridges-2. https://www.psc.edu/resources/bridges-2/.

[6] 2023. Expanse. https://www.sdsc.edu/services/hpc/expanse/.

[7] I Alan, E Arslan, and T Kosar. 2014. Energy-performance trade-offs in data transfer tuning at the end-systems. *Sustainable Computing: Informatics and Systems* 4, 4 (2014), 318–329.

[8] Ismail Alan, Engin Arslan, and Tevfik Kosar. 2015. Energy-aware data transfer algorithms. In *SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–12.

[9] William Allcock, John Bresnahan, Rajkumar Kettimuthu, Michael Link, Catalin Dumitrescu, Ioan Raicu, and Ian Foster. 2005. The Globus striped GridFTP framework and server. In *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*. IEEE Computer Society, 54.

[10] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke. 2012. Software as a Service for Data Scientists. *Commun. ACM* 55:2 (2012), 81–88.

[11] Md Arifuzzaman and Engin Arslan. 2021. Online Optimization of File Transfers in High-Speed Networks. In *High Performance Computing, Networking, Storage and Analysis, SC21: International Conference for*. IEEE.

[12] Engin Arslan, Kemal Guner, and Tevfik Kosar. 2016. HARP: predictive transfer optimization based on historical analysis and real-time probing. In *High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for*. IEEE, 288–299.

[13] Engin Arslan and Tevfik Kosar. 2018. High-Speed Transfer Optimization Based on Historical Analysis and Real-Time Tuning. *IEEE Transactions on Parallel and Distributed Systems* 29, 6 (2018), 1303–1316.

[14] Engin Arslan, Bahadir A Pehlivan, and Tevfik Kosar. 2018. Big data transfer optimization through adaptive parameter tuning. *J. Parallel and Distrib. Comput.* 120 (2018), 89–100.

[15] P. Balaprakash, V. Morozov, R. Kettimuthu, K. Kumaran, and I. Foster. 2016. Improving Data Transfer Throughput with Direct Search Optimization. In *2016 45th International Conference on Parallel Processing (ICPP)*. 248–257. https://doi.org/10.1109/ICPP.2016.36

[16] John Bresnahan, Michael Link, Rajkumar Kettimuthu, Dan Fraser, Ian Foster, et al. 2007. Gridftp pipelining. In *Proceedings of the 2007 TeraGrid Conference*.

[17] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-based congestion control. *Queue* 14, 5 (2016), 50.

[18] Mo Dong, Tong Meng, Doron Zarchy, Engin Arslan, Yossi Gilad, Brighten Godfrey, and Michael Schapira. 2018. {PCC} Vivace: Online-Learning Congestion Control. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*. 343–356.

[19] T. J. Hacker, B. D. Noble, and B. D. Atley. 2005. Adaptive Data Block Scheduling for Parallel Streams. In *Proceedings of HPDC '05*. ACM/IEEE, 265–275.

[20] Elad Hazan. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2, 3-4 (2016), 157–325.

[21] Takeshi Ito, Hiroyuki Ohsaki, and Makoto Imase. 2008. GridFTP-APT: Automatic parallelism tuning mechanism for GridFTP in long-fat networks. *IEICE transactions on communications* 91, 12 (2008), 3925–3936.

[22] T. Ito, H. Ohsaki, and M. Imase. 2008. On parameter tuning of data transfer protocol GridFTP for Wide-Area Networks. *International Journal of Computer Science and Engineering* 2(4) (Sept. 2008), 177–183.

[23] Rajkumar Kettimuthu, Gayane Vardoyan, Gagan Agrawal, and P Sadayappan. 2014. Modeling and optimizing large-scale wide-area data

transfers. In *Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on*. IEEE, 196–205.

[24] Yuanlai Liu, Zhengchun Liu, Rajkumar Kettimuthu, Nageswara Rao, Zizhong Chen, and Ian Foster. 2019. Data transfer between scientific facilities–bottleneck analysis, insights and optimizations. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. IEEE, 122–131.

[25] Zhengchun Liu, Rajkumar Kettimuthu, Ian Foster, and Peter H Beckman. 2018. Toward a smart data transfer node. *Future Generation Computer Systems* (2018).

[26] Zhengchun Liu, Rajkumar Kettimuthu, Ian Foster, and Nageswara SV Rao. 2018. Cross-geography scientific data transferring trends and behavior. In *Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing*. ACM, 267–278.

[27] MD SQ Zulkar Nine and Tevfik Kosar. 2020. A Two-Phase Dynamic Throughput Optimization Model for Big Data Transfers. *IEEE Transactions on Parallel and Distributed Systems* 32, 2 (2020), 269–280.

[28] Mohammad Javad Rashti, Gerald Sabin, and Rajkumar Kettimuthu. 2016. Long-haul secure data transfer using hardware-assisted GridFTP. *Future Generation Computer Systems* 56 (2016), 265–276.

[29] Pratiksha Thaker, Matei Zaharia, and Tatsunori Hashimoto. [n.d.]. Learning and utility in multi-agent congestion control. *optimization* 24, 10 ([n. d.]), 11–18.

[30] Esma Yildirim, Engin Arslan, Jangyoung Kim, and Tevfik Kosar. 2015. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. *IEEE Transactions on Cloud Computing* 4, 1 (2015), 63–75.

[31] Esma Yildirim, Engin Arslan, Jangyoung Kim, and Tevfik Kosar. 2016. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. *IEEE Transactions on Cloud Computing* 4, 1 (2016), 63–75.

[32] Daqing Yun, Chase Q Wu, Nageswara SV Rao, Qiang Liu, Rajkumar Kettimuthu, and Eun-Sung Jung. 2017. Data Transfer Advisor with Transport Profiling Optimization. In *Local Computer Networks (LCN), 2017 IEEE 42nd Conference on*. IEEE, 269–277.

[33] Liang Zhang, Phil Demar, Bockjoo Kim, and Wenji Wu. 2017. MDTM: Optimizing data transfer using multicore-aware I/O scheduling. In *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*. IEEE, 104–111.

[34] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 928–936.