Vcinity
ULTIMATE ACCESS

TECHNICAL WHITE PAPER

Rev F

Take Distance Out of Your Data

Remote Access to Geo-diverse Data with Vcinity's Radical X and Ultimate X

For Official Use Only (FOUO)

2055 Gateway Place | Suite 650 | San Jose, CA 95110 | T +1.408.841.4700 | info@vcinity.io | Vcinity.io

Contents

Executive Summary

Accessing data within an enterprise spread across geo-diverse locations challenges work productivity, time, and IT resources. Existing methods require data to be transferred and replicated leading to delayed business insights or even insights based on stale data. In addition, having to send a copy of data to every user that requires it leads to copy sprawl, data management challenges and compromised data security. Even with data transfer, data arrives at the destination in an unpredictable time and performance varies with data type/size and application. Utilizing network data mover appliances and applications is so cumbersome that even physical transportation of media or data is considered an acceptable solution. In summary, there is a need to access data without replication across geographic distances to support location-independent data centers, distributed workflow operations, real-time remote content creation, business continuity objectives, and content distribution in predictable time frames.

Vcinity™ enables enterprises to instantly access and operate on data sets over any distance, without copying them and with local-like performance. This is accomplished by transforming the enterprise-wide area network (WAN) into a Global local area network (LAN), enabling local application performance on data over global distances. This capability lets enterprises leverage modern business tools such as machine learning modeling and artificial intelligence innovations leading to more efficient business processes and a greater competitive advantage.

Vcinity's unique approach addresses various layers of the overall stack, through efficient transport, integration, and application transparency. This methodology combines proven, patented network processing technologies at both hardware and software levels, with open standards-based approaches to interoperability and implementation. The fundamental premise is to provide a more efficient access to data leveraging the existing enterprise infrastructure in a simple, easy-to-implement method. An open API-based interface coupled with powerful data management and synchronization tools offers a solution for enterprise customers to close the IT gap between applications and data and reduce operational costs commonly attributed to closed proprietary systems. Vcinity addresses these challenges with the approach outlined in Figure 1.

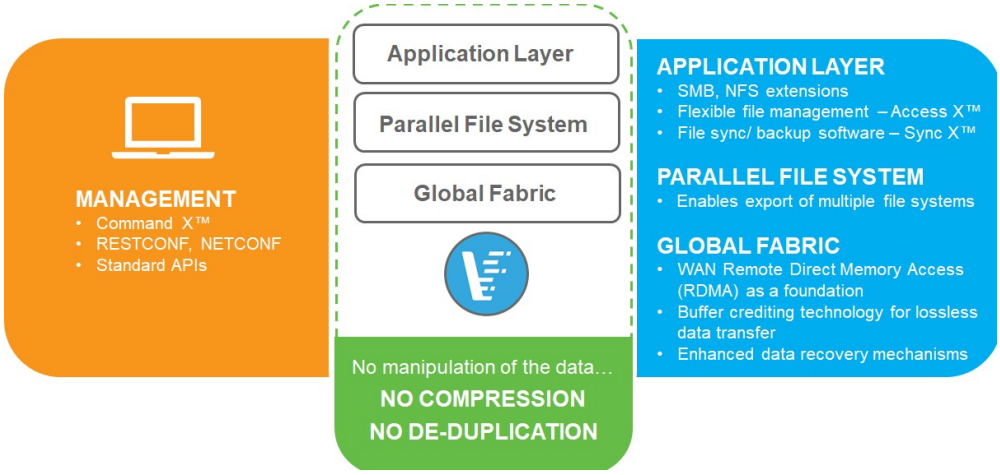

Figure 1. Vcinity's Methodology: Leveraging HPC Technologies to Remove Distance Barriers

The Vcinity Radical X™ (RAD X) solution enables the Global Fabric by extending InfiniBand (IB) or remote direct memory access (RDMA) over converged Ethernet (RoCE) fabrics outside the four walls of a data center. RAD X supports fabric extension seamlessly and transparently over networks ranging from 20Mbps to 100Gbps and is most commonly deployed in high performance computing (HPC) and Supercomputing environments.

As described in Figure 1 on page 3, the Vcinity Ultimate X® (ULT X) builds on RAD X and integrates with other HPC technologies and high-speed storage. It attaches to an enterprise standard similar to an industry standard network-attached storage (NAS) or transitional high-speed storage tier and, when connected over the metropolitan area network (MAN)/WAN, provides a high-performance, geo-diverse data exchange. ULT X essentially enables a global federated data platform for accessing data with or without replication, using global namespace and network-mapped drive volumes across geographically distributed enterprise storage.

Vcinity's Approach—RDMA over WAN

The concept of turning WAN into Global LAN and enabling a global fabric is achieved through RDMA over WAN.

It is commonly accepted that remote application execution is not possible when a high bandwidth delay product (BDP), i.e., the product of a link's capacity (in [x]bits per second) and its round-trip delay time (in [x]seconds), exists between the compute location and the data location. The BDP between the compute and data when they are within a data center, on a LAN, or in the same cloud provider, is acceptable for most common applications. However, as soon as the BDP increases traversing a WAN, these same applications are functionally unusable due to the time required to get the data to the location of the compute. BDP has major impact on traditional networks using the transmission control protocol (TCP)/internet protocol (IP) and various methods have been used to optimize or tune TCP/IP to make it more efficient with moderate results.

In most network infrastructures, TCP/IP is the dominant transport/network stack however TCP/IP is a poor protocol for modern networked systems with geographically dispersed data and compute resources. It is fundamentally limited data goodput over higher speed networks with significant latency, or long fat networks (LFNs—see RFC1323) as shown in Figure 2 where TCP/IP suffers performance degradation as latency and packet loss increases. Additionally, its inherent weaknesses is clearly seen when considering the Mathis Formula:

Rate <= (MSS/RTT)*(1 / sqrt{p}), and more accurately from Padhye et. al.:

Rate = MSS * [((1-p)/p) + w(p) + Q{p,w{p}}/(1-p)] /(RTT * [(w{p}+1)]+(Q{p,w{p}}*G{p}*T0)/(1-p))[1]

The traditional manner of networking using TCP/IP is no longer viable for LFNs when considering jitter, packet loss, collisions, and congestion.

[1] http://conferences.sigcomm.org/sigcomm/1998/tp/abs_25.html

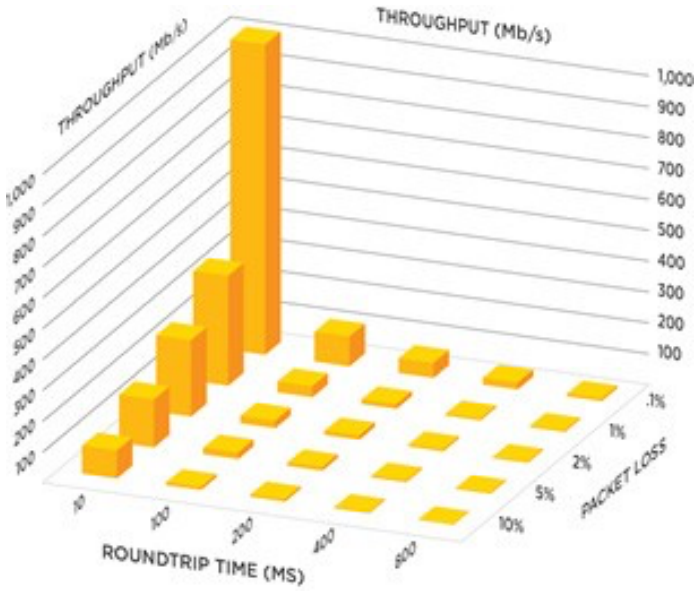

Figure 2. TCP Performance Over Latency

From an infrastructure perspective, the TCP/IP stack also consumes a great amount of resources, including CPU, memory, system bus, and even at the OS kernel level, which can be allocated for user and application workloads. The legacy TCP/IP stack incurs extra processing latency, which slows down application performance and ultimately its throughput. Modern data centers and cloud service providers (CSP) are adopting and migrating to a cost-effective transport/network paradigm to improve performance: RDMA; Figure 3 shows a comparison between the two.

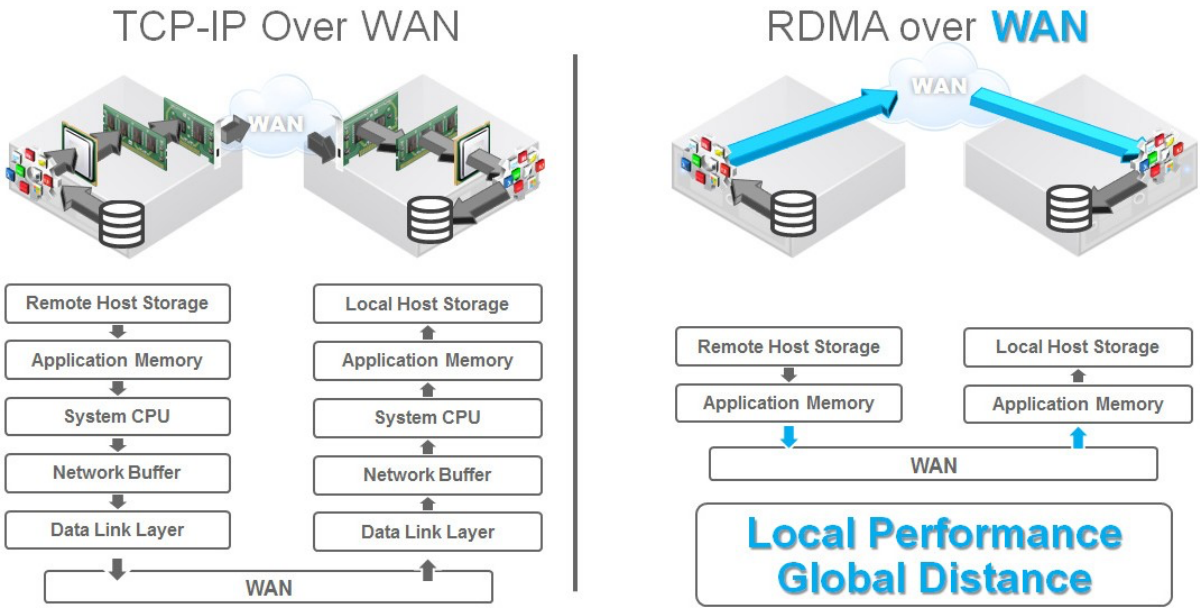

Figure 3. Technology Comparison

Due to its compelling efficiency and performance, RDMA-based solutions have earned their positions in cloud, HPC, and supercomputing infrastructures to address demanding workloads. However, most of the RDMA-based technologies—coexisting with TCP/IP, displacing TCP/IP or overlaying on top of TCP/IP—focus on improvement within data centers or campus area networks. As the footprint continues to expand, RDMA will become a favorable end-to-end transport/network solution where CPU offload, memory class storage and higher performance memory are needed. Vcinity's UTL X solution is designed to extend the RDMA-based infrastructure as well as the associated goodput over long distance to facilitate RDMA as a global transport/network paradigm.

Radical X Solution

The core functionality of the Radical X (RAD X) was originally created in the form of Application-specific Integrated Circuits (ASICs). Over time, those functions were converted to IP software libraries that are deployed into FPGAs to achieve hardware accelerated performance and deployed as software for use with standard virtual environments and cloud environments. The product portfolio consists of different options supporting a different number of interfaces and interface rates of 1/10/25/100GE RoCE and QDR/EDR InfiniBand. They primarily come in the PCIe form factor that can be plugged into the server and are powered by the powerful Intel® FPGA family. The RAD X solution is designed specifically to extend and manage RDMA-based protocols (e.g., RoCE and InfiniBand) over LFNs where other protocols would be ineffective. The following sections detail the functionality and methods employed to achieve this capability.

Network Fabric

The RAD X products support layer 3 networking utilizing L2TPv3. The data is encapsulated within Vcinity proprietary traffic engineered flows to manage the network transport for high latency and high bandwidth connections and accommodate large amounts of data in-flight.

The fabric is established through a single physical interface on the RAD X, over a unique network tunnel and single or multiple flows (e.g., one of the RAD X product family members—RAD X-1010er— supports up to 128 tunnels and 96,000 flows). As a function of the underlying protocol, these tunnels are lossless and deterministic in nature.

Routing

The RAD X supports routing via standard IP by adding an IP header to the encapsulated traffic to support layer 2/3 tunneling. In the event an implementation requires InfiniBand routing, RAD X supports IB Routing in accordance with the Open Fabrics Enterprise Distribution (OFED™)/Mellanox OFED (MOFED) standards.

Data Inflight Buffering/Crediting

RAD X contains significant buffer memory to ensure that all data inflight over an LFN is able to be persisted in the event of packet loss or other errors on the network. These buffers have the capacity to support up to 2.5s of latency for a link speed of 10Gbps, or 250ms at 100Gbps. These buffers are managed on a 'credit' basis in a manner similar to Fiber Channel, such that any system requesting data from another location reports that it is ready to receive and gets credits based on current traffic flow and available buffer memory. This enables a fundamentally lossless approach by only sending data when the recipient is ready to receive it.

Flow Control

With the protocol in use there are significant limits prescribed by the Max Full Rate Distance in flow control messaging. While these are sufficient for LAN connections it is not sufficient for a WAN/LFN. The current defined limits for InfiniBand are:

Full Rate Distance Limit: 52km (DDR), 13km (QDR), 10.5km (FDR-10), 7.7km (FDR), 4.3km (EDR)

While these distances are greater in a converged Ethernet implementation, they are still insufficient for an LFN considering the following limits:

Full Rate Distance Limit: 3355km (1GE), 336km (10GE), 134km (25GE), 84km (40GE), 67km (50GE), 34km (100GE)

RAD X manages the lossless tunnel with flow control between any two endpoints and includes forward congestion signaling downstream and backward congestion signaling upstream. This also includes multiple tunnel trunking for load balancing and resiliency. The tunnel scheduler schedules priority queues based on different policies; for example, weighted-round-robin, weighted-fair- queueing, round-robin, strict-round-robin, etc.

The system also manages the flight time differential within its flow control. The long-distance delay between the transmitter and receiver is viewed as outstanding committed data. At any instant, if the volume of the outstanding data in flight is greater than the available receiving space, the possibility of loss at the receiving side goes up.

The transmitter implements a counter which keeps track of the outstanding data, F(t), between time= t – ½ RTT and time= t. (t is the current time, and RTT is the round trip time of the connection). At any moment, the counter counts up with the data pushed into a tunnel, I(t), and deducts data which expires at ½ RTT the period, which is equal to the single direction delay. F(t) is pushed to memory for later comparison. The transmitter implements a counter which counts the incremental data, I(t), pushed into a tunnel in the current sample period. I(t) is pushed to memory for later calculation.
The receiver reports buffer space periodically to the transmitter B(t). The report takes ½ RTT time traveling from the receiver to the transmitter, therefore, the receiver receives B(t – ½ RTT) report at time=t. At time= t, F(t – 1/2RTT) is retrieved and compared against B(t – ½ RTT).
At time = t,

Receiver sends B(t);

Transmitter calculates $F(t) = F(t - 1) + I(t) - I(t - ½\ RTT)$;

Transmitter stops sending, if $B(t - ½\ RTT) + R1 - F(t - ½\ RTT) - F(t) < M1$; Transmitter resumes sending,

if $B(t - ½\ RTT) + R1 - F(t - ½\ RTT) - F(t) > M2$;

Before receiving B(t) report from the receiver, B(t – ½ RTT) is set to a default value, which is equal to the full receiving space. M1 and M2 are configurable values. M1 is the stop margin, and M2 is the resume margin.

R1 in the equation represents the committed or minimum drained volume of the receiving buffer. When set to 0, the calculation is conservative to assure lossless behavior under the worst condition,
i.e. the receiving buffer is completely blocked from draining instantaneously or permanently. The calculation works for smaller allocated buffer space which is less than capable of handling full tunnel bandwidth. The transmitter side is capable of self-regulating output adaptively.

Tunnel Congestion

Tunnel congestion is actively monitored to detect errors before any significant degradation can occur due to delays or loss. These are done in the form of a tunnel heartbeat, tunnel RTT monitoring, buffer occupancy threshold, and loss notifications.

Priority Classification

Within the RAD X queue management all incoming traffic is classified and binned. For example:

- Layer 2
  - InfiniBand: {SLID, DLID, virtual lane, service level} in local routing header
  - 802.1Q VLAN: {C-DA, C-SA, C-VID, priority code point (PCP) or differentiated code point (DSCP)}
- Layer 3
  - IP: {SIP, DIP, ToS field IPP, PHB, DSCP}
- Layer 4
  - TCP, UDP and port
  - RoCEv1 and RoCEv2
- Port ID
- Tag, label, tunnel ID carries the essence of priority and other information.

Packet/Flow Segmentation and Reassembly

Variable packet sizes cause unpredictable transmission quality in a lossless tunnel. Once scheduled, a large but lower priority packet can block resources from reallocating for latency or loss sensitive traffic.

The RAD X chops a packet into fixed-size segments to preempt blocking. The unified size increases accuracy for CoS and QoS which are a foundation for congestion immunity.

The system goes further to enable splitting of a single flow into multiple flows that is routed over different physical links of different latencies. In its current implementation, the system supports an 8x8 set of routes, specifically, a single tunnel and flow is split into eight separate paths—each of which is then split into eight separate paths. This is a function of inverse multiplexing wherein the split information is then reassembled by the receiver and variations in latency are adjusted for by storing differential content within the receiver's buffers.

The added benefit of this capability is that separate physical links are aggregated at the receiver endpoint to achieve the full capacity of all utilized connections rather than a single link.

Security

The RAD X supports LWC SIMON and all RAD X support industry standard industry standard AES-256 GCM encryption and authentication at line rate. The encryption is done within the FPGA to minimize the impact on data throughput and as a result, degrades performance by less than 1 percent. The standard default configuration is for static keys to be set by the administrator however, a KMIP interface is available via RESTful API to support any compliant key management and orchestration system.

One level of security is through enabling flow splitting over separate physical routes which are recombined by the receiving unit. When the packet size is larger than the segment size, each packet will be segmented to further obfuscate the underlying data. Figure 4 shows an example of two flows being spread across four physical links with segmentation of the larger packets. Depending on the number of routes chosen, a man-in-the-middle attack would only receive a percentage of the overall data being sent between locations. To create an even stronger security paradigm, each path can be encrypted with a separate key and those keys can be changed frequently by the key orchestration system. Vcinity has branded this functionality as DataPrizm™.

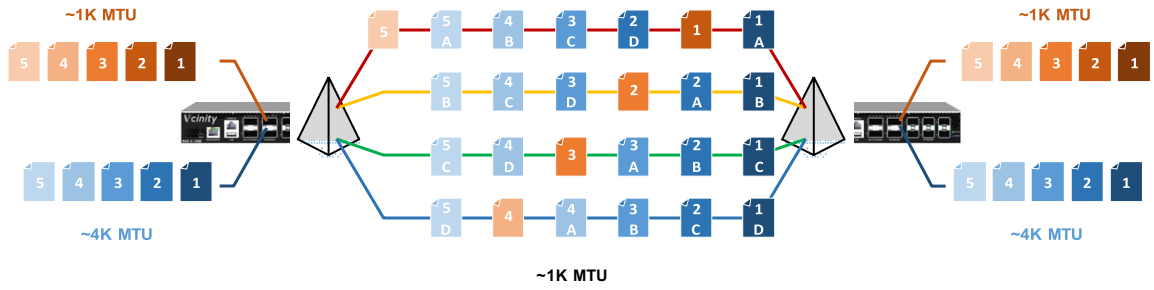

Figure 4. DataPrizm Concept

Ultimate X Solution

The ULT X solution combines one of the RAD X products with an industry standard X86 server or virtual machine running a global parallel file system and the management system software and APIs. The purpose of the server is to provide three main functions:

1. Interface between the LAN and RAD X:
    - Terminate LAN and Storage protocols such as NFS/SMB/TCP/IP
    - Connect to RAD X via InfiniBand or RoCE
2. Act as a storage overlay for legacy storage systems or standard POSIX:
    - Inbound data is received as NFS shares or SMB.
    - Outbound data requests are served in parallel per request.
        - Parallel data is pulled from striped drives to increase throughput greater than a single drive I/O can perform.
        - Data format output is RDMA based.
    - Each node acts as a high-speed data cache
    - Or, act as direct attached storage (DAS) if desired to scale beyond standard caching size
3. Permissions and Security:
    - Manages data access to the file system using standard access control lists (ACLs) across all nodes.
    - Capable of encryption of data at rest
    - Supports standard rights under lightweight directory access protocol (LDAP)/active directory (AD)

The server attaches to the LAN and appears like a standard NAS device. The LAN-side connection is through multiple standard Ethernet NICs, which access multiple storage subsystems within the network. The WAN connection through the RAD X is via a standard RDMA interface for the integrated NIC or virtual NIC. The WAN connection through the RAD X is via an InfiniBand or RoCE host channel adapter (HCA) for the external appliance. This allows for a highspeed data flow into the RAD X at 40G, 56G, or 100G from any RDMA-enabled device.

The product portfolio consists of:

- ULT X VM (ULT X 1000v)
    - Edge—ULT X KVM, ULT X ESXi®, and ULT X Hyper-V™ virtual appliances (~2Gbps WAN)
    - Cloud—ULT X AWS® for the AWS cloud and ULT X Azure® for the Azure cloud (~2Gbps WAN)
- ULT X SW (ULT X 1000s)—Virtual appliance for bare-metal (~5Gbps WAN)
- ULX X FPGA (ULT X 1000e)
    - Edge—ULT X FPGA appliance with FPGA NIC (multiple 1/10/25/100GE WAN based on the RAD X variant)
    - Cloud—ULT X AWS-F1 for FPGA-based AWS cloud instances (~10Gbps WAN)
- ULT X HW appliance (ULT X-1000) with RDMA NIC (up to 100GE WAN)

This creates a distributed cluster of nodes where a new node is easily added to the global cluster and integrated into the data shares. Each share is then available to authorized users via the remote mount point and appears as a standard network mapped drive, regardless of the latency between the two endpoints.

No data is cached at a given node until such time as a data request is made by a user or an application. At that time, any related metadata is pulled from the remote location and if file data is requested, it begins to transfer and prefetch based on selected criteria in the configuration. Frequently requested content may reside in the local cache however, the cache operates on a first in first out (FIFO) basis and is dependent upon the cache size and amount of data requests delivered.

By utilizing the RDMA protocol, data requests are delivered to memory using the inherent OS kernel bypass, and processor offload of the protocol. This same process occurs on the receiving node to ensure immediate application access with the same benefits—creating a more holistic approach to data access and acceleration than can be done via other means.

Comparison with Off-the-Shelf Solutions

The discussion of Vcinity's capabilities is incomplete without the context of certain solutions that we have been compared to in the marketplace and how Vcinity's technology differentiates in a fundamental way. Vcinity's unique feature is its ability to access remote data in-place without needing to pre-fetch, prioritize, manipulate, or move it ahead of time for local edge caching. There is no direct competition in this respect. The high-speed file transfer aspect of our solution is however, viewed as an alternative to edge caching or competing with WAN Optimizers and extreme file transfer (EFT) vendors.

Edge Caching

Edge caching solutions achieve performance by pre-staging or caching portion of data by predicting them ahead of running the workflow.

They claim to reduce total storage by up to 90 percent as only a portion of the data is cached. ULT X provides better savings as data does not need to be pre-staged or cached at all. It remotely accesses data, keeping it in-place inherently removing the need for data caching. Data is cached or replicated only if the workflow absolutely demands it. This results in a maximum of two copies of data including backup data required for disaster recovery or continuity of operations.

Edge caching solutions' claim of high performance with rapid file access for easy collaboration is true only for the pre-cached data. Performance is compromised if data is not in the cache. ULT X provides predictable performance for accessing any and all of the data regardless of where it may be located within the enterprise infrastructure.

By definition, edge-caching addresses data in the cloud and compute on-prem with their on-prem caching appliances. ULT X leverages the economics and scale of the cloud more effectively by both leveraging scalable compute in the cloud while keeping data on-prem and leveraging on-prem compute by reaching into the data in the cloud avoiding the need to move the data close to the compute. Remotely accessing (or reaching in to) on-prem data provides CSPs new ways to address customers not willing to or unable to commit data to the cloud for regulatory, compliance, intellectual property, or other business reasons.

Edge caching's approach primarily addresses latency issues for collaboration with global data centralization and sharing of data in the cloud. ULT X inherently deals with data anywhere including the cloud, data centers, on-prem or remote offices. It adapts to the customer's IT strategy and supports hybrid including multi-cloud environments.

WAN Optimization

WAN optimizers don't allow for remote execution on data but rather focus on optimizing large numbers of lossy TCP/IP flows and reducing the amount of data traversing the WAN. They perform compression, de-duplication or pre-processing of data to achieve the target application performance. This compromises data integrity and doesn't work for certain data types like video or encrypted data. Also, it adds additional pre-processing delay. ULT X provides a seamless data fabric without needing any pre-processing, preserves the data integrity and handles any type of file -based data.

Extreme File Transfer

EFT solutions generally use UDP-based protocols for fast transfers while managing data loss recovery by requiring special software be placed on both ends of the data transfer and has limited scalability beyond 2-3Gbps. ULT X takes a fresh, holistic approach by using HPC protocols to remove distance out of data to create a global data fabric but still seamlessly integrate into existing IT infrastructure with industry standard interfaces. ULT X performance also scales predictably regardless of distance and bandwidth. EFT requires installing and maintaining the software at every server/client whereas once an ULT X-based fabric is established, there is little maintenance required for any additional client.

2055 Gateway Place | Suite 650 | San Jose, CA 95110 | T +1.408.841.4700 | info@vcinity.io | Vcinity.io