SEAGATE

eBook

Enterprise Data Transfer Playbook

Find the Most Efficient Way to Transfer Mass Data for Your Business

Table of Contents

Introduction

This playbook explores a simple issue with extensive ramifications: enterprises often need a variety of mass data transfer technologies to help them address specific performance, scale, security, and cost challenges.

Whether organizations are using mass data sets for in-depth analytics, gathering dozens of terabytes a day of seismic survey results for oil exploration, or shooting hours of 4K video at a remote location, their operations rely on data. Organizations spend hundreds of millions of dollars making sure that the right data is accessible to the right users and services at the right time, for the right cost, with the right protections.

One-size-fits-all mass data transfer technologies aren't suitable for all requirements. The rapid growth of data, emerging use cases, narrowing data transfer windows, and heightened security concerns mean that organizations require an evolving, multifaceted toolkit for mass data transfer that offers added versatility and flexibility.

Depending on their requirements, an organization might use one mass data transfer technology for scenario A, another for scenario B, and a third for scenario C. Every week, we hear from customers that they need something different to augment and extend their mass-data transfer capabilities. What they have isn't offering the right balance of performance, capacity, simplicity, safety, and value. They're 50% or 80% of the way there, but they need help achieving essential goals.

Fortunately, technology companies are offering new approaches to mass-data transfer, offering capabilities to accommodate shifts in what's possible, desirable, and needed. This playbook will give a better understanding of the pros and cons of conventional enterprise data transfer technologies, discuss alternatives, and help identify suitable technologies for a range of distinctive requirements.

What's Changing with Enterprise Mass Data Transfer

Why do organizations need to move massive amounts of data? Data movement stems from the impact of four critical, intersecting areas:

1 No data left behind

Keep data available to power unanticipated opportunities

2 Analytics

Prescriptive, sifting petabytes for forward-looking decisions

3 Customer Experiences

Differentiated, personalized, everywhere

4 Artificial Intelligence

Powering, sustaining, accelerating digital transformation

1 No Data Left Behind

In many cases from the endpoint to the edge, to protect and move data over connections that might not be ideal, organizations have resorted to data modification technologies (like encryption) and data manipulation (compression, editing). However, there's a tendency, over time, for new requirements to emerge around data sets. Operational technologists often require access to raw data alongside the modified and manipulated data. Users may have to go back to the raw data to accommodate emerging needs. Examples include:

a. A last-minute requirement for uncompressed video to create a new scene in a movie.

b. Digging into a full-resolution IoT stream to identify a failure in a manufacturing line.

c. Going beyond preliminary analytics findings by analyzing a full data set instead of a sample.

More and more organizations across all verticals need two ways to move a data set—an initial data movement for modified, reduced data sets, and another method that allows every byte to be migrated back into server closets, edge computing nodes, and ultimately to data centers or clouds, without any data reduction or data loss. There's a balancing act between accelerating time-to-data access while ensuring no data is ever left behind.

2 Analytics

Analytics is central to the value of data. Organizations are shifting from descriptive analytics that explain what's happened to prescriptive analytics that empower organizations with accelerated access to forward-looking decision making. This shift is predictable on a foundation of fast access to massive data sets. Whether data is captured and stored at endpoints, at the edge, in the cloud, or in the data center, the ability to rapidly acquire, sift, and process massive (hundreds of terabytes to dozens of petabytes) data sets will become a new normal for forward-thinking organizations.

3 Customer Experiences

There's another area of rapid innovation around customer-facing experiences. Whether we're talking about smart cities or augmented reality shopping in the metaverse, organizations want to build unprecedented experiences on a foundation of massive data sets and distributed infrastructure. Differentiated, personalized experiences will become a distinguishing feature of leading-edge organizations. These require large data sets and higher performance computing near end users.

4 Artificial Intelligence

Finally, AI is already transforming the nature of business and IT. 61% of telecom providers are planning AI projects for edge computing because the reality of today's massive data sets, distributed infrastructure, increasing reliance on analytics, and desire to deliver differentiated experiences means that action can't be managed manually. AI will power meta-analysis from the web of sensors and collections of data we're acquiring today, giving enterprises access to connections they weren't looking for and new opportunities to apply the data. Distributed infrastructures will rely on decisions by code and actions by autonomous agents. The next generation of digital transformation will be powered by AI, sustained by AI, and made better by AI.

These four areas of opportunity and innovation can emerge, grow, and provide value because of two fundamental shifts in the IT landscape.

The first is—the world is shifting toward IT 4.0. For reasons that include performance, reliability, agility, flexibility, scale, and impact, IT infrastructure is moving away from a centralized model that includes a single data center or a few data centers or a data center and single public cloud combination, and instead moving towards a distributed IT ecosystem to support disruptive customer experience, analytics, and AI. IT 4.0 is characterized by:

- Multicloud and hybrid cloud.
- Data capture, processing, and storage at the edge.
- Data is constantly transferred from the edge to the core or specialized cloud for further processing and analysis.
- Updates and changes are pushed from cloud or core to edge.

This approach adds performance, versatility, flexibility and resilience to IT, and organizations that shift toward IT 4.0 have advantages over organizations that don't.

The second shift is—data sets are growing exponentially. Data interactions (creation, capture, copying, and utilization of data) have increased 5000% between 2010 and 2020.[1] During that time, data usage increased from 1.2 million petabytes to almost 60 million petabytes across all forms of utilization, including consumers and businesses.[2] IDC predicts that data creation and replication, across all applications, will experience a compound annual growth rate (CAGR) of 23% over the 2020-2025 forecast period.[3]

Of course, exponential data growth isn't exactly a problem. In fact, it's a good thing—more data means more opportunity for insight, customer engagement, and delivering compelling business outcomes. The technical problems caused by data growth can be resolved with the right technologies applied in the right way.

Data analytics, innovative customer experiences, and AI are all powered by the integration of massive data across distributed IT. And today, in the world of enterprise data transfer, new challenges are emerging because the right data needs to be in the right place at the right time, *and often…isn't.*

Data usage has increased almost

60

Million PB

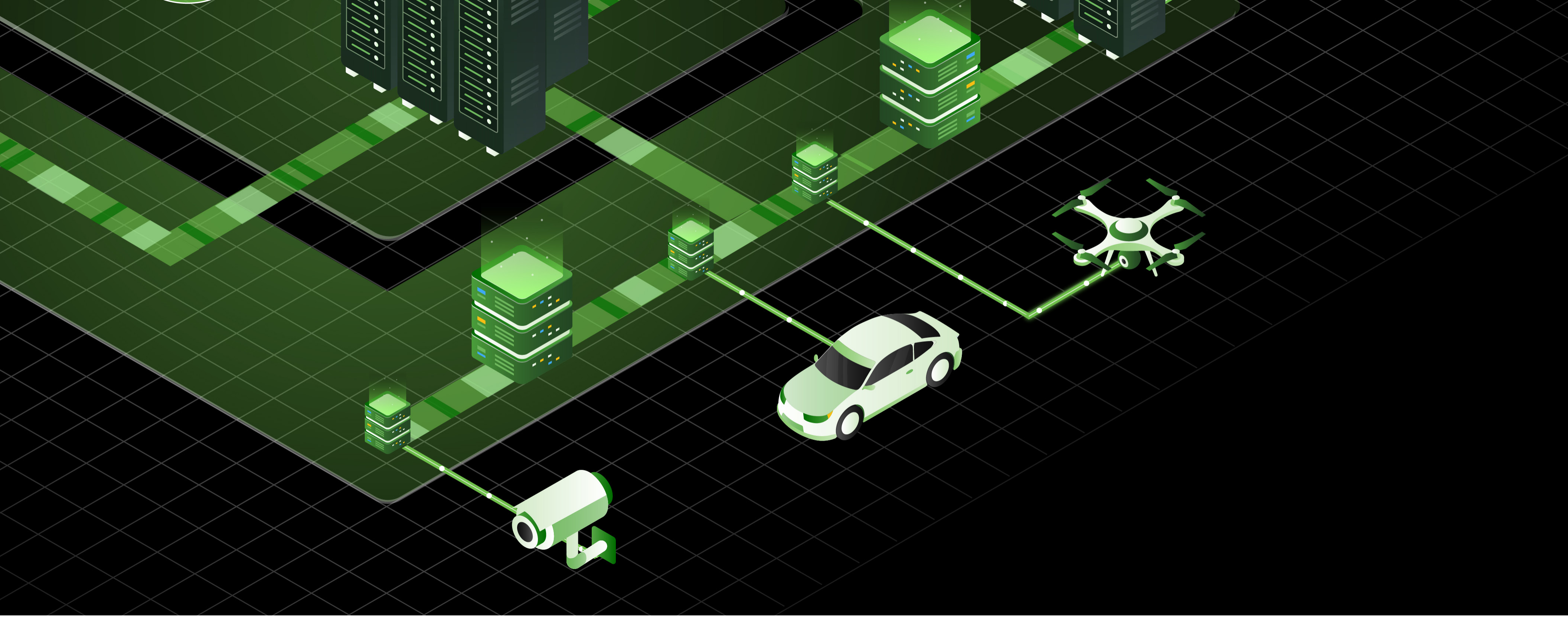
The Challenges of Mass Data Transfer

There are several critical considerations when it comes to mass data movement:

- **Data volume**—Will the data movement technology support the volume of data? Will it cope with increases in data set size as data sets grow over time?
- **No data ever left behind**—Every byte must be captured, stored, and made usable.
- **Transfer frequency**—Will the data movement technology support the frequency of data movement? An enterprise might have an hourly or daily requirement, a weekly requirement, or might only need to transfer data once a month.
- **Transfer speed**—Can the technology transfer data in a timely fashion? If there's a twelve-hour backup window for a data center, can the data be migrated from the edge data center quickly enough to coincide with the backup window?
- **Security**—Is the transfer technology sufficiently secure? Could it be hacked? Are there user/administrator access controls? Is data encrypted? Are any data transfer devices physically protected from threat, breach, or tampering? Do the devices offer secure erasure?
- **Cost**—Do the costs of the technology fit within budgets? Is it structured in a tolerable way (CapEx vs OpEx, short-term commitments vs long-term)? Are the costs easy to understand and manage?

When we're exploring data transfer across distributed infrastructures, it's helpful to think about the critical considerations listed above across three different categories:

1. **Endpoint to edge**
2. **Edge to core**
3. **Edge and core to cloud**



Endpoint to edge

Data transfer challenges and opportunities start with the endpoints—the devices organizations use in the trenches to present and gather data. Exploring the role of endpoints is critical for understanding the mass data transfer problem because, according to IDC, 65% of all data is created by endpoints. These could be mobile phones, IoT devices, video surveillance cameras, or anything else that creates, captures, and temporarily stores large amounts of data. They're transferring data, whether constantly or on a periodic basis, to edge infrastructure. Endpoints are connected to edge infrastructure through last mile connectivity, whether wired or wireless.

Endpoint

A data capture, storage and computing device that's localized, gathering data from a single data source or user as an endpoint on a network. E.g., IoT sensor, mobile phone.

Edge Computing

A data capture, storage, and computing device that gathers data from a few or many endpoints across a distributed network. E.g., IoT gateway, edge compute node, server closet.

Challenges around this requirement stem from the size of these data sets—some of which are so large, certain data movement technologies aren't able to support requirements, especially if raw data sets must be captured and transferred without editing, compression, or other modes of reduction.

It's common for endpoints to be connected to edge data centers by wireless networks—in the U.S. 4G or 5G. Wireless networks are widely available, aren't restricted to a single provider, encrypt data-in-motion for better security, and can be cost-effective, but come with challenges. Organizations considering wireless connectivity for endpoint-to-edge data transfers need to evaluate:

Speed—4G networks can, under controlled conditions, manage 50MB/s download speeds and 10MB/s upload speeds, but in the real world, 4G is able to deliver download speeds of 5-12MB/s and upload speeds between 2 and 5MB/s. To put that into context, a 5MB/s upload of a 1TB data set would take twenty-one days, eight hours, thirty-eight minutes, thirteen seconds. A 5G network with upload speeds of up to 58.1MB/s would allow the same data set to be transmitted in just under two days. These speeds can be adequate for some requirements and not others.

Consistency—As we all know, performance for wireless networks can be unpredictable or consistent. Weather, distance from a tower, and simple network congestion (too many devices using the same bandwidth) can cause performance degradation.

Reliability—Wireless networks can suffer from reliability issues. Building construction and line-of-sight obstacles can cause disconnections.

Distribution—As of December 2021, more than half of the country lacks access to fixed 5G wireless networks. Wireless connectivity may not be an option if endpoint devices aren't in a major city or situated along an interstate.

There are ways to compensate for these issues. Using multiple wireless providers and creating private wireless networks are two approaches that help mitigate the limits of wireless networks. Unfortunately, both approaches contribute to the last problem:

Cost—A single commercial 5G connection costs between $70-100/month. Hundreds of thousands of connections drive costs through the roof, can be hard to manage, and come with fixed charges regardless of whether endpoint devices are consuming bandwidth. Efforts to mitigate problems with speed, consistency, and reliability by adding backup wireless networks or private networks can drive up costs to astronomical levels.

In some circumstances, endpoints can be connected via low-Earth satellites with 0.25-1Gb/s class connections. These were tested in the 1990s and found lacking, but a handful of companies have recently begun deploying thousands of satellites aimed at building global communications networks in space. These networks are new, and so far, users have encountered:

Delays in receiving equipment—Both residential and commercial users have found it impossible to receive the satellite network transceivers quickly. One vendor has reported multi-month delivery delays due to silicon shortages that slowed production.

Consistency—Performance varies from vendor to vendor and from time to time, as dozens or hundreds of connections are provided by a single satellite, with variations in the number of connections and the traffic per connection. These shared connections don't offer predictable performance. In addition, heavy rain and snow can reduce bandwidth or even block satellite network connections entirely.

Speed—Connection speeds vary across vendors. One provider can reach real-world speeds of nearly 105MB/s while others oscillate around 20MB/s. The fastest upload speeds are around 22MB/s, which means that transferring a 1TB file would take weeks.

Distribution—For economic reasons, satellite network providers concentrate their efforts on the Northern hemisphere.

Costs—Pricing varies from vendor to vendor and requirement to requirement, ranging from $49.99 to $500 per connection. Satellite internet is the most expensive wireless connection available.

Due to the limitations of different types of wireless network connectivity, it's not unusual for endpoints to be connected by last-mile wired network infrastructure, which could be DSL, cable, coax, or fiber optics. Obviously, these are well-established technologies, widely available from a range of providers, and come (at least with commercial contracts) with the kinds of enterprise insight and management features organizations need.

Wired last-mile connectivity is useful as another technology for mass data transfer, but it isn't a panacea for endpoint-to-edge data transfer, as it has its limitations as well:

Access—Wired broadband isn't universally available. The FCC provides grants to encourage state and regional broadband networks for healthcare institutions simply because access in many areas is still lacking.

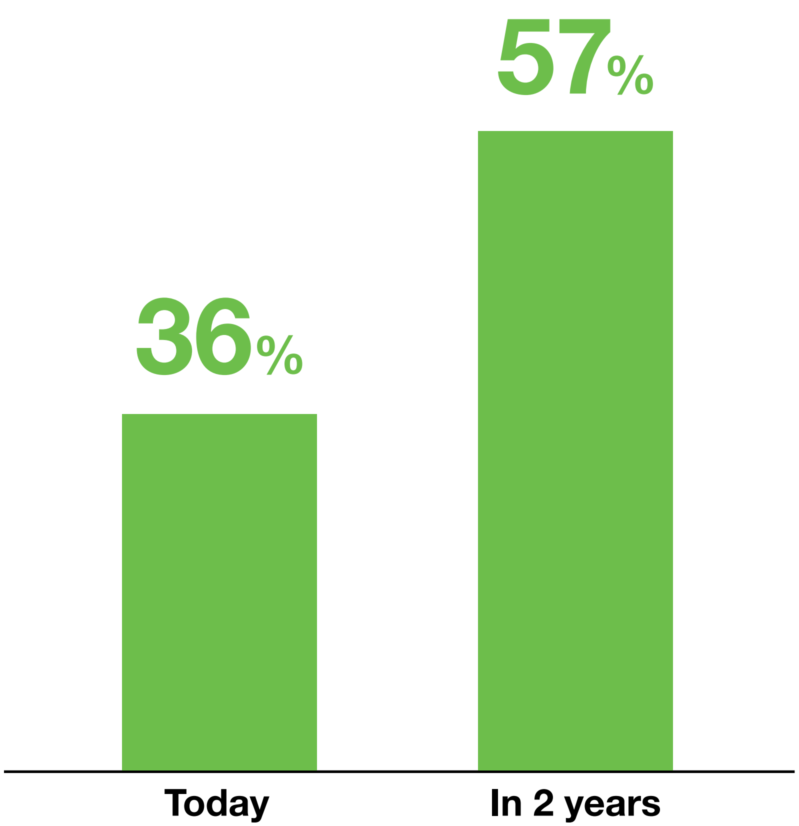
Performance—According to data from Microsoft, in some U.S. states, most counties lack access to connectivity that's faster than 15MB/s. For example, according to the National Broadband Map, in areas just outside New York City, the fastest available wired endpoint connections top out at 30MB/s. If an endpoint device was there, it would take three days and thirteen hours to move a 1TB file.

Cost—Many use cases need intermittent or occasional connectivity, not a constant monthly bill, and commercial last-mile wired broadband costs $75-150 per month, per connection, regardless of use. To make matters worse, if operations are in unserved or underserved areas, wired connectivity improvements are incredibly expensive. A developer and owner of master-planned business parks was faced with a $4 million bill to bring wired connectivity improvements to one business park.[4]

If planning to use last-mile connectivity for endpoint-to-edge transfers, enterprises need to ensure that the services they need are available and sufficiently performant, and costs are acceptable.

In the real world, most organizations use some mix of wireless and wired connectivity to manage their endpoint-to-edge data transfer. Many organizations are doing well enough with a mix of wired and wireless connectivity, but it's clear that the shift toward distributed infrastructures and exponential data growth could (and does) put pressure on wired and wireless networking. In many cases, the available networks aren't entirely sufficient for the task, particularly as data sets grow over time, and unedited, unaltered data sets are needed. Many organizations use workarounds, including aggregating multiple connections, which can help address bandwidth and consistency issues with the downside of increased costs. Some organizations wish they had an alternate technology that could move multi-terabyte data sets from any location, as intermittently or frequently as needed, in any location, and with high levels of predictable reliability and performance. Increasingly organizations search for better ways to store, protect, and move massive, raw data sets from endpoints to the edge at any time, under any circumstances, without any obstacles caused by technological complexity that's impossible to manage in the field.

Organizational data transfer from edge to core



Edge to core

Having explored the endpoint-to-edge challenges, let's explore the challenges of moving large data sets from the edge to core data centers. Whether for analytics, backup and archive, or service delivery, edge to core transfers are important to organizations, and data volumes are growing. According to a survey we recently commissioned, organizations now periodically transfer about 36% of edge data to the core. Within only two years, this number will grow to 57%. The volume of data immediately transferred from edge to core will double, from 8% to 16%, with many organizations transferring multi-terabyte data sets each week. This means enterprises will have to manage a lot more data in motion and will be testing and evaluating core technologies for data transfer.

All the challenges we've covered earlier apply to edge to core, because edge micro data centers may rely on wireless and wired broadband connections, but due to the amount of data that's aggregated at the edge, many organizations rely on technologies like middle-mile fiber infrastructure.

As we already noted, high-performance last-mile connectivity isn't available in many areas, and it's mainly limited to 1GB/s transfer speeds, suitable for many applications, but not for frequent multi-terabyte transfers. Middle-mile infrastructure is something different…and important. We've seen it defined as "the connections from national and major regional internet backbones to local networks," and that's as good a definition as any. It's simply the high-performance network that spans out from backbones toward (without reaching) end-users.

Fiber connectivity is the backbone, literally and figuratively, of the Internet and all major enterprise data transfers. Backbone connectivity scales to 800GB/s connections in major metropolitan areas, allowing gargantuan amounts of data to be simply, reliably, and safely transferred from point to point. Branching off those backbone connections, middle-mile fiber infrastructure is the key enabler for edge-to-core data transfers because it delivers speeds up to 10GB/s.

However, middle-mile fiber isn't perfect. It has some issues:

Availability—The reality of fiber is that availability is patchy outside of major metropolitan areas. Substantial parts of the country don't have adequate middle-mile infrastructure, resulting in sub-standard broadband performance and/or reliability, or simply no access at all. The recently passed Infrastructure Investment and Jobs Act includes billions of dollars in grants to help providers build out both last-mile and middle-mile fiber infrastructure.

Provider geographical variability—today, organizations that are attempting to create widely distributed IT infrastructures discover that middle-mile fiber access isn't necessarily available from their preferred provider in a given location. Each fiber provider has geographical areas where they have significant offerings, and other areas where they lack capacity or capability.

Cost—Another problem with fiber is cost. Securing a link means that an enterprise is paying for a resource, whether or not it's being used. A typical 10GB/s fiber connection costs thousands of dollars, often as an up-front cost, in other forms as an ongoing expense such as a multi-year lease. Resulting operating costs can easily get out of hand.

Features and functionality—Finally, unlike conventional wireless and wired last-mile infrastructure, which is full of competition and largely commoditized (with standard pricing and features) not all fiber providers are created equal. Each has a mix of capabilities, differences in reach and scale, and differences in price models and contract terms. Picking the right one for edge-to-core mass data transfer requirements can be more of an art than a science.

Edge and core to cloud

Finally, we come to edge-and-core-to-cloud data transfers.

We all know that cloud is a ubiquitous technology. More than 90% of organizations use public cloud, and more than 80% of those have multi-cloud environments. On average, organizations use 4.8 clouds, running 32% of workloads in public cloud and 45% in private cloud. We also know that core-to-cloud data transfer is a ubiquitous requirement. Whether they're aggregating data at the edge or the core or both, most enterprises need to transfer large data sets into the cloud.

On average, organizations use

4.8

Clouds

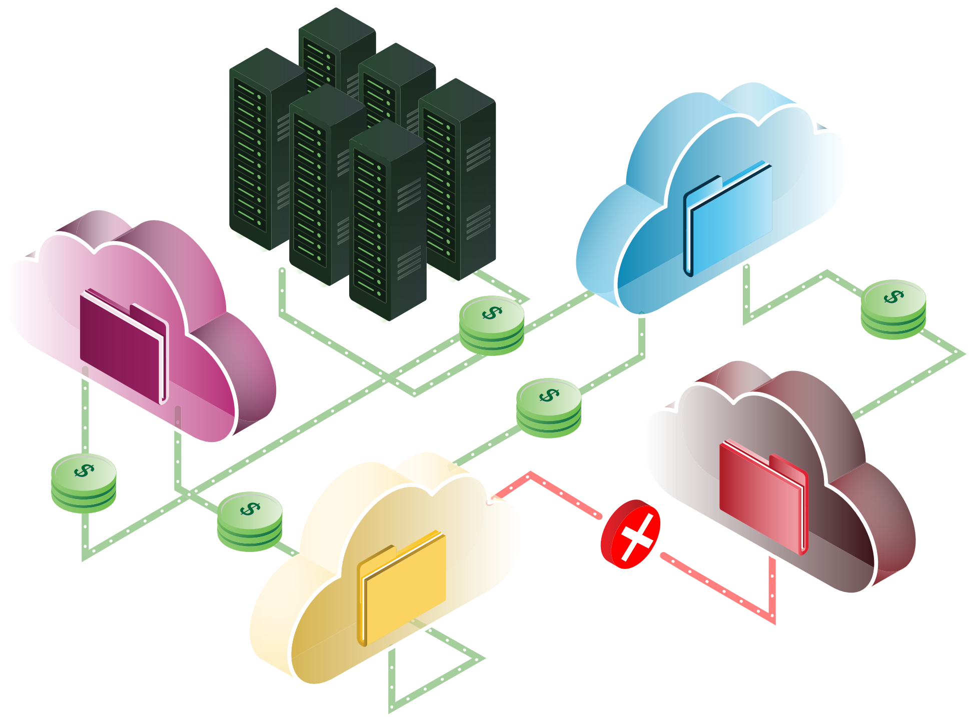
Cloud providers know this. They offer massive bandwidth (100GB/s backbones and hundreds of terabits of capacity) as well as dedicated connections (called cloud onramps) from every major fiber provider and every major public cloud. Public cloud providers also have mass data transfer functionality that includes centralized job management, physical and digital security and reliability functionality, and granular control over costs, transfer times, and scheduling. Cloud providers aim to make it easy for organizations to move multi-terabyte data sets into their clouds. Data movement won't be a problem if the enterprise has the bandwidth needed from their edge sites or core data centers to the preferred cloud or clouds.

However, cloud data transfers have one substantial problem, and that's pricing. Most enterprises discover, over time, that cloud data transfers come with an unpleasant set of surprising, unplanned, difficult-to-understand costs. Whether data is being moved into cloud storage from one storage type to another or from one cloud to another, complicated, confusing cost models apply.

Examining the pricing page for a leading cloud provider, for one data center region on a single cloud, an organization would have to manage:

- More than a dozen different storage tier price points, depending on data set size or number of objects.
- A pricing chart with eighty-five cells for request and data retrieval charges.
- Nearly sixty different data transfer charges.

In other words, mass data transfer with public clouds can force organizations to accept a pricing model that's almost impossible to understand and incredibly difficult to manage. Cloud providers seem to be devoted to complex models for mass data transfer that make it easy to move data INTO their clouds. Still, moving data from one storage type or from one cloud to another comes with costs that can quickly become intolerable. As data sets grow and as needs change, costs become more unpredictable, spiral out of control, and ultimately result in budget overruns and executive frustration.



Exploring Physical Mass Data Transfer

In the previous section, we explored all the issues of typical data transfer technologies and the impacts their limitations have. The fundamental problem is that data sets are exploding, organizations need effective access to their data, and many technologies can't get the job done.

For the sake of argument, let's say that an enterprise has a 20TB dataset at the edge, and the analytics team needs it tomorrow to run a prescriptive analysis to support a business case for a new product line. They have a dedicated 1GB/s link from the edge data center to the core. Can the data be moved in time?

No. Under perfect conditions, assuming no downtime and no lost performance, moving that dataset would take two days.

All the technologies previously covered have their places, are a good fit for some requirements, and not such a good fit for others.
Our customers tell us, repeatedly, that their last-mile and middle-mile connectivity, in some circumstances, with some data sets:

- cannot provide adequate bandwidth for their requirements.
- isn't physically available where they need it.
- comes with too many dropped connections.
- forces them to accept long-term, unsustainable costs.

Rather than moving data via radio waves, light, or electrons down wires, organizations are beginning to turn to physical data transfer technologies to mitigate these issues. These devices or platforms are becoming popular as an alternative form of data movement when existing data transfer technologies can't support an emerging or ongoing need. Mass data transfer platforms are typically comprised of a high-capacity drive or drives, an interconnect (could be PCIe, iSCSI, USB 3.2, or Ethernet), and software to facilitate data transfer.

- First, these mass data transfer devices are shipped to a source location (often at the edge or a core data center).
- Next, an admin transfers data to them.
- Then the devices are physically moved (driven, flown, or shipped) to a destination, usually a core data center or cloud provider.
- And finally, data is removed from the mass storage device, which can be shipped back for reuse.

Due to exponential data growth, emerging demands for data, and an accelerated need to utilize data quickly, more and more organizations are using mass data movement tools and platforms to augment and extend their data transfer capabilities.

These tools come from many different vendors and come with many features and capabilities that may differ from platform to platform. Evaluating different platforms isn't easy, but we have some recommendations. When evaluating physical data transfer tools, technologies, and platforms, critical criteria include:

1. **Capacity**—Does the technology have the capacity needed for the data sets, and does it have overhead room to accommodate data growth?
2. **Time-to-ingest**—Can the dataset be moved to the device in the time allotted? Are data transfer speeds sufficient?
3. **Time-to-access**—Can the device be physically moved to its destination quickly enough so that the data can be used?
4. **Frequency**—Can the devices be used according to the required frequency (once a day, once a week, once a month)?
5. **Security**—Are the devices secure? Could they be tampered with on-site or during movement from one location to another? Are they encrypted? Do they meet government standards for data protection? Do they have role-based access controls or other access controls that reduce the risk of login and data access by unauthorized users?
6. **Ruggedness**—Are the devices built for frequent physical transport, or would they fail if dropped or jostled?
7. **Open infrastructure/vendor lock-in**—Do they use open standard technologies—or are they proprietary? Can they work with any operating system and cloud provider—or are they restrictive?
8. **Costs**—Are they easy to purchase? Are buyers stuck with high fixed costs, or can they easily increase or decrease their costs as needs change? Are they purchasing products, or purchasing a service?

Let's explore a few examples of physical mass data transfer tools to explore these considerations.

External drives

These drives often come with 8-16TB capacities, connect via USB 3.2 Gen II (up to 20GB/s), and have a set of basic features for data transfer. Some offer encryption and basic administration features.

Though these are widely available, many organizations that try them discover that they're not the best option for enterprise mass data for several reasons.

PROBLEMS INCLUDE:

Support

They are mainly consumer-grade products. Enterprise support is lacking.

Performance

Performance can be unpredictable. Some of these products, especially the larger capacity ones, use 5400 or 7200 RPM hard drives, resulting in data transfer rates as low as 160MB/s (1.28GB/s).

Security

Many lack any security measures, leaving data at risk.

Fragility

Many of these aren't rugged enough to take the jostling and impact of endless shipping from edge to core and back again. Data loss in transit is a real possibility.

Cost

These are products that are purchased and kept, meaning that companies end up with an upfront capital investment regardless of product use. Scaling up and down becomes a problem.

Cloud vendor data transfer devices

With the limitations of conventional external drives, some public cloud companies have developed alternative edge storage devices for mass data transfer. One of these offers 8-14TB of usable capacity per device. The devices are built for enterprise use; they're ruggedized, secure, and portable. However, customers who use these report some problems.

PROBLEMS INCLUDE:

Lock-in

The vendors don't support use with other clouds right now. If an infrastructure is multicloud, the platform may not be able to move data to every cloud provider.

Bi-directional movement

Currently, some vendors don't support data movement from the cloud to the device—only device to cloud.

Costs

These devices come with a complicated set of costs, including setup fees per job, per day fees, and data transfer fees priced per gigabyte—plus a $5,000 charge if a device is lost in shipping.

THESE SOLUTIONS ARE:

High capacity
Up to 96TB per device.

High performance
Up to 22GB/s throughput.

Highly secure
Built with enterprise-class security, including secure, self-encrypting drives, AES 256-bit encryption / FIPS approved, off-appliance encryption keys, and user-based permissions.

Rugged
Built with military-grade ruggedization, includes a high-performance case for safe data movement.

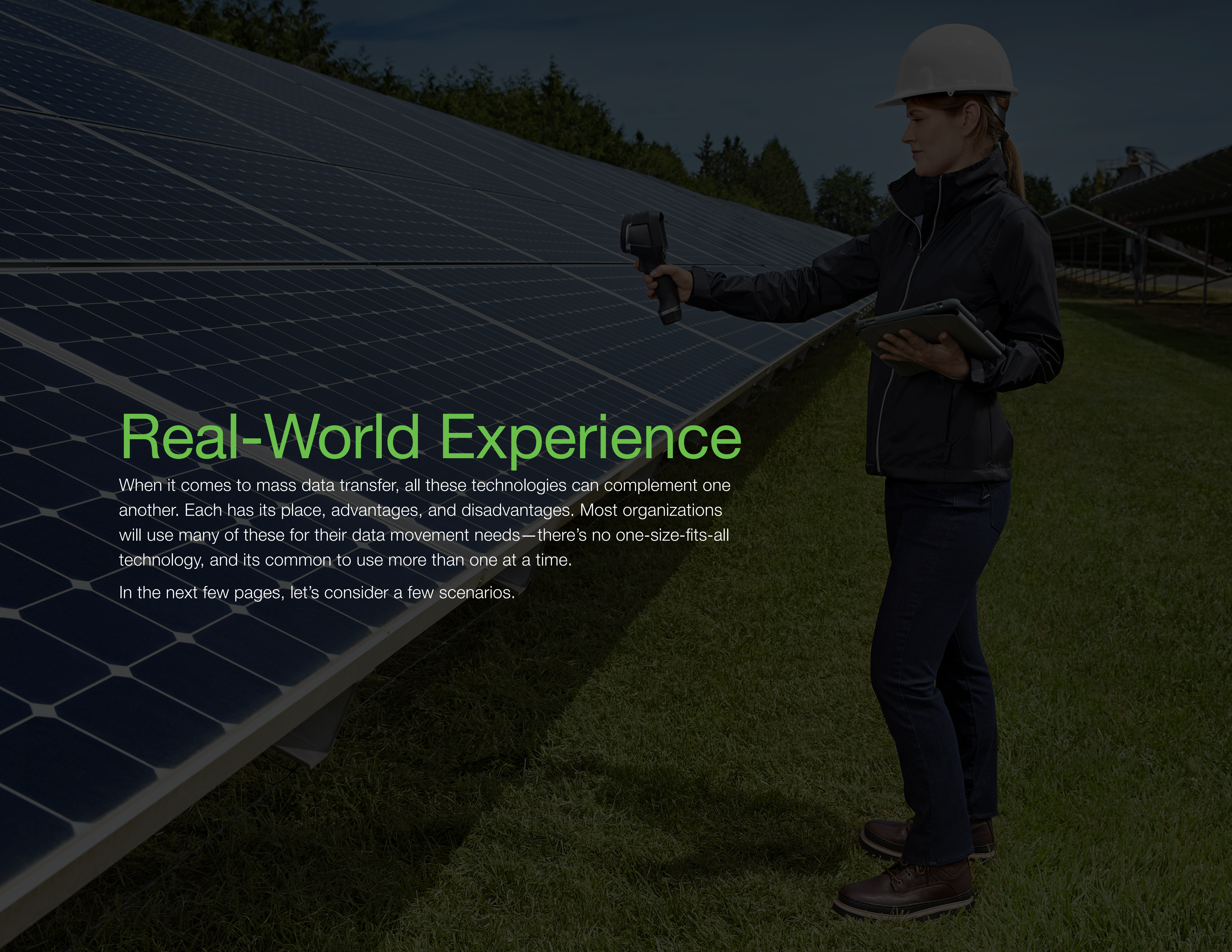
Open infrastructure
Built with open standards and compatible with most ecosystems

Simplified
Easy connectivity, touchscreen drag-and-drop data transfer.

Cost-effective
Procured through an as-a-service model that allows fine-tuning of the service month by month, resulting in reduced CapEx and OpEx over time.

Enterprise data shuttles as a service

To compensate for these weaknesses, a few vendors like Seagate have built suites of enterprise data shuttles with varying capabilities to serve the needs of organizations that have regular mass data transfer requirements.



Real-World Experience

When it comes to mass data transfer, all these technologies can complement one another. Each has its place, advantages, and disadvantages. Most organizations will use many of these for their data movement needs—there's no one-size-fits-all technology, and its common to use more than one at a time.

In the next few pages, let's consider a few scenarios.

Media and entertainment

A film production company has a second unit, filming scenery in the deep desert of Arizona. It's creating a few terabytes of digital video each day. The production house in California urgently needs that video to continue working on the movie. One problem: the second unit is in a dead zone; there's no wireless network, much less access to fiber. Without the video, the production house cannot do its job, delaying the process of finishing the movie.

What are the parameters of this situation?

- A need to move extensive data sets.
- Movement must happen every day—it's a very time-sensitive requirement.
- A limited window of opportunity to download the data for transport because filming happens during daylight hours every day.
- Limited time to ingest the data when it arrives.
- There is budget to move the data, but there are limits.

In this scenario, wired and wireless data movement technologies aren't an option. Without a physical data movement technology, the film can't move forward. A technology that enables the second unit to physically ship dozens of terabytes of data overnight to the production house is what's needed.

But not all physical data transfer technologies are suitable.

External drives

- They may have adequate capacities.
- Ingest rates might be too slow—at 0.29GB/s, transferring 3TB would take just under three hours.
- Durability may not be rated for this scenario—will they hold up to damage caused in transport?
- They come with sunk costs: once the drives are purchased, they're owned, even if they sit on a shelf after the film is over.
- If a drive fails, is enterprise support available 24 × 7? Are the drives even supported for this intensive application, or does the consumer-grade warranty not apply?

Cloud vendor on-premises storage transfer devices

- Good capacities—dozens of terabytes.
- Faster ingest rates—at 0.4GB/s, transferring 3TB would take over two hours.
- They're ruggedized, suitable for shipping.
- They come with enterprise-class support.

However, these devices only work with a single cloud provider. Suppose the production house has in-house effects applications. In that case, they're going to be forced to transfer data into the cloud then download it to their facility, resulting in a more complex workflow and added cloud data transfer costs.

Enterprise data shuttles offer:

- Large capacities—dozens of terabytes.
- Excellent transfer rates: 3TB can be transferred in eighteen minutes.
- Ruggedized, safe for shipping.
- Open standards, can be shipped directly to the production house.
- Enterprise-class 24 × 7 support.
- Provided as a service, pay as you go, stop paying when the requirement is over.

In this case, enterprise data shuttles seem to be the only effective technology.

Healthcare

Consider another example. A technology company has an augmented reality application for surgical visualization they want to deliver to rural hospitals, and they must deploy edge infrastructure in various cross-county regions to accommodate the need.

The edge infrastructure nodes must be connected to cloud infrastructure, but connection speeds vary from location to location—some locations are close to cloud on-ramps or fiber backbones, others are limited to 25MB/s connectivity.

The problem is that every day, if the platform is being fully utilized from each location, the application needs to transfer between 2-4TB of data back to the cloud infrastructure for backup and archive. Data sets need to move to your cloud as quickly as possible, hopefully within a day.

What are the considerations in this scenario?

A mix of technologies could address the requirement.

- Some sites have access to 10GB/s fiber, which is perfect for this requirement.
- Other locations have 1GB/s fiber, which would take a little over two hours to transfer 1TB, assuming there was no other traffic on the link. That might be adequate, or it might not.
- A few locations have slow links that couldn't be utilized for this requirement.

All technologies must be easily deployed and utilized at an edge location

- There are obvious advantages for the always-on connectivity of a wired link.
- If utilizing a physical transfer technology, staff will be needed on site to manage download, packaging, and shipping devices.

They utilize a single cloud for backup and archive

- Any data movement technology needs to work well with the cloud.
- If an organization uses external hard drives or data shuttles, a team will be responsible for uploading data to the cloud. If utilizing a cloud vendor device, their team will do the work. Any data movement technology that reduces data transfer costs and complexity is a good thing.

If data set requirements grow, the technology has to keep up. It's possible that over time, data sets will expand, putting added pressure on 1GB/s sites.

One approach could be to send critical data over a fast link, and less critical data via data shuttle.

- Data reduction technologies at edge locations could play a role.
- Some data sets might simply not be backed up, or only backed up on location.

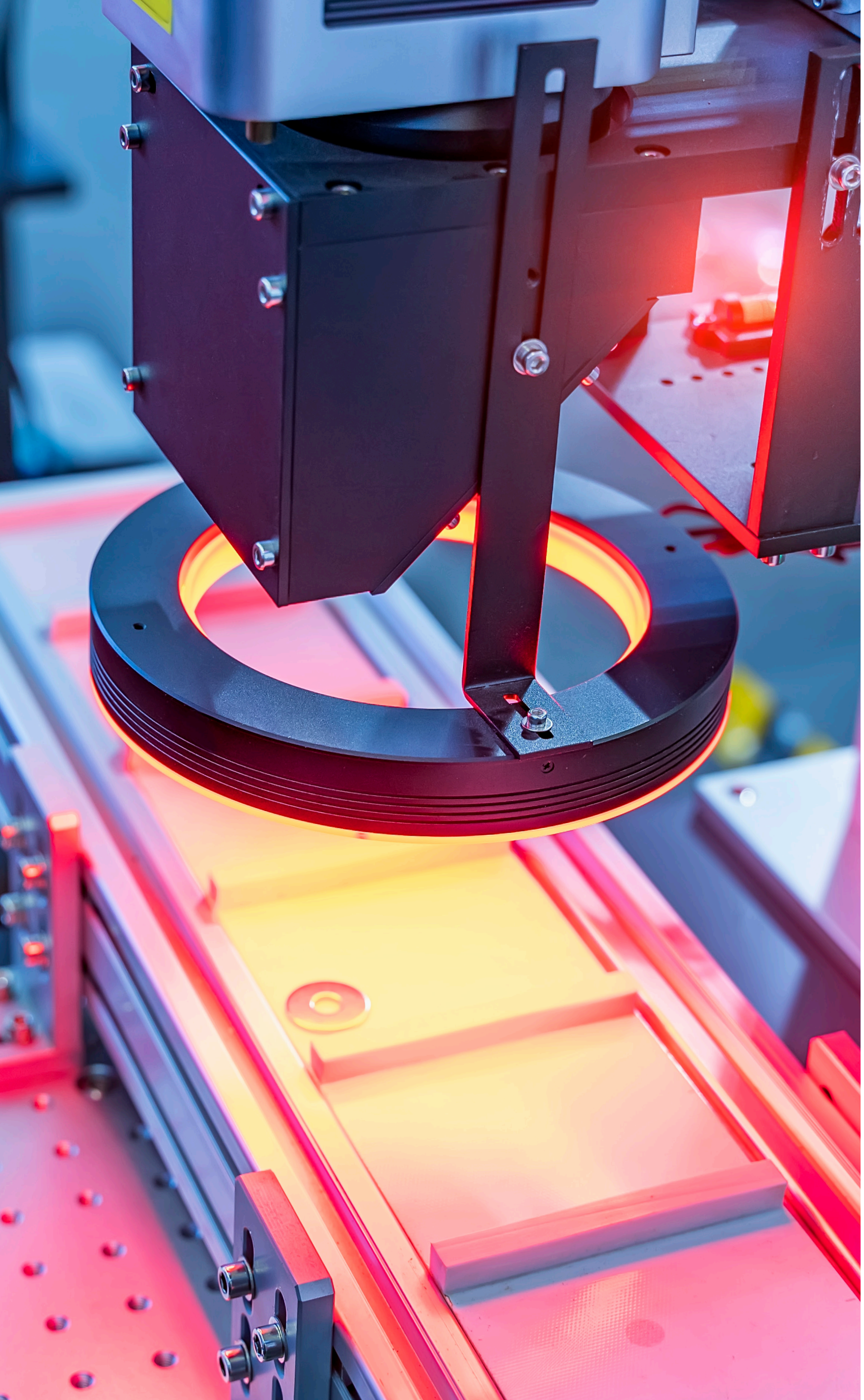
This scenario calls for a mix of fiber links and physical data transfer devices, probably cloud vendor devices. Since there's a single cloud for backup, in this situation where fiber links aren't adequate, it makes sense to use the cloud vendor devices since the cloud vendor will be responsible for adding new data to cloud storage for backup and archive.

Manufacturing

Consider a third scenario. A manufacturing company with several locations spent millions of dollars on a digital transformation initiative, adding IoT sensors at all parts of the manufacturing process.

Each site produces 15TB of data a week, and that volume is growing. They need to transfer that data to a specialist data analytics firm for processing and predictive analysis, which they use to optimize their operations.

They've tried using 5G, LEOs, and 1GB/s fiber for the requirements but discovered that reliability issues (bandwidth restrictions, connectivity problems, utilization by other applications and users) made conventional connectivity a sub-optimal solution. Some days, transfers are successful, and other days, they aren't.



Smart factories increasingly struggle to analyze all the data they can capture. Again and again, they find out that their data creation capacity outstrips their data transfer capabilities.

They need:

- A platform that can transfer dozens of terabytes with predictable ingest rates and transfer rates.
- A platform built on open standards. They don't need to be locked into a single ecosystem or model because as requirements change or they change data analysis providers, they can't afford to transfer data from a closed platform to another closed platform.
- A platform that offers reliability. It needs to function as expected, can be reused dozens or hundreds of times, and when problems arise, comes with 24 × 7 enterprise-class support.

In this scenario, enterprise data shuttles are the logical choice. They offer the consistency, speed, and reliability needed. They also support open standards, making them a flexible platform that can be used by any data analytics provider, not just the one currently selected.

Enterprise data center consolidation

In a fourth scenario, a large enterprise has decided to consolidate their data centers, moving from fifteen data centers, scattered around the world, to four—in Italy, one in South Africa, one in China, and one in the U.S. Each data center has hundreds of terabytes of data, ranging from production data, data used by legacy applications, file and object storage, and even online backups.

As the IT teams consider their options, some facts emerge:

These data sets are in different formats, running on different operating systems, and file sizes range from a few megabytes to terabytes. Whatever they use for the migration, it has to support different formats, operating systems, and file sizes.

- Some of these data centers are connected to middle-mile,100GB/s fiber infrastructure, and others are connected using 10GB/s links.
- That fiber infrastructure is already constantly utilized for all sorts of traffic—applications, file movement, replication, backup, and archive.
- Administrators don't want to saturate those links during business hours with massive data migrations.

Administrators could, in some cases, purchase additional capacity from their fiber network providers for this data migration initiative. Still, if that capacity isn't already lit, it could take days or weeks for additional ports to be turned on, and costs could be very high.

In this case, the enterprise will likely use a combination of existing links (after hours), new links (where they can be quickly brought online at an appropriate cost), and enterprise data shuttles to migrate the data to their new data centers. Enterprise data shuttles are likely to offer a compelling cost/benefit ratio compared with new high-performance fiber links that will lead the IT teams, in the final analysis, to rely heavily on data shuttles for the migration.

Autonomous vehicles

Let's wrap up with a final example. A pioneer in autonomous, self-driving freight trucking needs once-a-week access to somewhere between 60-75TB of data collected by each truck to optimize fuel consumption and refine their algorithms for steering and obstacle avoidance. However, the trucks are going places where 5G networks are spotty and not entirely reliable.

How could they cope?

- They would need a device that fits on a truck, is rugged, and reliable.
- It would need to constantly ingest and store data.
- It would have to be easily removed and used.
- It would have to be able to transport data anywhere—possibly to a regional hub, possibly to a centralized location.

This is a situation where enterprise data shuttles are an ideal solution. A ruggedized enterprise data shuttle with high-performance ingest can be shipped once a week relatively easily.

In these scenarios, you can see some of the challenges forward-looking, creative organizations face when it comes to data movement. Today's reality is that we're constantly putting pressure on our approaches to mass data movement, and sometimes, the technologies can't serve the requirements. Fortunately, many options exist, and new ways to solve lingering problems with mass data transfers are evolving.

Conclusion

We hope this playbook offers insight into the emerging state of play for how to best solve specific mass data transfer technology needs. A wide variety of options exist, and there are new ways to work around technology capabilities and limitations. In summary, the available technologies include:

Option	Transfer Speed	Pros	Cons
Wireless	Up to 1GB/s	Widely available, always on, multiple vendors, relatively low cost per connection	Huge gap between theoretical and real performance, mass data bottlenecks, reliability affected by conditions, fixed costs regardless of utilization
Last-mile wired	From 25MB/s to 10GB/s	High reliability, always on, mature technology, open standards	Variable performance from location to location, fixed costs regardless of utilization
Middle-mile wired	Up to 800GB/s[5]	Massive performance and reliability	Not widely available, extremely high costs
Cloud transfer	Up to 100GB/s[6]	High performance, easily provisioned through cloud console	Vendor-specific, not open, highly complex cost model
External drives	Up to 2.32GB/s[7]	Easily available, low cost, open standards	Often not ruggedized, limited or no security features, limited by physical shipping speeds, requires on-premise administration
Cloud vendor devices	Up to 3.2GB/s[8]	Built for enterprise use, rugged, mature	Vendor lock-in, complex costs, limited by physical shipping speeds, limits on bidirectional transfer, requires on-premise administration
Enterprise data shuttles	Up to 22.4GB/s[9]	Flexible, rugged, can be faster than network, easy to align supply/demand, built for enterprise, open standards	Limited by physical shipping speeds, requires on-premise administration

As long-standing market leaders, we at Seagate are excited to help organizations shift from workarounds and restrictions to approaches that deliver better business value. All these technologies have a place in mass data movement, offering a mix of capabilities, advantages, and disadvantages that give enterprises more flexibility, versatility, and possibility.

References

[1] https://www.forbes.com/sites/gilpress/2021/12/30/54-predictions-about-the-state-of-data-in-2021/?sh=46845807397d

[2] Ibid.

[3] https://www.idc.com/getdoc.jsp?containerId=prUS47560321

[4] https://blog.mho.com/how-businesses-are-addressing-last-mile-problems

[5] https://potsandpansbyccg.com/2021/10/08/zayo-installs-800-gbps-fiber/

[6] https://aws.amazon.com/premiumsupport/knowledge-center/s3-maximum-transfer-speed-ec2/

[7] https://www.lacie.com/products/d2/#specs

[8] https://aws.amazon.com/about-aws/whats-new/2021/08/aws-snowball-edge-storage-optimized-nfs-data-transfer/

[9] https://www.seagate.com/files/www-content/datasheets/pdfs/lyve-drive-mobile-array-DS2044-3-2201US-WW-en_US.pdf

Ready to Learn More?

Visit us at
www.seagate.com/products/data-transport/