# Throughput Analytics of Data Transfer Infrastructures*

Nageswara S. V. Rao[1], Qiang Liu[1], Zhengchun Liu[2], Rajkumar Kettimuthu[2], and Ian Foster[2]

[1] Oak Ridge National Laboratory, Oak Ridge, TN USA
{raons,liuq1}@ornl.gov
[2] Argonne National Laboratory, Argonne, IL USA
{zhengchun.liu,kettimut,foster}@anl.gov

**Abstract.** To support increasingly distributed scientific and big-data applications, powerful data transfer infrastructures are being built with dedicated networks and software frameworks customized to distributed file systems and data transfer nodes. The data transfer performance of such infrastructures critically depends on the combined choices of file, disk, and host systems as well as network protocols and file transfer software, all of which may vary across sites. The randomness of throughput measurements makes it challenging to assess the impact of these choices on the performance of infrastructure or its parts. We propose regression-based throughput profiles by aggregating measurements from sites of the infrastructure, with RTT as the independent variable. The peak values and convex-concave shape of a profile together determine the overall throughput performance of memory and file transfers, and its variations show the performance differences among the sites. We then present projection and difference operators, and coefficients of throughput profiles to characterize the performance of infrastructure and its parts, including sites and file transfer tools. In particular, the utilization-concavity coefficient provides a value in the range [0,1] that reflects overall transfer effectiveness. We present results of measurements collected using (i) testbed experiments over dedicated 0-366 ms 10 Gbps connections with combinations of TCP versions, file systems, host systems and transfer tools, and (ii) Globus GridFTP transfers over production infrastructure with varying site configurations.

## 1 Introduction

Data transport infrastructures consisting of dedicated network connections, file systems, data transfer nodes, and custom software frameworks are being deployed in scientific and commercial environments. These infrastructures are critical to many High Performance Computing (HPC) scientific workflows, which

* This work is funded by RAMSES project and the Applied Mathematics Program, Office of Advanced Computing Research, U.S. Department of Energy, and by Extreme Scale Systems Center, sponsored by U. S. Department of Defense, and performed at Oak Ridge National Laboratory managed by UT-Battelle, LLC for U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

increasingly demand higher data volumes and sophistication (e.g., streaming, computational monitoring and steering) [22]. The data transfer performance of such infrastructures critically depends on the configuration choices of:

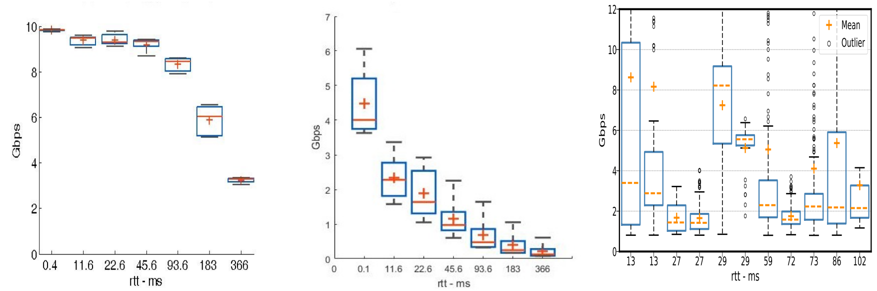
(i) data transfer host systems, which can vary significantly in terms of number of cores, Network Interface Card (NIC) capability, and connectivity;
(ii) file and disk systems, such as Lustre [23], GPFS [10], and XFS [38] installed on Solid State Disk (SSD) or hard disk arrays;
(iii) network protocols, for example, CUBIC [33], H-TCP [35], and BBR [7] versions of Transmission Control Protocol (TCP); and
(iv) file transfer software such as Globus [4] and GridFTP [3], XDD [34, 37], UDT [11], MDTM [26], and Aspera [6], and LNet extensions of Lustre [29].

Our main focus is on workloads with sufficient data volumes to require a close-to-full utilization of the underlying file, IO and network capacities.

Big data and scientific applications are becoming increasingly distributed, and often require coordinated computations at geographically distributed sites that require access to memory data and files over Wide-Area Networks (WANs) [17, 22]. Memory transfers are supported by TCP, with performance depending on its version and parameters such as buffer size and number of parallel streams. For example, Data Transfer Nodes (DTNs) used in the U. S. Department of Energy (DOE) infrastructure typically employ H-TCP and buffer sizes recommended for 200 ms Round Trip Time (RTT), and use Globus [4] to drive multiple streams for a single transfer. Typically, file systems are installed at local sites, and wide-area file transfers are carried out by using transfer frameworks [8], mounting file systems over WAN [14, 27], and extending IB Lustre to WAN using LNet routers [29]. In this paper, we specifically consider file transfers over shared and dedicated connections, such as those provisioned by ESnet's OSCARS [28] and Google's Software Defined Network (SDN) [16], and use TCP for underlying data transport. In our infrastructures, site DTNs and file system vary significantly but H-TCP and Globus are dominant options.

The transport performance of an infrastructure $\mathcal{S}$ is characterized by its *throughput profile* $\hat{\Theta}_A^{\mathcal{S}}(\tau)$ over connections of RTT $\tau$, where the modality $A = T$ corresponds to memory-to-memory transfers using TCP and $A = E$ to disk-to-disk file transfers. Such a profile is generated by aggregating throughput measurements over site connections, and is extrapolated and interpolated to other values of $\tau$ using regression or machine learning methods. It captures the combined effects of various sites, components and their configurations; in particular, for file transfers, it reflects the composition effects of file systems, network connections and their couplings through buffer management, which vary significantly across the sites of production infrastructures studied here.

Consider three scenarios: (a) XDD transfers between identical testbed sites with XFS file systems mounted on SSD storage, (b) file copies between identical sites with Lustre file systems extended with LNet routers, and (c) Globus transfers between various pairs of DOE production sites, each with its own WAN connectivity, local network architectures, network interfaces, file system (e.g., Lustre or GPFS), and storage system. In the first two cases, communications



(a) XDD transfers: XFS mounted over SSD

(b) LNet-routed Lustre

(c) Production Globus transfers with throughput over 100 Mbps

**Fig. 1: Measurements used for throughput profiles $\hat{\Theta}_E$ of file transfers. The concave profile in near-optimal case of XDD transfers in (a) and the convex profile in (b) due to LNet router limits are both discernable. For Globus transfers over production infrastructure, the site differences lead to larger variations with respect to RTT.**

occur over dedicated 10GigE connections, for $\tau \in [0, 366]$ ms; in the third case, they are over the production ESnet infrastructure with $\tau \in [0, 105]$ ms. Figs. 1a-1c show profiles for these three scenarios. We observe that the two profiles over dedicated connections, (a) and (b), are quite different in their peak throughputs (10 and 4.5 Gbps) and their *concave* and *convex* shapes, respectively. While peak throughput is a direct indicator of performance, the concave-convex geometry is a more subtle indicator [19, 30]: a concave (convex) profile indicates intermediate-RTT throughput higher (lower) than linear interpolations. The third profile, (c), in contrast, is highly non-smooth, where each point represents measurements between a pair of DOE sites. From an infrastructure throughput perspective, a smooth and concave profile similar to Fig. 1a is desired, which is achieved by (i) enhancing and optimizing sites so that their profiles closely match, thereby making the infrastructure profile smooth, and (ii) selecting TCP version and transport method parameters to make the profile as concave as possible.

We present operators and coefficients of profiles for a part or version $\mathcal{S}'$ of infrastructure $\mathcal{S}$ to assess its component combinations:

- *Profile Calculus*: The projection and difference operators provide profiles $\Theta_A^{\mathcal{S}'}$ corresponding to individual sites and their collections, and to the separate contributions of TCP and file transfer tools.
- *Performance Coefficients*: Coefficients of a profile $\Theta_A^{\mathcal{S}'}$ capture the overall utilization and the extent of concavity-convexity, and the combined overall effects of those two elements.

We propose the *utilization-concavity coefficient* $\mathcal{C}_{UC} \in [0, 1]$, a scalar metric that captures both the peak throughput and the concave region of $\hat{\Theta}_A^{\mathcal{S}'}(\tau)$, and thus enables an objective comparison of different versions $\mathcal{S}'$ of $\mathcal{S}$. By combin-

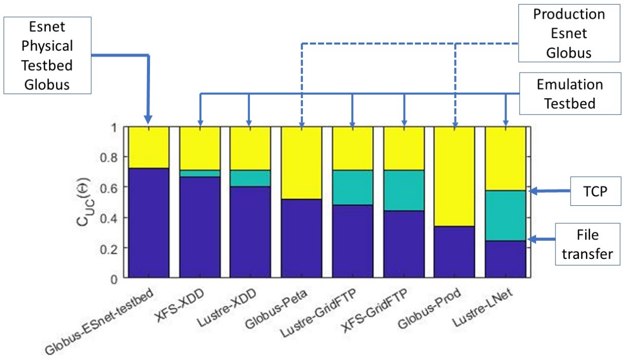


**Fig. 2: Summary of $\mathcal{C}_{UC}$ for eight file transfer configurations. In each case, the lower (blue) region is measured disk-to-disk performance; the middle (green) region is additional throughput achieved by TCP memory transfers (when measurements are available); and the upper (yellow) region is network capacity not used by transfers.**

ing the profiles and coefficients of throughput measurements from testbeds and production deployments, we analyze the performance of data transfer infrastructures in terms of current and newer components. A summary of our analytics is illustrated in Fig. 2 with $\mathcal{C}_{UC}$ for eight different file transfer configurations, wherein the performance decreases from left to right. The two left-most bars are for Globus transfers over the ESnet testbed DTNs for 0–150 ms connections, and XDD transfers over dedicated, emulated 0–366 ms connections (data from Fig. 1a). They represent near-optimal configurations. The right-most bar represents the worst case corresponding to Lustre LNet transfers over emulated infrastructure (data from Fig. 1b). The performance of Globus transfers over production infrastructure lies in between but below 0.5, indicating potential for site improvements reflected by $\mathcal{C}_{UC}$'s of individual sites in Section 4.2. In general our analytics lead to the selection and performance optimizations of various sites and parts of the infrastructure. These measurements have been collected over a five-year period to cover various testbed configurations and log production transfers; individual $\mathcal{C}_{UC}$ computations typically use 10 measurements at each RTT collected within a few hours. In this paper, we only present summaries that highlight the significant roles of (i) individual sites and sub-infrastructures, (ii) buffer sizes, IO limits and parallelism in TCP, GridFTP, XDD and LNet routers, and (iii) TCP versions and file transfer methods and their parameters.

The organization of this paper is as follows. Testbed and Globus infrastructure used for measurements are described in Section 2. In Section 3, we present throughput profiles for various scenarios, and describe their operators and coefficients. Throughput measurements and analytics are presented in Section 4. Related work is described in Section 5 and we conclude the paper in Section 6.

## 2 Testbed and Production Infrastructures

The analyses performed in this paper use a mix of log data for transfers performed by the production Globus service and measurements performed in various testbed environments.

For the **Globus log data**, we focus on transfers among the six sites shown in Fig. 3, each of which has one or more Globus-enabled DTNs, a high-speed ESnet connection, and various other systems deployed, for example, Lustre and GPFS file systems at OLCF and ALCF, respectively. This dataset thus comprises performance data for many transfers with different properties (e.g., number and size of files) and end system types, performed at different times.

To enable more controlled studies in a similar environment, we also perform experiments on an **emulation testbed** at ORNL comprising 32-core (`feyn1`–`feyn4`, `tait1`, `tait2`) and 48-core (`bohr05`, `bohr06`) Linux workstations, QDR IB switches, and 10 Gbps hardware connection emulators. We conduct experiments in which hosts with identical configurations are connected in pairs while RTT is varied from 0 to 366 msec. We include the 366 ms RTT case to represent a connection spanning the globe, as a limiting case. We perform memory-to-memory TCP throughput measurements with *iperf* [15] and measure the performance of other transfer tools by running them on the end system computers. Typically, we use 1–10 parallel streams for each configuration, set TCP buffer sizes to the largest value allowed by the host kernel to avoid TCP-level performance bottlenecks (resulting in the allocation of 2 GB socket buffers for iperf), and repeat throughput measurements 10 times. The emulation testbed also includes a distributed Lustre file system supported by eight OSTs connected over an IB QDR switch. Host systems (`bohr*`, `tait*`) are connected to the IB switch via Host Channel Adapter (HCA) and to Ethernet via 10 Gbps Ethernet NICs. In addition, our SSD drives are connected over PCI buses on `bohr05` and `bohr06`, which mount local XFS file systems. We also consider configurations in which Lustre is mounted over long-haul connections using LNet routers on `tait1` and `bohr06`; in that case, we use IOzone [1] for throughput measurements for both site and remote file systems.

We also study a **petascale DTN network** comprising the Globus endpoints and associated DTNs at ALCF, the National Center for Supercomputing Applications (NCSA), NERSC, and OLCF. As shown in Table 1, these endpoints differ in their configurations, but each operates multiple DTNs (compute systems dedicated for data transfers in distributed science environments [21]) to enable high-speed data transfers and all are connected at 100 Gbps. We conduct experiments on the petascale DTN network transferring a portion of a real science data generated by cosmology simulations [12]. The dataset consists of 19,260 files totaling 4.4 TB, with file sizes ranging from 1 byte to 11 GB.

Finally, we conduct experiments between the **ESnet testbed data transfer nodes** [2] attached to the production production ESnet network. These DTNs are deployed at NERSC, Chicago, New York and CERN, and are primarily intended for performance testing. Together, these testbed and production infrastructures provide flexible configurations under which various combinations

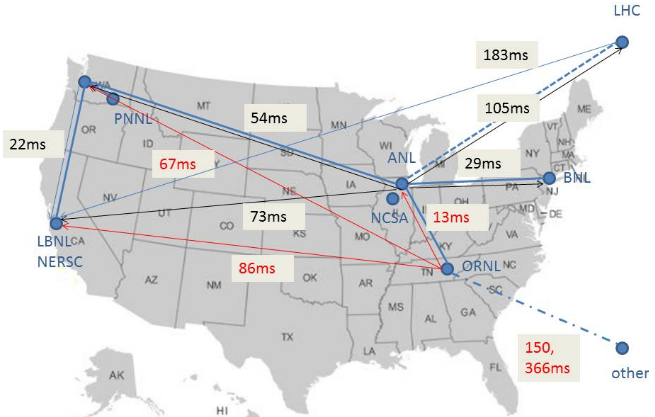


**Fig. 3: File transfer infrastructure, showing the six Globus sites and their connections, in blue, plus other connections (in red) considered in emulation studies. The sites are the Argonne Leadership Computing Facility (ALCF) at Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (LBNL), Pacific Northwest National Laboratory (PNNL), Oak Ridge Leadership Computing Facility (OLCF) at Oak Ridge National Laboratory (ORNL), and Large Hadron Collider (LHC) at CERN in Europe.**

**Table 1: Network configurations at the four petascale DTN network sites. All have 100 Gbps WAN connectivity**

| Institution | ALCF | NCSA | NERSC | OLCF |
|---|---|---|---|---|
| No. of DTNs | 12 | 23 | 9 | 8 |
| Filesystem | GPFS | Lustre | GPFS | Lustre |

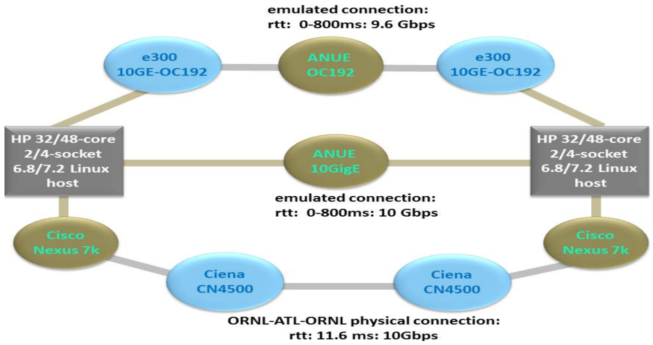


**Fig. 4: Physical and emulated connections between hosts**

of file systems, TCP versions, XDD, and GridFTP, and their parameters can be assessed. The many combinations resulted in the collection of several Terabytes of measurements data, which we analyze in this paper.

## 3 Profiles and Coefficients

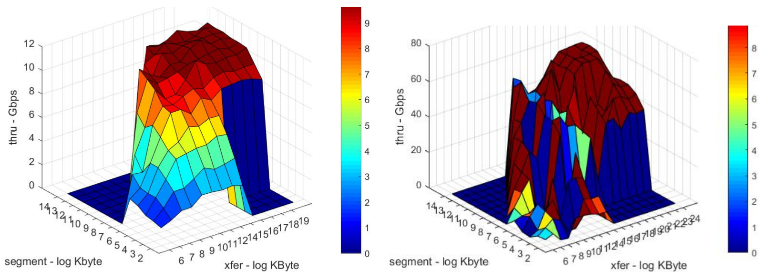
We consider an infrastructure $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, where each site $S_i$ comprises file system $F_i$ and transfer host $H_i$, and is connected to other sites $S_j, j \neq i$. For example, $N = 7$ for the emulation scenario with sites, ANL, BNL, NERSC, PNL, ORNL, CERN and round-the-earth site, in Fig. 3. Let $S_{\tau_i}$ denote the set of RTTs of connections used by site $S_i$ to support transfers, and $\mathcal{S}_\tau$ is set of all connections of all sites of $\mathcal{S}$. The *throughput profile* $\hat{\Theta}_A^{\mathcal{S}}(\tau)$ is generated by using measurements at selected RTT $\tau_{i,j} \in \mathcal{S}_\tau$ between sites $S_i$ and $S_j$ and "extended" to other $\tau$, for example, using measurements to additional client sites and computational methods such as piece-wise linear extrapolation. This concept generalizes the throughput profiles to infrastructures with multiple sites from its previous use for single client-server connections [32].

### 3.1 Variety of Throughput Profiles

The throughput profile $\hat{\Theta}_A^{\mathcal{S}}(\tau)$ is a complex composition of profiles of host and file systems, TCP over network connections, and file transfer software, which may vary across the sites due to their choices and configurations.

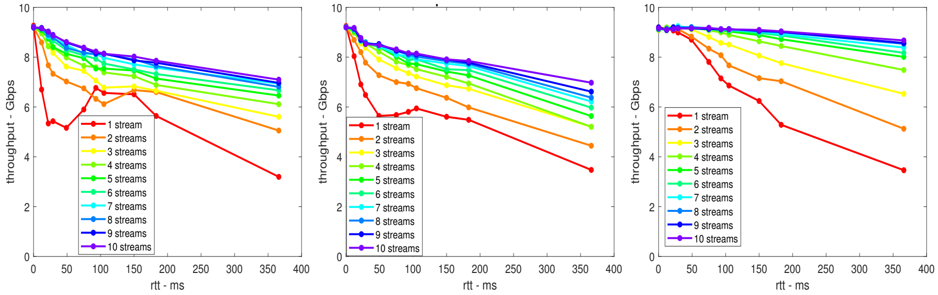
**Host and File IO Profiles** IO profiles of distributed Lustre and GPFS file systems at OLCF and ALCF with peak write rates of 12 and 60 Gbps, are shown in Figs. 5a and 5b, respectively. Such variations in peak rates will be reflected in file transfer throughput over connections with these file systems as end points, typically through complex relationships based on the underlying rate limiting factors. When limited by disk or HCA speeds, they represent peak file transfer throughput which is not improved by parallel streams; in particular, they expose throughput limits for large transfers that exceed the host buffers, since in some cases smaller transfers are handled between buffers at a higher throughput. On the other hand, they represent per-process limits in the case of distributed file systems such as Lustre. File transfer throughput then can be improved significantly by multiple stripes and parallel streams, for example by using higher CC (number of files transferred concurrently) and P (number of TCP streams per file) parameters [21], respectively, for GridFTP.

**TCP Profiles** We show throughput profiles for CUBIC, H-TCP, and BBR in Fig. 6 for largest allowed buffer sizes on the hosts that are connected by SONET connections. In all cases, more streams lead to higher throughput, and H-TCP, which is currently deployed in DTNs, performs better than CUBIC (Linux default). Recently developed BBR outperforms others with less than 5% decrease in parallel throughput for 366 ms connections, compared to more than 10% decrease of H-TCP.



(a) OLCF Lustre (b) ALCF GPFS

**Fig. 5: Throughput profiles of four file systems used in our experiments**

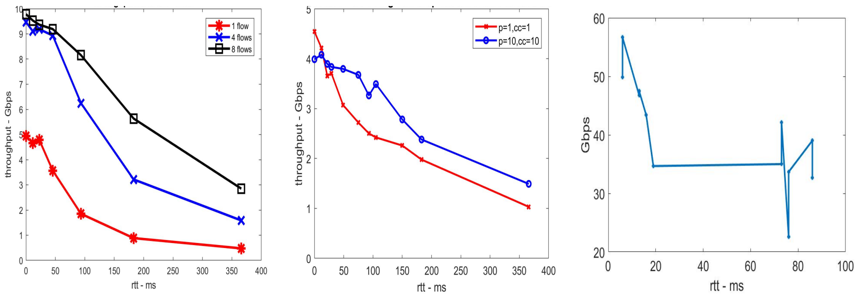


(a) CUBIC (b) H-TCP (c) BBR

**Fig. 6: Throughput profiles for four TCP versions between `feyn3` and `feyn4` over SONET**

**File Transfer Profiles** File transfer profiles vary significantly based on the file systems at source and destination and the transfer tool used to move data between sites. Transfer performance using XDD matches TCP throughput with the XFS file system on SSD, as shown in Fig. 7a, whereas GridFTP transfers between Lustre and XFS achieve much lower throughput, as shown in Fig. 7b, in part due to Lustre limits. Figs. 1c and 7c illustrate profiles for Globus transfers on the DOE infrastructure and Petascale DTN testbed, respectively. The latter uses DTNs with optimized configurations, whereas the former features more diverse DTNs. Overall, we see in Figs. 7a–7c that as we move from controlled and dedicated testbeds to heterogeneous, shared production systems, profiles become more complex.

The infrastructure profiles show significant variations and complex dependencies on the component systems, which motivates the need for simple measures that capture the overall performance, even if further analysis requires deeper investigations. Despite the variability, these profiles also satisfy important stability properties with respect to RTT in emulations in that profiles of its smaller subsets provide reasonable approximations; as illustrated in Fig. 8, a profile with five RTTs is within 2% of that with 11 RTTs. Consequently, we choose smaller

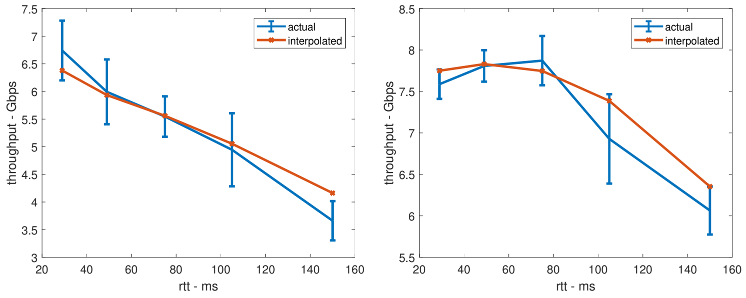


(a) XDD on XFS with SSD

(b) GridFTP: Lustre – XFS

(c) Petascale DTN testbed transfers

**Fig. 7: Throughput profiles of file transfers.**



(a) 1 stream

(b) 5 streams

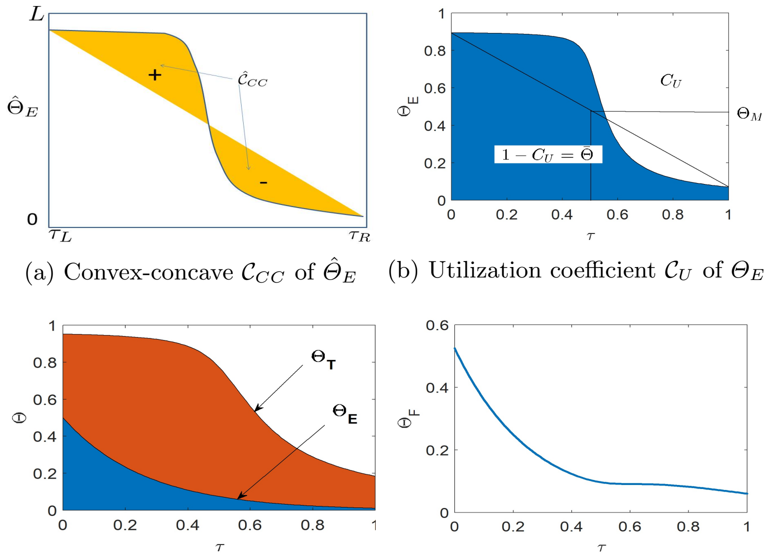
**Fig. 8: Profiles constructed from measurements performed on the `bohr5`–`10GigE`–`bohr6` emulation testbed at 11 RTT values (blue lines, with error bars) vs. profiles constructed from measurements at just five RTT values (red lines).**

representative sets of RTTs out of 28 possible values in the analysis of emulation scenarios, which significantly reduces the measurement collection time.

### 3.2 Profile Calculus

Consider an infrastructure $\mathcal{S}'$ whose sites have the same configurations as those of $\mathcal{S}$, for example, its sub-infrastructure. The *projection operator* $\mathcal{R}$, generates throughput values for $\mathcal{S}'_\tau$, given those for $\mathcal{S}_\tau$ such that $\mathcal{R}\left(\hat{\Theta}^{\mathcal{S}}, \mathcal{S}'\right) = \hat{\Theta}^{\mathcal{S}'}$. This operator can be used to infer a profile for an individual site in $\mathcal{S}$, $\hat{\Theta}^{\mathcal{S}_i} = \hat{\Theta}^{\{\mathcal{S}_i\}}$, and also for a future site $S_C$ of $\mathcal{S}$ as $\hat{\Theta}^{\{\mathcal{S}_C\}}$ based on the RTTs of the new site's connections.

We define the *difference operator* for two profiles $\hat{\Theta}_1$ and $\hat{\Theta}_2$ as $\left(\hat{\Theta}_1 \ominus \hat{\Theta}_2\right)(\tau) = \hat{\Theta}_1(\tau) - \hat{\Theta}_2(\tau)$. By using different profiles, this operator can be used to provide, for example, an incremental profile of a file transfer tool, $F$, as $\Theta_F = \Theta_{E|T} =$



(a) Convex-concave $\mathcal{C}_{CC}$ of $\hat{\Theta}_E$ (b) Utilization coefficient $\mathcal{C}_U$ of $\Theta_E$

(c) Network $\Theta_T$ and end-to-end $\Theta_E$(d) Transfer $\Theta_F(\tau) = \Theta_E(\tau)/\Theta_T(\tau)$

**Fig. 9: Throughput profiles and coefficients**

$\hat{\Theta}_T \ominus \hat{\Theta}_E$. When used with $\hat{\Theta}_T$ for different TCP versions, it characterizes the effectiveness of file transfer tool $F$ under a given TCP version. In ideal cases, $\Theta_{E|T}$ is close to the zero function as in the case of XDD transfers using XFS on SSD as shown in Fig. 7a.

### 3.3 Utilization and Convex-Concave Coefficients

Let $L$ represent the connection capacity, and $\tau_L$ and $\tau_H$ denote the smallest and largest RTTs, respectively, of the infrastructure. Then, we define the *under utilization coefficient* of $\hat{\Theta}$ as $\mathcal{C}_U(\hat{\Theta}) = \int_{\tau_L}^{\tau_R} \left( L - \hat{\Theta}(\tau) \right) d\tau$. By applying to memory and file transfer throughput, we have $\mathcal{C}_U(\hat{\Theta}_T)$ and $\mathcal{C}_U(\hat{\Theta}_E)$ that represent the unused connection capacity by TCP and the file transfer method, respectively, as shown in Fig. 2 for emulation testbed configurations. Then, the file transfer method $\mathcal{C}_U(\hat{\Theta}_T \ominus \hat{\Theta}_E)$ captures its effectiveness by using TCP profile as a baseline.

The convex and concave properties of $\hat{\Theta}$ are specified by the area above and below the linear interpolation of $\hat{\Theta}(\tau_L)$ and $\hat{\Theta}(\tau_R)$, respectively. This area is positive for a concave profile and negative for a convex profile. We define this area as the *convex-concave coefficient* of $\hat{\Theta}$, as illustrated in Fig. 9a, and it is given by

$$\begin{aligned}\mathcal{C}_{CC}\left(\hat{\Theta}\right) &= \int_{\tau_L}^{\tau_R} \left( \hat{\Theta}(\tau) - \left[ \hat{\Theta}(\tau_L) + \frac{\hat{\Theta}(\tau_R) - \hat{\Theta}(\tau_L)}{\tau_R - \tau_L} \tau \right] \right) d\tau \\ &= (\tau_R - \tau_L) \left[ \bar{\hat{\Theta}} - \hat{\Theta}_M \right].\end{aligned}$$

Let $\tilde{\Theta} : [0,1] \mapsto [0,1]$ denote a normalized version of $\hat{\Theta}$ such that throughput values are scaled by $L$, and the operand $\tau$ is translated and scaled from interval $[\tau_L, \tau_R]$ to $[0,1]$. We now combine both utilization and convex-concave properties and define the *utilization-concavity coefficient* as

$$\mathcal{C}_{UC}\left(\hat{\Theta}\right) = \frac{1}{2}\left(\left[1 - \mathcal{C}_U\left(\tilde{\Theta}\right)\right] + \left[\frac{1}{2} + \mathcal{C}_{CC}\left(\tilde{\Theta}\right)\right]\right).$$

It takes a much simpler form $\mathcal{C}_{UC}\left(\hat{\Theta}\right) = \bar{\tilde{\Theta}} - \tilde{\Theta}_M/2 + 1/4$ such that $\bar{\tilde{\Theta}}$ is the average and $\tilde{\Theta}_M/2$ is throughput at midpoint, which are closely related. The variations with respect to RTT in profiles, as observed in production Globus transfers infrastructures in Fig. 1c, lead to lower $\bar{\tilde{\Theta}}$ and hence lower $\mathcal{C}_{UC}$. These variations are due to differences in site systems which lead to lower $\mathcal{C}_{UC}$ compared to smoother emulation profiles of infrastructures with identical host systems in Fig. 7a, namely, all with XFS on SSD. However, smoother profiles can also be achieved for transfers between distributed Lustre and XFS on single SSD device as shown in Fig. 7b for CC=10 for GridFTP.

### 3.4 File Transfer Method Profiles

Let us assume that we have performed performance measurements for both memory-to-memory and disk-to-disk transfers over some infrastructure, and normalized the results to [0,1]. We may see something like the network profile $\tilde{\Theta}_T(\tau)$ and file transfer profile $\tilde{\Theta}_E(\tau)$ shown in Fig. 9c. The gap between the two profiles illustrates the effects of the file system, storage system, and file transfer method. Since $\tilde{\Theta}_E(\tau)$ encompasses TCP profile $\Theta_T(\tau)$ and host file IO limits, its lower values may be due to file IO throughput limits and not necessarily due to how well the transfer method, for example, manages IO-network transfer buffers. In particular, lower values of $\tilde{\Theta}_E(\tau)$ could be due to lower file IO throughput of the file system $\hat{\Theta}_{F_i}$ and/or host system $\hat{\Theta}_{H_i}$ at site $S_i$. We can assess the performance of the file transfer method $F$ itself by suitably normalizing with respect to the site profiles. If site throughputs are higher, we define the throughput profile of the file transfer part as $\tilde{\Theta}_F(\tau) = \tilde{\Theta}_E(\tau)/\tilde{\Theta}_T(\tau)$, for $\tau \in [0,1]$, as illustrated in Fig. 9d. We consider that TCP profile is a non-increasing function of $\tau$. When file or host system at a site limits TCP throughput most among the sites such that $\min\{\hat{\Theta}_{H_i}, \hat{\Theta}_{F_i}\} < \hat{\Theta}_T(0)$, we define

$$\tilde{\Theta}_F(\tau) = \frac{\tilde{\Theta}_E(\tau)L}{\tilde{\Theta}_T(\tau)\min\{\hat{\Theta}_{H_i}, \hat{\Theta}_{F_i}\}}, \tag{2.2}$$

for $\tau \in [0,1]$. Then, the utilization-concavity coefficient of the file transfer part is given by $\mathcal{C}_{UC}\left(\hat{\Theta}_F\right)$.

## 4 Throughput Profiles and Analytics

Production deployments of data transfer infrastructure spanning multiple sites could be quite heterogeneous; for example, DOE infrastructure employs different file systems (Lustre and GPFS), protocols (H-TCP on DTNs and CUBIC on cluster nodes), network connections (10GigE nd 40GigE) and file transfer software (GridFTP, bbcp and others). The projection and difference operations enable us to extract the profiles and coefficients of parts of the infrastructure its profile estimated from measurements. However, they are limited to the existing components, and assessment of other, in particular, newer components is not practical since the sites are independently operated and maintained. Our approach combines emulations, testbeds, and production infrastructures to support these tasks. We use four types of resources to generate datasets that drive the analytics, as described in Section 2.

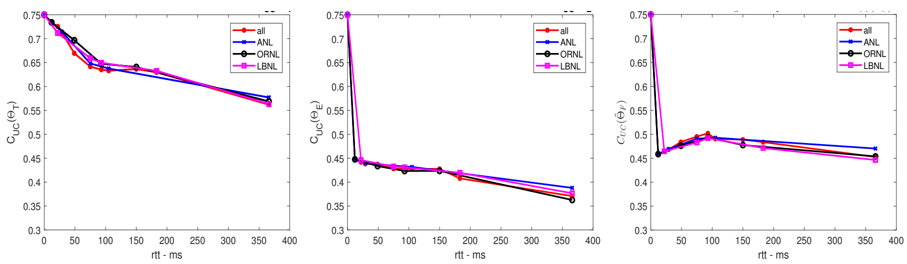
The emulation testbed enables broad and flexible configurations but only a limited reflection of production aspects. In contrast, Globus measurements provide a true reflection of production transfers, but offer limited scope for testing potentially disruptive technologies such as LNet-routed Lustre system. The ES-net DTNs offers more flexible configurations but its connections are limited by its much smaller footprint, and it does not support cluster nodes and Lustre file system. The Petascale experiments are constrained by the site systems and footprint of current infrastructure but offers several optimizations to select DTNs. By combining results from all four systems, we gain new insights into current and future data infrastructures, as discussed subsequently in this section.

### 4.1 Testbed Measurements

The dedicated ORNL testbed and its support for Lustre and for XFS on SSD enables us to test specialized scenarios such as BBR-enabled infrastructure, LNet-based wide-area extensions of Lustre, and data transfers between DTN-class hosts and cluster nodes.

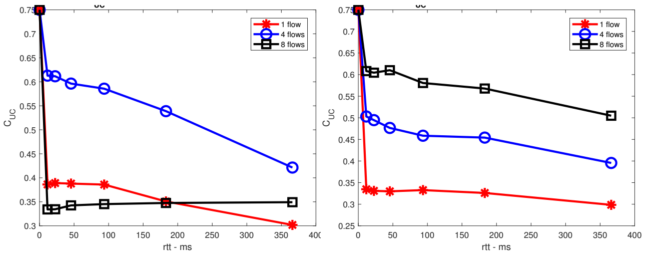
**Site Profiles** The $\mathcal{C}_{UC}$ values for TCP and GridFTP file transfer throughput for the `bohr5—10GigE—bohr6` configuration with H-TCP and 10 parallel streams are shown in Figs. 10a and 10b respectively. In addition to infrastructure profiles (labeled "all"), we also extract site-specific transfer throughput performance (i.e., transfers to/from a particular site) for ANL, ORNL, and LBNL, and show their individual $\mathcal{C}_{UC}$ values in these figures. We now focus on GridFTP software component at each site. The normalized filesystem throughput profile, namely, the ratio of the end-to-end file transfer and iperf throughput in Eq. (2.2), characterizes the performance of GridFTP specific to sites, and their $\mathcal{C}_{UC}(\hat{\Theta}_F)$ plots are shown in Fig. 10c. We see that these individual site $\mathcal{C}_{UC}$ curves closely match the infrastructure curve in all three cases, which is an artifact of the site systems being similar in the testbed.

**File Transfers** We consider transfers similar to those plotted in Fig. 7a but with Lustre filesystem, where two stripes are used for default and direct IO options. Fig. 11 plots the $\mathcal{C}_{UC}(\hat{\Theta}_E)$ values for both options with 1, 4, and 8 flows. It is interesting to note that while for direct IO Lustre, the transfer performance



(a) iperf: $\mathcal{C}_{UC}(\hat{\Theta}_T)$ (b) GridFTP end-to-end: $\mathcal{C}_{UC}(\hat{\Theta}_E)$ (c) GridFTP filesystem: $\mathcal{C}_{UC}(\hat{\Theta}_F)$

**Fig. 10: $\mathcal{C}_{UC}$ plots for `bohr5`–`10GigE`–`bohr6` configuration with H-TCP and 10 parallel streams: Overall and site profiles**



(a) Default IO (b) Direct IO

**Fig. 11: $\mathcal{C}_{UC}(\hat{\Theta}_E)$ for XDD transfers with two stripes and direct IO Lustre**

improves with higher flow counts, as evidenced by the higher $\mathcal{C}_{UC}(\hat{\Theta}_E)$ curves, the default IO Lustre option does not share the same characteristics: using 4 flows yields the best performance, which has also been observed in [31].

Finally, we plot $\mathcal{C}_{UC}$ of profiles for TCP, end-to-end transfer, and file transfer mechanism for eight-stripe LNet Lustre configuration in Fig. 12. The underlying LNet file transfer mechanism is significantly different from GridFTP and XDD software, and its throughput is limited by LNet peer credits which in turn results in lower and more convex profile compared to the above XDD cases. Indeed, the $\mathcal{C}_{UC}(\hat{\Theta}_E)$ values are lower and drop to below 0.4 even at lower RTTs, reflecting the inferior LNet performance. Here, $\mathcal{C}_{UC}(\hat{\Theta}_F)$ for LNet component is obtained using the difference operator, which shows the use of calculus to estimate its effect on infrastructure throughput, as discussed in Section 3.4.

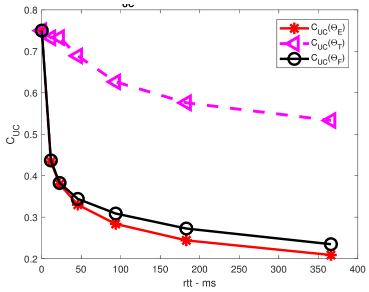
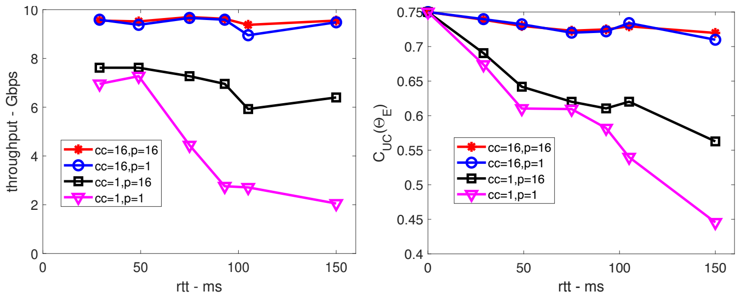


**Fig. 12: $\mathcal{C}_{UC}$ values for measurements of Lustre file copies**



(a) Throughput (b) $\mathcal{C}_{UC}(\hat{\Theta}_E)$

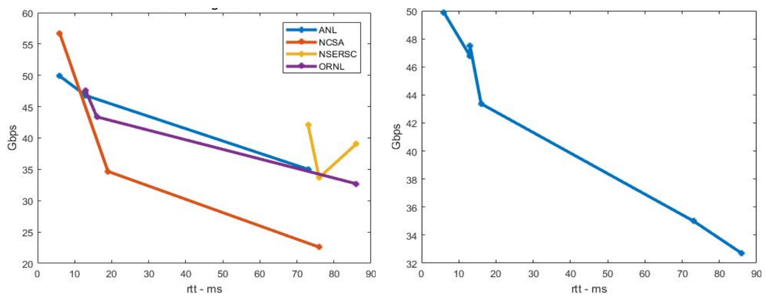
**Fig. 13: ESnet testbed DTNs: Globus file transfer throughput and corresponding $\mathcal{C}_{UC}(\hat{\Theta}_E)$ profiles for different concurrency ($CC$) and parallelism ($P$) values.**

**ESNet DTNs** For ESnet testbed, we consider Globus file transfers among DTNs located at four sites. We vary both the concurrency $CC$ and parallelism $P$ values within $\{1, 2, 4, 8, 16\}$ with a total of 25 possible $CC$ and $P$ combinations for each of the 16 connections. Fig. 13a shows throughput profiles for four selected configurations with smallest and largest $CC$ and $P$ values. The configuration with $CC = 1$ exhibits much reduced throughput compared to that with $CC = 16$, whereas in the former case, using $P = 16$ leads to significantly higher throughput compared to using $P = 1$, especially for higher RTTs. The $\mathcal{C}_{UC}(\hat{\Theta}_E)$ plots in Fig. 13b also show a superior performance with higher $CC$ values, as evidenced by the nearly level $\mathcal{C}_{UC}$ values with increasing RTT, in stark contrast to the precipitous drop in $\mathcal{C}_{UC}$ when $CC = P = 1$.

### 4.2 Globus Measurements

We processed Globus logs from the four Petascale DTN sites to obtain throughput profiles, and computed $\mathcal{C}_{UC}$ values in Fig. 14(c) with 100Gbps network capacity. The profiles of individual sites are shown in Fig. 14a, whose profiles differ both in peak throughput and rate of decrease; they all have convex profiles with $\mathcal{C}_{UC}$ values below 0.5. The infrastructure profile shown in Fig. 7c interestingly has a higher $\mathcal{C}_{UC}$ value compared to the sites. The convexity of these profiles indicates that network flows have not reach their full rates and indeed the throughput is limited by site IO or file systems). Among the four sites, ANL and ORNL have file systems with the highest measured throughput (Fig. 5), and the profile of ANL-ORNL sub-infrastructure is shown in Fig. 14b; although still convex, its profile is smoother and its $\mathcal{C}_{UC}$ is higher than that of both sites.

The second set of Global logs are collected over ESnet production infrastructure for a wide variety of transfers that included DTNs and other servers from five sites ANL, BNL, NERSC, PNNL, and ORNL. Our interest is mainly in high transfer rates, and we computed profiles using transfers with rates of at least 100 Mbps as shown in Fig. 15. Unlike the previous case, not all servers have been configured for high throughput, which resulted in a wide range of



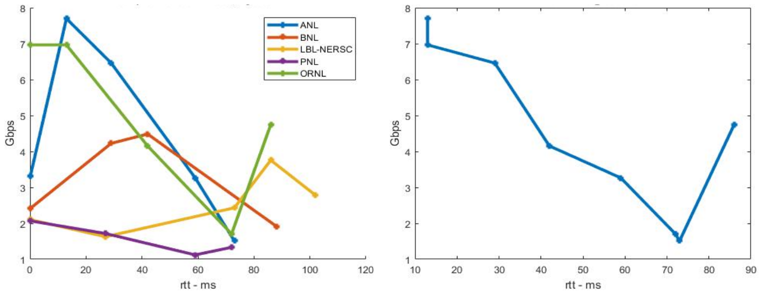
(a) site profiles

(b) ANL-ORNL profile

| 4 sites | ANL-ORNL | ANL | NCSA | NERSC | ORNL |
|---|---|---|---|---|---|
| 0.480 | 0.442 | 0.440 | 0.346 | 0.422 | 0.424 |

(c) $\mathcal{C}_{UC}$

**Fig. 14: Using Globus transfer log data to characterize 100 Gbps-connected Petascale DTNs at ANL, NCSA, NERSC, and ORNL. (a) Profiles for each site, based on average performance to each other site. (b) An aggregate profile based on just measurements from ANL and ORNL to each other site (including each other). (c) $\mathcal{C}_{UC}$ values for the aggregate of all sites, just the transfers in (b), and each site in (a).**



(a) Site profiles: $\geq$ 100 Mbps

(b) ANL-ORNL profile

| 5 sites | ANL-ORNL | ANL | BNL | NERSC | PNNL | ORNL |
|---|---|---|---|---|---|---|
| 0.343 | 0.459 | 0.653 | 0.436 | 0.414 | 0.358 | 0.551 |

(c) $\mathcal{C}_{UC}$

**Fig. 15: Using log data for Globus transfers of $\geq 100$ Mbps to characterize 10 Gbps-connected servers at ANL, BNL, NERSC, PNNL and ORNL. See Fig. 14 for a description of each subfigure.**

site profiles shown in Fig. 15a. To account for them in computing $\mathcal{C}_{UC}$, we use a lower connection capacity of 10 Gbps. This profile of the infrastructure with five sites is much more complex than its counterpart in testbed measurements, reflecting the more varied nature of site configurations; however, the overall profile decreases with RTT and is convex as indicated by $\mathcal{C}_{UC} = 0.343$. Among the sites, ANL and ORNL achieve higher throughput and contain concave regions as reflected by their $\mathcal{C}_{UC}$ values above 0.5. However, the profile of ANL-ORNL sub-infrastructure has additional concave regions as shown in Fig. 15b, and their combined profile has $\mathcal{C}_{UC} = 0.459$, which is lower than that of either site. Thus, the effect of site profiles on infrastructure profiles is the opposite of the previous case. The overall convexity of profiles indicates potential improvements in throughput by overcoming IO and file throughput limits and also improving the network throughput to achieve concave regions. It is possible to achieve broader concave network profiles with large buffers and more parallel streams but the IO and file throughput limits must also be mitigated to ensure higher and more concave infrastructure profiles for file transfers.

Relating these results with corresponding testbed emulations we make the following inferences about the production infrastructure:

(i) *Site Systems:* Smooth infrastructure profile indicates well-aligned sites, and the variations in profile indicate critical differences among the sites, which in turn lead to lower $\mathcal{C}_{UC}$. By enhancing all sites to match the top ones, smoother profiles and hence higher $\mathcal{C}_{UC}$ will be achieved; for example, in ANL-ORNL sub-infrastructure for Petascale scenarios between GPFS and Lustre file systems as shown in Fig. 14b. Such enhancements are needed for these Globus transfers, and emulations indicate that they are indeed feasible.

(ii) *Network Transfer Aspects:* The convex regions of profiles indicate buffer or IO limits which in turn prevent from full TCP flows that have concave profiles. In addition, file transfer methods that translate between IO and network flows lead to rate or buffer limits that in turn result in convex profiles, which can be overcome by matching parallel IO and TCP flows as illustrated in Fig. 7a with 8 flows. Again, our testbed emulations indicate that these improvements are feasible, for example by using BBR TCP and suitable CC and P values for GridFTP.

(iii) *Measurements:* The profiles and corresponding $\mathcal{C}_{UC}$ values are generated from Globus logs of transfers which are collected non-intrusively. They provide the above valuable information about sites and network transfers without needing extensive site instrumentation.

The inferences based on available Globus logs do not necessarily lead to the best possible performance improvements, which might indeed require measurements of transfers specifically designed to exercise the specific site configurations.

## 5 Related work

Parallel TCP flows are commonly used to transfer large data over wide-area networks. Hacker et al. [13] examine the effects of using parallel TCP flows to improve end-to-end network performance for distributed data-intensive applications. However, their experiments do not involve storage systems and thus only partially capture the factors determining end-to-end file transfer performance. To transfer ATLAS experiment data over a high RTT (∼290 ms) wide area network, Matsunaga et al. [25] test various combinations of GridFTP parameters, such as the number of parallel streams and TCP window size. They conclude that careful parameter optimization is needed for bulk data transfer, especially over high-RTT networks because default configurations are usually optimized for short RTTs and large RTTs lead to quite different behavior. Kosar et al. have also investigated the impact of such parameters and developed automated parameter selection methods [5, 39, 40]. In a different take on the modeling problem, Liu et al. [22] introduce methods for evaluating potential design points of a distributed multi-site infrastructure. Their use case demonstrates the benefits of building such an infrastructure, as well as the requirements of profiling it for better estimation of end-to-end data movement performance.

A complete understanding of a distributed infrastructure requires explanations for each individual subsystem and their interactions. Liu et al. [20] extract features for endpoint CPU load, NIC load and transfer characteristics, and use these features in linear and nonlinear transfer performance models. Some studies have focused on profiling of subsystems [19, 30, 31] including network, IO, and host systems. In particular, both Rao et al. [30] and Liu et al. [19] have investigated conditions under which the overall memory transfer throughput profile exhibits the desirable concave characteristic, whereas in Rao et al. [31] extensive XDD file transfer throughput performance is discussed. This paper extends the above concavity-convexity analysis to infrastructure data transfers for the first time, whose variations across sites lead to non-smooth profiles. These findings are important: although over the decades, several detailed analytical models have been developed and experimental measurements have been collected for various network transport protocols, e.g., several TCP variants, these conventional models [24, 36] provide entirely convex throughput profiles. Thus, they underestimate the throughput, and furthermore do not accurately reflect the superior TCP performance over linear interpolations, at least for smaller RTTs. Liu and Rao [18] apply similar metrics to describe memory data transfer performance for client-server connections using a suite of transport protocols but not to much more complex infrastructure-level file transfers.

For data transfer infrastructures, we know of no reliable analytical methods in the literature for characterizing the data transfer performance for bottleneck detection and accurate performance prediction. Largely motivated by recent proliferation of high-performance distributed computing and networking in scientific and commercial applications, this is our first attempt at filling the void in characterizing data transfer performance by using a single metric applicable to disparate infrastructures to compare them, and to the sites and file transfer mechanisms to pinpoint parts to be improved.

## 6 Conclusions

We have presented analytics of throughput measurements of wide-area data transfer infrastructures that support science and big data distributed computations. These measurements include both testbed and production infrastructures with the GridFTP and XDD file transfer tools, and the Lustre file system extended with LNet routers. The throughput measurements were quite varied due to the complexities of host, file, IO, and disk systems, and their interactions, which makes performance assessment and optimization of infrastructures or their parts challenging. We presented unifying analytics based on the convexity-concavity geometry of throughput regression profiles, and proposed the utilization-concavity coefficient that characterizes the overall transfer performance as a scalar in [0,1] range. The profiles and their coefficients extracted using measurements from structured testbeds and production infrastructures provide a high-level, summary performance assessment and also indicate potential areas for further deeper investigations. Our results also provided guidelines for performance optimizations by highlighting the significant roles of individual site configurations, and buffer sizes and utilization, and parallelism implemented by network protocols and file transfer methods.

Further investigations, including additional test configurations and examination of additional parameters, are needed to further improve throughput performance over shared connections. In addition to throughput considered here, other parameters and derived quantities may be studied for objective performance comparisons of varied configurations. Of particular importance are the quantities that indicate the levels of optimizations of a given configuration and provide insights for further investigations. It would be of future interest to extend the calculus of throughput profiles described here with a deeper focus on subsystems and broader aspects to encompass data streaming infrastructures.

## References

1. Iozone file system benchmark, 2018 (accessed March 28, 2018). `http://www.iozone.org`.
2. *Energy Science Network Data Transfer Nodes*, accessed March 28, 2018. `https://fasterdata.es.net/performance-testing/DTNs/`.
3. W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster. The Globus striped GridFTP framework and server. In *ACM/IEEE Conference on Supercomputing*, pages 54–64, Washington, DC, 2005. IEEE Computer Society.
4. Bryce Allen, John Bresnahan, Lisa Childers, Ian Foster, Gopi Kandaswamy, Raj Kettimuthu, Jack Kordas, Mike Link, Stuart Martin, Karl Pickett, and Steven Tuecke. Software as a service for data scientists. *Communications of the ACM*, 55(2):81–88, February 2012.
5. Engin Arslan and Tevfik Kosar. High speed transfer optimization based on historical analysis and real-time tuning. *IEEE Transactions on Parallel and Distributed Systems*, 2018.
6. Aspera transfer service, accessed March 28, 2018. `http://asperasoft.com`.

7. N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson. BBR: Congestion based congestion control. *ACM Queue*, 14(5), Dec. 2016.
8. K. Chard, E. Dart, I. Foster, D. Shifflett, S. J. Tuecke, and J. Williams. The modern research data portal: A design pattern for networked, data-intensive science. *Peer Journal of Computer Science*, 4(6), 2018.
9. Sally Floyd. HighSpeed TCP for large congestion windows. RFC 3649, IETF, 2003. `https://tools.ietf.org/html/rfc3649`.
10. General Parallel File System, `https://www.ibm.com/support/knowledgecenter/en/SSFKCN/gpfs_welcome.html`.
11. Y. Gu and R. L. Grossman. UDT: UDP-based data transfer for high-speed wide area networks. *Computer Networks*, 51(7), 2007.
12. Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope, and Katrin Heitmann. HACC: extreme scaling and performance across diverse architectures. In *International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '13, pages 6:1–6:10, New York, NY, USA, 2013. ACM.
13. T. J. Hacker, B. D. Athey, and B. Noble. The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network. In *16th International Parallel and Distributed Processing Symposium*, 2002.
14. R. Henschel, S. Simms, D. Hancock, S. Michael, T. Johnson, N. Heald, T. William, et al. Demonstrating Lustre over a 100Gbps wide area network of 3,500km. In *Internationl Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–8, Nov. 2012.
15. https://iperf.fr/. *iPerf - The ultimate speed test tool for TCP, UDP and SCTPs*, 2018 (accessed March 28, 2018). `https://iperf.fr/`.
16. S. Jain, A. Kumar, S. Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jonathan Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat. B4: Experience with a globally-deployed software defined WAN. *SIGCOMM Comput. Commun. Rev.*, 43(4):3–14, Oct. 2013.
17. Rajkumar Kettimuthu, Zhengchun Liu, David Wheelerd, Ian Foster, Katrin Heitmann, and Franck Cappello. Transferring a petabyte in a day. In *4th International Workshop on Innovating the Network for Data Intensive Science*, page 10, November 2017.
18. Q. Liu and N. S. V. Rao. On concavity and utilization analytics of wide-area network transport protocols. In *Proc. 20th IEEE Conference on High Performance Computing and Communications (HPCC)*, Exeter, U.K., Jun. 2018.
19. Q. Liu, N. S. V. Rao, C. Q. Wu, D. Yun, R. Kettimuthu, and I. Foster. Measurement-based performance profiles and dynamics of UDT over dedicated connections. In *International Conference on Network Protocols*. Singapore, Nov. 2016.
20. Zhengchun Liu, Prasanna Balaprakash, Rajkumar Kettimuthu, and Ian Foster. Explaining wide area data transfer performance. In *26th International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '17, pages 167–178, New York, NY, USA, 2017. ACM.
21. Zhengchun Liu, Rajkumar Kettimuthu, Ian Foster, and Peter H. Beckman. Towards a smart data transfer node. In *4th International Workshop on Innovating the Network for Data Intensive Science*, page 10, November 2017.
22. Zhengchun Liu, Rajkumar Kettimuthu, Sven Leyffer, Prashant Palkar, and Ian Foster. A mathematical programming- and simulation-based framework to evaluate cyberinfrastructure design choices. In *IEEE 13th International Conference on e-Science*, pages 148–157, Oct 2017.

23. Lustre Basics, `https://www.olcf.ornl.gov/kb_articles/lustre-basics`.
24. M. Mathis, J. Semke, J. Mahdavi, and T. Ott. The mascroscopic behavior of the TCP congestion avoidance algorithm. *Computer Communication Review*, 27(3), 1997.
25. H Matsunaga, T Isobe, T Mashimo, H Sakamoto, and I Ueda. Data transfer over the wide area network with a large round trip time. *Journal of Physics: Conference Series*, 219(6):062056, 2010.
26. Multi-core aware data transfer middleware, accessed March 28, 2018. `mdtm.fnal.gov`.
27. S. Michael, L. Zhen, R. Henschel, S. Simms, E. Barton, and M. Link. A study of Lustre networking over a 100 gigabit wide area network with 50 milliseconds of latency. In *5th International Workshop on Data-Intensive Distributed Computing*, page 4352, 2012.
28. On-demand Secure Circuits and Advance Reservation System. http://www.es.net/oscars.
29. N. S. V. Rao, N. Imam, J. Hanley, and O. Sarp. Wide-area Lustre file system using LNet routers. In *12th Annual IEEE International Systems Conference*, 2018.
30. N. S. V. Rao, Q. Liu, S. Sen, J. Henley, I. T. Foster, R. Kettimuthu, D. Towsley, and G. Vardoyan. TCP throughput profiles using measurements over dedicated connections. In *ACM Symposium on High-Performance Parallel and Distributed Computing*, Washington, DC, July-August. 2017.
31. N. S. V. Rao, Q. Liu, S. Sen, G. Hinkel, N. Imam, B. W. Settlemyer, I. T. Foster, et al. Experimental analysis of file transfer rates over wide-area dedicated connections. In *18th IEEE International Conference on High Performance Computing and Communications (HPCC)*, pages 198–205, Sydney, Australia, Dec. 2016.
32. N. S. V. Rao, Q. Liu, S. Sen, D. Towsley, G. Vardoyan, I. T. Foster, and R. Kettimuthu. Experiments and analyses of data transfers over wide-area dedicated connections. In *26th International Conference on Computer Communications and Network*, 2017.
33. I. Rhee and L. Xu. CUBIC: A new TCP-friendly high-speed TCP variant. In *3rd International Workshop on Protocols for Fast Long-Distance Networks*, 2005.
34. B. W. Settlemyer, J. D. Dobson, S. W. Hodson, J. A. Kuehn, S. W. Poole, and T. M. Ruwart. A technique for moving large data sets over high-performance long distance networks. In *IEEE 27th Symposium on Mass Storage Systems and Technologies*, pages 1–6, May 2011.
35. R. N. Shorten and D. J. Leith. H-TCP: TCP for high-speed and long-distance networks. In *3rd International Workshop on Protocols for Fast Long-Distance Networks*, 2004.
36. Y. Srikant and L. Ying. *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective.* Cambridge University Press, 2014.
37. XDD – The eXtreme dd toolset, accessed March 28, 2018. `https://github.com/bws/xdd`.
38. XFS, http://xfs.org.
39. Esma Yildirim, Engin Arslan, Jangyoung Kim, and Tevfik Kosar. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. *IEEE Transactions on Cloud Computing*, 4(1):63–75, 2016.
40. Esma Yildirim, Dengpan Yin, and Tevfik Kosar. Prediction of optimal parallelism level in wide area data transfers. *IEEE Transactions on Parallel and Distributed Systems*, 22(12):2033–2045, 2011.