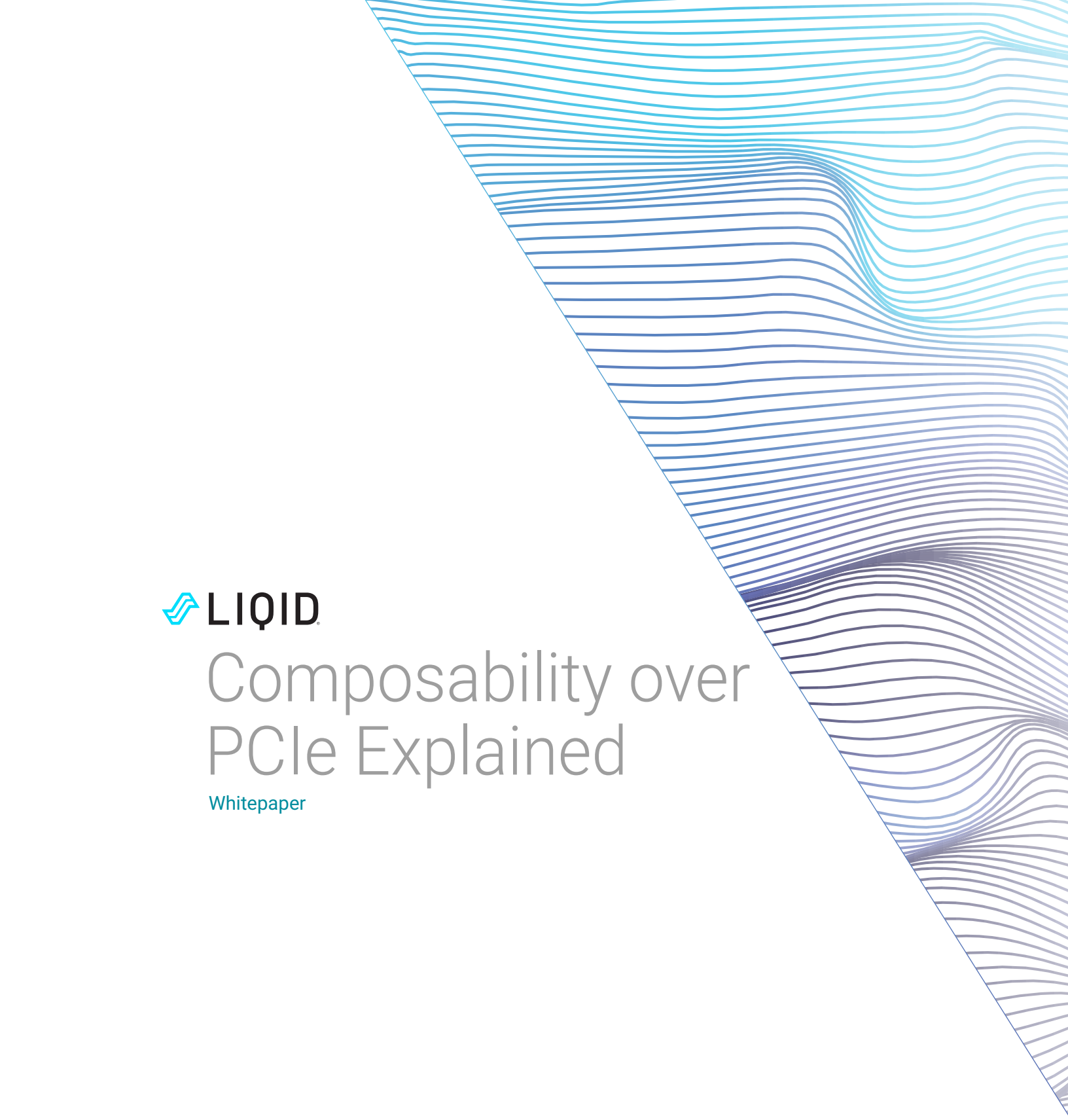


LIQID

Composability over PCIe Explained

Whitepaper

INTRODUCTION

Today's workloads look very different from the typical workloads seen even 5 years ago. Previously it was rare to find customers that were running large scale HPC, AI/ML, Inferencing etc. According to Gartner, "By 2025, AI will be the top category driving infrastructure decisions, due to the maturation of the AI market — resulting in a tenfold growth in compute requirements."[1] These modern workloads have new infrastructure demands, and organizations are seeking solutions that are flexible, agile, and efficient to solve them.

Modern workloads have exposed the Achilles heel of traditional server architecture, the sheet metal that surrounds them. With high GPU, Storage and/or FPGA resource requirements, success is effectively limited to what can fit in a box. Customers are forced to purchase larger, more expensive servers than needed for new deployments, or forklift upgrade to them when a server is maxed-out. Don't forget they take weeks to order and deploy and scaling is disruptive. This also interrupts the typical refresh cycle for servers due to the extra cost, and other business outlays such as software licensing for the larger servers as well as power/cooling/space constraints in the data center.

A new technology called Composable Disaggregated Infrastructure (CDI) is eliminating the sheet metal limitations affecting traditional datacenters, and Liqid is at the forefront of this charge. Liqid Matrix™ software allows organizations to address any workload need in seconds by configuring, deploying, and scaling bare-metal systems in seconds from pools of existing compute, storage, and accelerator resources. This paper serves as a CDI primer, and explores core technical concepts that surround this exciting, new technology.

CORE CDI CONCEPTS

The goal of CDI is to extract cloud-like value from the datacenter by applying a new architecture to conventional servers and their resources. With any new technology, there are some new concepts that must be learned, and with CDI it can be broken down into three key ideas:

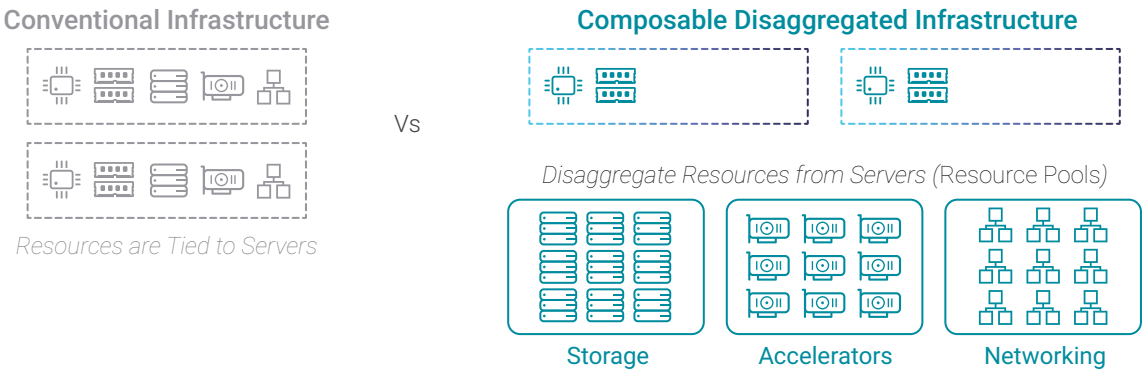
- The disaggregation of a datacenter's constituent infrastructure.
- A fabric to connect those disaggregated components.
- Software that allows the configuration, deployment and scaling of bare-metal systems.

[1] Gartner, "The Current State of Artificial Intelligence and Its Strategic Direction", Whit Andrews, et al, 29 April 2021

Disaggregation

If server sheet metal is the limiting factor to optimally deploying and scaling servers for modern workloads, disaggregation is the first step in thinking outside the server box. This first concept is quite simple. As the name implies, with Liqid CDI PCIe devices including GPU's, FPGA's, SSD's, Storage Class Memory, NIC's are not installed in the server chassis, but are instead disaggregated, and placed into external PCIe enclosures, called expansion chassis. Additionally, servers are now considered compute resources, that contain at minimum CPU and RAM.

Figure 1: Instead of residing in the servers, resources are disaggregated into pools.

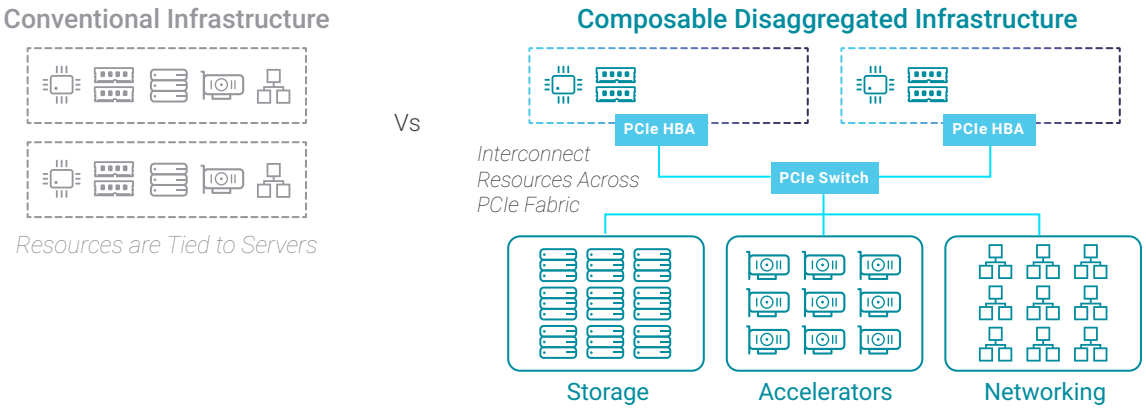


An organization can have as many expansion chassis as are required to hold their storage, accelerators and/or networking resources. Liqid is vendor agnostic, meaning customers get to choose what resources they compose. Liqid's hardware compatibility list is always growing and can be found on our website here.

Fabric

Fabric is what interconnects all the resources to be composed, providing every composable resource direct access to each other. For connections between chassis and racks to be transparent to workloads, Liqid leverages high bandwidth technology, including PCIe (Gen 4 and Gen 3). While Liqid also supports Ethernet and InfiniBand (Eth/IB) fabrics, they are not covered in the scope of this document.

Figure 2: Resources are interconnected via PCIe fabrics

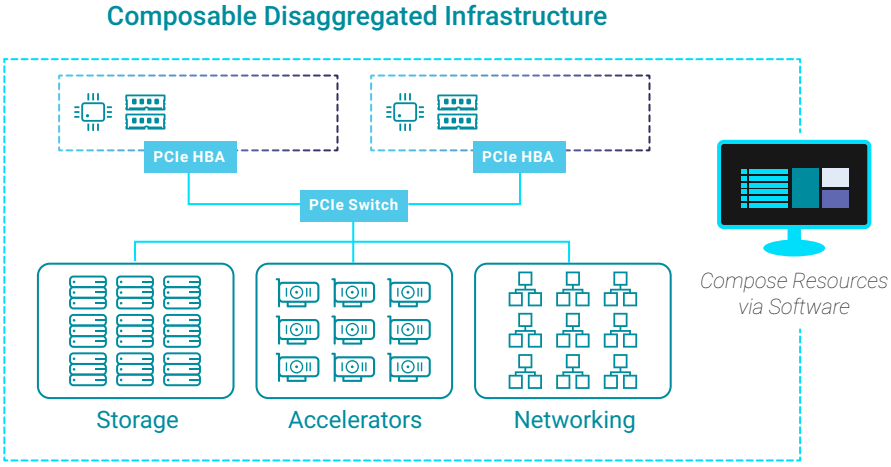


Organizations choose the fabric type that best meets their composability needs and can even create multi-fabric environments that support them all. All expansion chassis support PCIe and a subset support either PCIe or Eth/IB connectivity. Compute resources (servers/blades) are connected to the fabric via PCIe HBA and/or 100GbE network cards. All composable resources connected to high-speed fabric switches, either PCIe or Ethernet.

Software

Once resources are disaggregated and connected over distributed fabric(s), we've essentially flattened the datacenter and turned it into massive computing, accelerator, and storage pools. At this point, Liqid's composable software is used to create bare metal servers composed with resources tuned to meet any workload need, in seconds.

Figure 3: Liqid Matrix software allows for the creation of bare metal servers in seconds.



Liqid Matrix™ composable software controls the fabric though an out-of-band connection, and brokers resource connections via UI, API or CLI. Additionally, Liqid Matrix can be controlled by SLURM Workload Manager and other tools via northbound partner integrations

PCIe PRIMER

One of the primary fabrics that Liqid leverages to disaggregate and compose is PCIe, but what does it mean to have a PCIe fabric? Let's start by defining Peripheral Component Interconnect Express (PCIe). It is a low latency, high data transfer rate serial expansion bus standard for connecting one or more peripheral devices to a computer. Each device with a PCIe link has its own dedicated point-to-point connection, meaning they do not compete for bandwidth with other devices. All of the devices connect to the root complex, which acts as an interconnect between the CPU and memory subsystem, and the bus is expanded to endpoint devices through one or more switches.

The PCIe link between two devices commonly ranges from 1 to 16 lanes based on bandwidth requirements, with each lane composed of two signaling pairs, one for receiving data and one for transmitting. Thus, each lane is composed of four wires or signal traces. In a multi-lane link, the packet data is striped across lanes, and peak data throughput scales with the overall link width. These signaling pairs also allow PCIe to offer full-duplex communication.

While physical PCIe links can contain from 1 to 16 lanes, they must follow powers of 2 (1, 2, 4, 8, or 16). Lane counts are written with an "x" prefix (spoken as "by") – for example x8 for an 8-lane card or slot. Since each lane requires distinct signaling pairs, physical slot sizes also vary based on the number of lanes and are downward-compatible. For example, a x16 slot (the largest in common use) can also support x8, x4, x2 and even x1 cards if required. The reverse is not always the case. For example, a x16 card may not fit in a x8 slot unless it is specially designed to accept it.

As of the time this writing, PCIe 4.0 (Gen4) is the latest iteration of the PCIe protocol commonly available, which sees a doubling in throughput per lane as compared to PCIe 3.0 (Gen3). The below chart shows the previous and future theoretical performance metrics:

Version	Introduced	x1 Transfer Rate	x16 Transfer Rate	Line Code
PCIe 1.0	2003	2.5 GT/s (250 MB/s)	40 GT/s (4.0 GB/s)	8b/10b
PCIe 2.0	2007	5.0 GT/s (500 MB/s)	80 GT/s (8.0 GB/s)	8b/10b
PCIe 3.0	2010	8.0 GT/s (984.6 MB/s)	128 GT/s (15.75 GB/s)	128b/130b
PCIe 4.0	2017	16.0 GT/s (1969 MB/s)	256 GT/s (31.51 GB/s)	128b/130b
PCIe 5.0	2019	32.0 GT/s (3938 MB/s)	512 GT/s (63.02 GB/s)	128b/130b

Commonly, peripheral cards are manufactured in several sizes:

- Low Profile/Slim
- Half Height, Half Length (HHHL)
- Full Height, Half Length (FHHL)
- Full Height, Full Length (FHFL)
- U.2 - Liqid manufactures a Gen 3 x4 host bus adapter (HBA) in the U.2 formfactor for blade servers without PCIe card slots.

LIQID COMPONENTS

Component overview

The following is a short overview of the Liqid components that make up a CDI architecture. For more detailed specifications and descriptions please see https://www.liqid.com/products#elements.

- PCIe HBAs – each compute host needs at least one host bus adapter (HBA) installed to connect to the Liqid PCIe fabric. These typically provide up to 4 x4 ports for a total of x16 PCIe lanes. Supported connectivity options are: 1 port (x4), 2 ports (x8) and 4 ports (x16). There are options for both Gen3 as well as Gen4 HBAs. The cables used to connect the HBAs to the PCIe fabric switches are mini-SAS HD cables, and come in various lengths from 1m to 3m.

» Liqid Expansion Chassis – these hold the various accelerator cards that will be composed to the compute hosts. Liqid has an ever-expanding list of chassis options designed to support a wide range of accelerator and other add in cards including GPUs, FPGAs, NICs, SSDs, etc. These chassis are connected to the PCIe fabric – typically via multiple x16 PCIe connections.

» Liqid Fabric Switches – these provide the connectivity between servers and expansion chassis. Each "port" on a fabric switch can consist of 1 or more physical connections each operating as a x4 PCIe link. For example, a x8 port would consist of 2 physical connections (cables) and a x16 port would consist of 4 physical connections. Unlike an ethernet switch, each port needs to be configured for the type of device that will plug into it. In general, ports on a Liqid switch are configured as one of the following four types:

- **Host Port** - these are the ports configured to connect to the HBA cards installed in the compute hosts.
- **Downstream Port** – these are typically connected to expansion chassis that do not require any advanced management capabilities, for example a NVMe-only storage chassis.
- **Management Port** – these ports are connected to the Liqid Director management appliance. They are used for management of the switch and some chassis by the Liqid Matrix software.
- **Fabric Port** – these ports are connected to expansion chassis that require advanced capabilities such as Peer-to-Peer (P2P) mode. This capability allows devices within a chassis to communicate directly with one another without having to interrupt the server.

» Matrix Software – this is the software that manages connectivity between the various devices on the Liqid fabric. This software runs on a hardened appliance and requires connectivity to the fabric switches and any expansion chassis that requires advanced capabilities such as P2P mode. This connectivity is provided via a dedicated PCIe link to each of the components. Access to the Liqid Matrix software is via a web interface or API accessed via an ethernet network.

A SAMPLE FABRIC

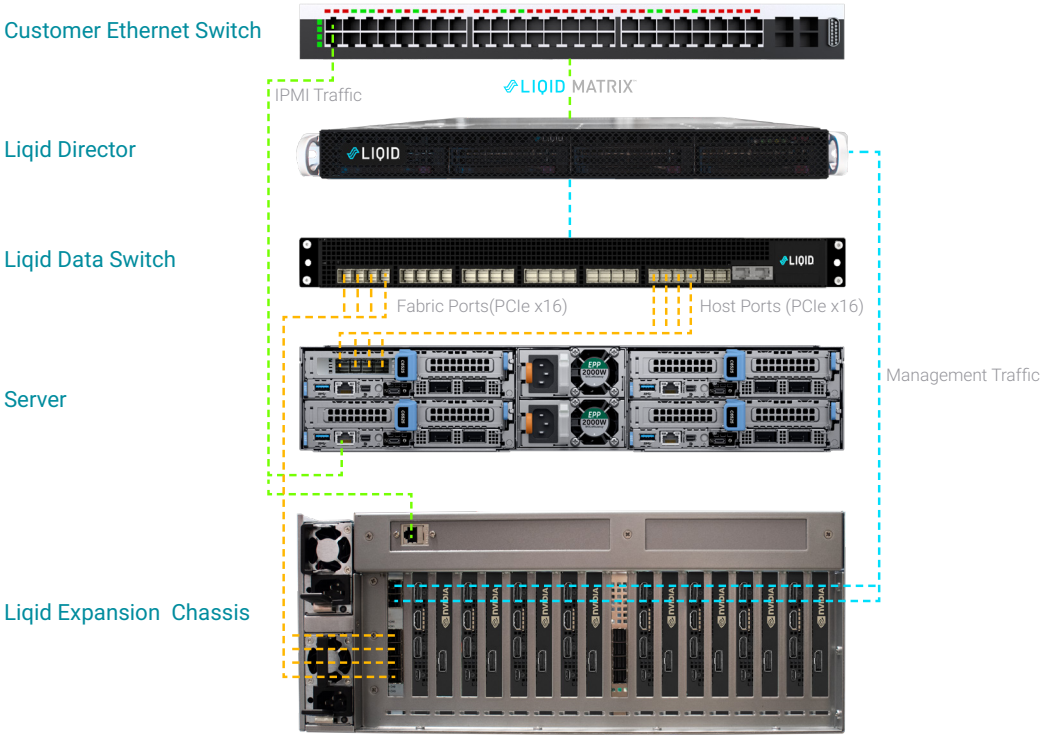
A simple diagram is shown below consisting of the following:

- Liqid Director
- Liqid Data Switch
- Liqid Expansion Chassis
- Single server/host with Liqid HBA card installed

The following lists some of the connectivity between the components:

- The Liqid Director physically connects to the data switch and the expansion chassis via a single dedicated PCIe link to each. This allows the Liqid management software to monitor/manage these components.
- The compute host is connected to the Liqid Data Switch via 4 x4 connections in this example – providing a potential throughput of 32GB/s.
- The expansion chassis is connected via 4 x4 connections in this example for simplicity. Typically, the chassis would be connected via 8 x4 connections (2 x16) to ensure there is no bottleneck.
- The connectivity to the Liqid Web Interface as well as the IPMI management connections would be provided via the customers' existing ethernet infrastructure.

Figure X: Sample CDI Fabric with GPU Resources



ARCHITECTURE CONSIDERATIONS

Device Discovery

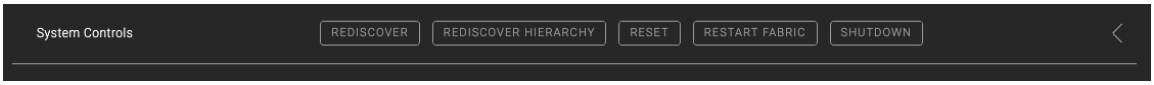
Liqid leverages a standard PCIe discovery process – very similar to how the discovery process works in a standard server. The big difference is that devices are now distributed across the Liqid fabric and not sitting directly in the server. In a typical server, the BIOS enumerates all of the installed PCIe devices and discovers their capabilities, e.g. whether they are end-point devices, bridge devices, what memory registers they use, etc. In a Liqid environment, there are a couple of extra steps needed to ensure things work correctly:

- Proper boot order. When the Liqid system is started (or restarted), it is important to start things in the correct order. This ensures that components such as accelerator cards are powered up and available to be discovered. The proper startup order is as follows:

    1. Power on any expansion chassis in the fabric and wait 20-30 seconds
    2. Power on any PCIe data switches in the fabric and wait 20-30 seconds
    3. Power on the Liqid Director management appliance

A quick note on ensuring proper boot order. In a large environment – or in environments where physical access to the hardware is difficult, Liqid offers a solution. We have partnered with several manufacturers of “Smart PDUs” to create an automated solution for restarting the Liqid fabric.

Figure Y: Liqid Matrix System Controls



With this solution, when the “Restart Fabric” button is pressed in the Liquid UI, Liqid will make API calls to the smart PDU and tell it to power down the system, and power things on in the proper order with a pause in between each component. This can make recovering from events such as a power outage much easier.

One additional thing to consider is the fact that most PCIe devices (specifically GPUs) require dedicated memory space to be allocated in the server’s memory for various uses such as configuration of the device or to load code into. This allocated space is known as the Base Address Register (BAR) in the PCIe specification. In a

Liqid environment, Liqid allocates this address space to any server connected to the Liqid fabric. This is a tunable parameter in Liqid and can be adjusted based on the type and number of specific devices in the system.

Ensure any accelerator devices used are on Liqid's Hardware Compatibility List (HCL) and they have an entry in the Matrix local device database. Liqid utilizes this device database to help it determine what devices are available on the fabric and their capabilities. As Liqid starts up it will also begin a device enumeration across all the devices in each of the chassis. This process allows Liqid to understand what devices are present and functioning on the fabric. Liqid can also query servers and other equipment via IPMI in order to gather information about the devices capabilities.

To see a list of all of the devices that Liqid has currently tested, please see the Hardware Compatibility List at: https://www.liqid.com/resources/all. New devices are added constantly based on customer request or new products/technologies entering the market.

PCIe Device Limits

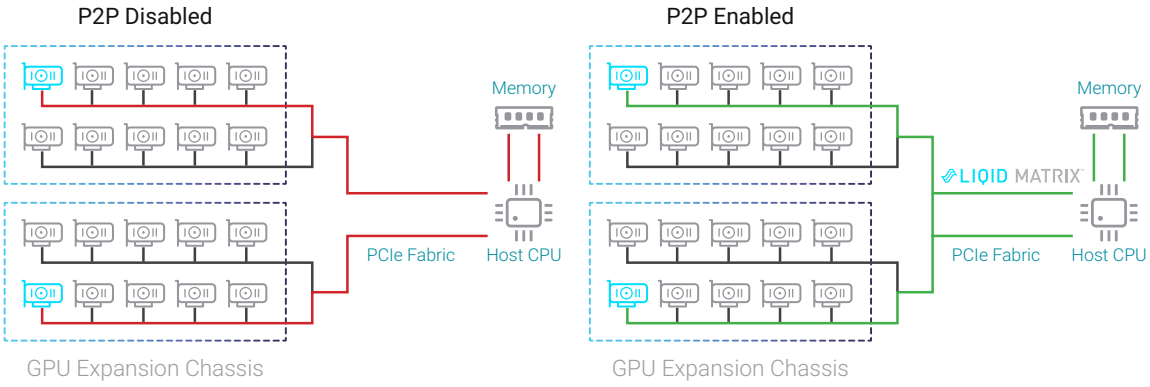
The current PCIe specification limits the number of devices that can be on a single PCIe bus to 255. This is one of the constraints to keep in mind when designing a Liqid solution as it can change how larger systems are designed and deployed. This does not typically impact smaller solutions however, for larger solutions such as HPC deployments, we typically break the larger solution into multiple smaller "pods". Each pod would be a "composability domain", meaning any server in that pod could compose any device in that pod, however a server in one pod could not compose devices from a different pod. This pod architecture also allows for maintenance of part of the overall system without impact to other pods.

Peer to Peer

Customers who leverage GPUs to perform data processing are most likely familiar with Nvidia's NVLink – which "links" one or more GPUs together via a high-performance bridge to allow very fast data transfers between the GPUs. Liqid leverages the native capabilities of the PCIe specification to deliver similar functionality. When this functionality enabled, communications between GPUs or even between GPUs and NVMe devices within a chassis can occur directly without

having to go back to the server. This feature can greatly improve transfer speeds, but does require that the customers' application support this capability. A quick diagram of how this functionality works is shown below:

Figure Z: Liqid P2P enables direct communication between GPUs



Data must flow through x86 Root Complex

Data bypasses CPU increasing performance

The performance improvement using this technology is significant, nearly a 300% increase in bandwidth between devices and response time is improved by more than 1000%.

This functionality has recently been extended to devices outside of a single chassis as well. In this instance a device in one chassis would be able to communicate directly with a device in a 2nd chassis. Communication would take place via the upstream switch and would continue to bypass the compute node.

SR-IOV

SR-IOV or Single Root IO Virtualization is a native feature of the PCIe specification and allows virtual functions of a PCIe device to be shared. This is most commonly used to share a physical PCIe device among a number of virtual machines. In the case of VDI, a physical GPU can be shared among a number of virtual systems allowing better resource utilization. Liqid Matrix allows a PCIe device to be composed via the Liqid fabric and also allow the composed compute host to make use of SR-IOV functionality on the composed devices.

PUTTING IT ALL TOGETHER

An Example

Let's walk through an example. We will use the same topology we used previously and assume the expansion chassis has several GPUs installed. In our example, we will walk through the process of composing GPUs from the expansion chassis to a server connected to the Liqid fabric as shown below. The process is as follows:

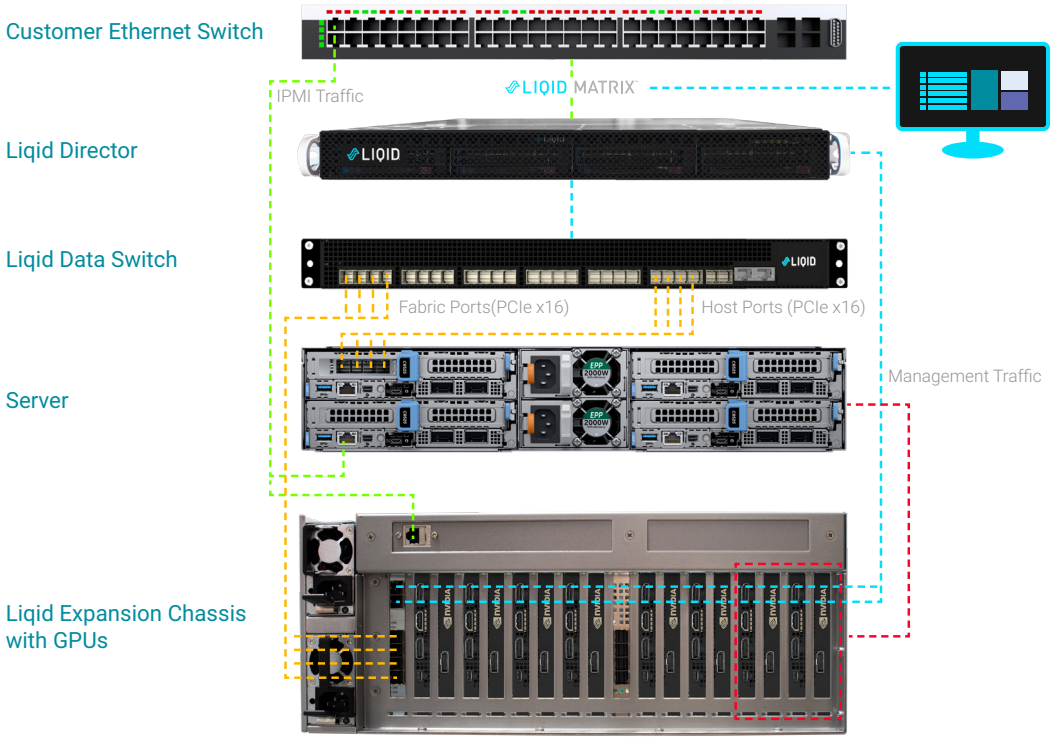
1. The customer logs into the Liqid Command Center web UI and locates the appropriate server that requires the GPUs. Resources such as accelerator devices and servers are partitioned into groups in the Liqid interface. This allows for multi-tenant capabilities and ensures that a user only has access to the resources assigned to that group. The customer will highlight the proper server and assign one or more accelerators from the resources assigned to that group.

2. The customer then clicks "Reprogram" button. This causes Liqid to communicate with each of the devices in the path (expansion chassis, PCIe fabric switch and HBA) and establish an electrical connection via the PCIe fabric. This is analogous to physically plugging the device into a server and takes ~15-20 seconds. There is no software installed on the server at any time during this process.

3. The desired devices will appear on the server's PCIe bus ready for use. At this point the customer can install any software or drivers needed to operate the device and utilize it as if it were a physically installed device. Performance will be nearly identical to a locally installed device as the Liqid PCIe switch only adds ~100 nanoseconds of latency.

Figure ZZZ: Understanding the Composability Process

1. Users request to assign GPUs to a server console.
2. Liqid OS communicates with chassis, switch and HBA to establish direct electrical connection.
3. Requested GPUs are presented to the server over the PCIe links without any drivers.



For more detailed deployment scenarios, download the ESG Technical Validation of Liqid here.

SUMMARY

This paper was written to provide a deeper understanding of how Liqid's CDI solution operates behind the scenes, and considerations for designing a PCIe fabric powered by Liqid Matrix.

The Liqid Matrix has the power to change how organizations acquire, deploy, and scale infrastructure. While public cloud continues to be a popular direction in many situations, Liqid allows organizations to build flexible private cloud solutions that drive superior resource utilization from their existing infrastructure and are often more flexible and cost effective than public cloud options.

COMPOSABILITY OVER PCIe EXPLAINED

The era of composable infrastructure has begun, and organizations are continually discovering new ways to shape their datacenters, whether on-prem, remote or at the edge. We are on the brink of new technological breakthroughs that will unlock more flexibility and performance than ever possible. These include:

» PCIe Gen5 – doubles the performance of Gen4 PCIe
» PCIe Gen6 – again doubles the performance of Gen5
» Compute Express Link (CXL) - https://www.computeexpresslink.org/
» GenZ - https://genzconsortium.org/

Liqid will continue to develop our software and compatibilities in a way that delivers maximum value, and exceeds modern workloads demands. If you would like learn more, please visit our website at https://www.liqid.com and request a demo.

About Liqid

Liqid provides the world's most-comprehensive software-defined composable infrastructure platform. The Liqid Composable platform empowers users to manage, scale, and configure physical, bare-metal server systems in seconds and then reallocate core data center devices on-demand as workflows and business needs evolve. Liqid Command Center software enables users to dynamically right size their IT resources on the fly.