

# Technical Insight Report

## From Underutilization to Optimization Liquid GPU On-Demand Customer Stories

**By Dave Raffo, Sr. Analyst**

**February 2023**



**Evaluator Group**

*Enabling you to make the best technology decisions*

## Overview

Graphics Processing Units (GPUs) have become a crucial element of high performance computing technology, especially for AI, deep learning and visualization applications. GPUs share the work of CPUs in servers, improving performance and increasing the ability to run calculations or processes simultaneously for parallel applications by supercomputers and other HPC devices.

However, conventional server form factors limit GPU density and scalability, and their siloed design trap and underutilize them. Underutilization of these costly GPUs can place a financial burden on academics and non-profit research labs.

Liquid, a Colorado-based startup company, provides GPU capabilities on demand through software and PCIe switching. Liquid's GPU On-Demand technology enables the software provisioning of bare-metal GPUs into servers to focus on HPC applications. It separates physical GPU, compute and sometimes storage resources, and allocates them through software to where they are most needed. This process – sometimes referred to a composable infrastructure – can return these resources to a common pool so they can be reallocated when demands change.

Allowing customers to provision, scale and migrate GPUs in real-time via software can:

- Maximize performance by provisioning and scaling GPU to meet the most demanding workload requirements.
- Increase agility by deploying GPUs when and where needed in seconds, and then migrate them to other servers when the demands change.
- Improve efficiency by increasing GPU utilization through provisioning exact quantities and migrating them as needed. It also reduces time consuming manual configuration tasks.

Liquid GPU On-Demand includes:

- Liquid Matrix software that enables and manages shared infrastructure across fabrics such as PCIe, Ethernet, and InfiniBand
- Liquid SmartStack systems that allow the sharing of up to 30 GPUs with up to 16 servers
- IO Accelerators which are NVMe devices that includes 3D NAND Intel Optane memory
- Liquid PCIe switches with 24 or 48 PCIe ports and 1U Liquid Directors that run Liquid Matrix software.
- Liquid expansion chassis that contain a customer's disaggregated GPUs and Liquid's IO Accelerator NVMe flash and Intel Optane memory devices.

Liquid's technology is designed to create high-performance, rack-scale infrastructures for applications, such as AI and ML. Where most server chassis can only support a limited number of GPUs, Liquid's composable architecture can incorporate large numbers of GPUs into a dedicated server instance.

Evaluator Group spoke with three Liquid customers who described the real-world benefits of using its GPU On-Demand technology in production. The customers confirmed that Liquid helped in the areas of management efficiencies, performance, cost and sustainability.

## User Case Studies

### Case Study No. 1 – Orange Silicon Valley: Liquid allows teams to work in parallel without conflicts

Orange Silicon Valley – the U.S. subsidiary of the French-based Orange Group telco – used Liquid GPU On-Demand technology to build one of the world’s fastest single-node deep learning supercomputers. In 2021, Orange put together a supercomputer using a Dell server, Liquid Matrix CDI software and composable fabric and 16 Nvidia A100 GPUs – eight in two separate JBOGs (Just a Bunch of GPUs) with Liquid’s 16 TB NVMe for Deep Learning cache for storing training data. Orange’s Liquid Matrix system achieved some of the fastest deep learning performance benchmarks ever documented.

Soumik Sinharoy, a principal in the Orange Silicon Valley Technology Group, said Orange also has about 150 data scientists sharing a system using Liquid technology in France. He said the composability allows multiple teams to work in parallel “without stepping on each other’s foot.”

Sinharoy said Orange uses Liquid to meet increased demands for GPU with existing server investments. Before deploying Liquid, Orange built a supercomputer for researchers with 2 CPUs and 20 GPUs. When multiple teams shared GPUs, they also had to share the same CPU, memory and bus architecture. That meant two teams had to divide the GPUs, so if one team needed more resources for a project it could not use GPUs attached to another system, even if they were not required by the second team at the time.

With Liquid’s GPU On-Demand approach, Orange has 20 GPUs in an external GPU tray so they can be dynamically allocated to bare metal servers based on needs of project teams. If multiple projects share the same GPUs and one team’s project becomes a top priority, the teams can quickly reallocate resources from the other projects.

“A team working on a high priority project gets more GPUs assigned, and then GPUs are taken away when they no longer need them,” Sinharoy said. “It provides a lot more wiggle room.”

He said that flexibility gives Orange at least 90%, and sometimes 100% utilization of the GPU assets.

“I’d rather have a server go idle than GPU go idle,” Sinharoy said. “If Team A and Team B were using the GPUs from the GPU tray, and Team B has no job, the server for Team B goes idle. Now we can allocate all 20 GPUs to Team A. That way a server is idle instead of having idle GPUs. The researchers don’t care about idle servers because CPU and memory is not that important for deep learning projects. Most of the work is GPU-intensive.”

### **Operational cost reduction**

As a telco, Orange Silicon Valley is a CAPEX-heavy organization that invests a lot in infrastructure. More efficient use of infrastructure resources provides a return of investment in the long-term. “Idling servers cost us less money than idling GPUs because GPUs cost so much more money,” Sinharoy said.

Sinharoy said without GPU On-Demand, the bottleneck caused by sharing the same host reduced GPU ROI. That prompted Orange to look for a new way of sharing GPU resources. But competing GPU-optimized servers and workstations would require them to buy more GPUs at about \$10,000 to \$15,000 per GPU.

“It was cheaper for us to add more servers than adding more GPUs in a workstation,” he said. “For workloads that require high GPU density, expensive GPU optimized servers are needed. With Liquid we were able to build the same compute capability at 50% cost of a GPU chassis. Plus, we get the additional benefit of flexibility of resource sharing. For machine learning/AI workloads, we have GPUs at 100% utilization. When the same GPU server [without Liquid] is shared between multiple processes sharing the same GPU system, contention for server CPU and memory resources drops the utilization rate 30-40% per GPU.”

### **Sustainability**

With operations in France as well as the United States, Orange is mindful of power consumption. Sinharoy said the Liquid technology helps there, too, by reducing cooling. France and the European Union have ambitious energy reduction goals.

“The cooling is more efficient, because in the multi-GPU servers with CPU, memory, and GPUs all packed into one, you have to use the highest speed on the fan to cool everything together,” Sinharoy said of his Liquid system. “But the problem is, GPUs will probably need higher CFM (Cubic Feet Per Minute) than the CPU and the memory needs. So, we were using extra cooling power for something that is not required. Once we use separate servers and use GPU trays only designed to host and cool and commit to GPUs, we’re basically appropriating cooling power as required to a specific enclosure. We haven’t measured it with a metric yet, but it is obvious that we have increased the cooling efficiency for our overall rack.”

## Case Study No. 2 -- Electronic Visualization Laboratory (EVL) at University of Illinois Chicago: Higher utilization, lower administration costs for workload diversity

The EVL at UIC used a \$1 million major research instrument grant from the National Science Foundation in 2019 to purchase its COMPaaS DLV system with Liquid technology, Dell compute and Nvidia GPUs.

The system consists of 64 GPUs (32 Nvidia V100s and 32 Nvidia T4s), 24 CPU nodes, NVMe storage, 100G networking and Intel Optane memory.

“Adding 64 GPUs to the campus gave us the largest number of GPUs of any installation at the time,” said Lance Long, senior research programmer at EVL.

COMPaaS DLV is an acronym for Composable Platform as a Service Instrument for Deep Learning & Visualization. EVL specializes in research, development, deployment, technology transfer and training in visual data science. The main focus is high-performance visualization, analytics, virtual reality and advanced computing and networking.

Long said COMPaaS DLV users include researchers, faculty and students from computer science, engineering and other groups. COMPaaS DLV serves a diverse set of needs – some users run containerized workflows on Kubernetes, some require bare metal servers and others use a Jupyter Notebook web-based interactive computing platform.

### Administrative Efficiency

Long said composability allows him to support researchers on one type of system. “Instead of having a rack of bare metal servers that require manual changes to components for every researcher, or having just a Kubernetes system with container workflows, or large CPU nodes providing Jupyter resources. This allows me as the only system administrator to manage everything, especially remotely. I can change hardware for users remotely and quickly, instead of spending a day reconfiguring a rack of bare metal servers.”

### Increased Resource Utilization

Long said with previous systems, he would look at how many GPUs can fit on the motherboard, and then fill the computer up with whatever it could handle. But in practice, one researcher may tie up many of the GPUs, CPUs or the network, and no one else could use the system. His lab would need separate systems for container workflows and bare metal research. With Liquid, he can use the same system and have one admin to manage it. Instrument grants are primarily for equipment only, so they would need another grant to hire more admins at a cost of more than \$100,000 per admin. With COMPaaS DLV, two undergrad students handled most of the additional administration needs.

Because Liquid is designed for AI and machine learning workloads, Long said his researchers complete their analyses faster, they can do more iterations, and can prepare their data for larger environments faster. It allows them to create better data results and run their software faster.

“You give somebody more GPUs and better GPUs, hopefully they’re going to get better results,” Long said.

### **Flexibility**

The Liquid GPU On-Demand technology also allows students to take advantage of the cutting edge, data center GPU technology. Without it, researchers often buy students a computer with a low-end or mid-range GPU to keep costs down. The student would have to set up and maintain the computer, as well as conduct research. Long said he was often responsible for supporting many repetitive problems that the students encounter.

With Liquid, “now the students have immediate access to a large number of GPUs that they did not have access to before. And these GPUs are designed for AI and machine learning workloads. The GPUs that they normally use on their desktop system are not.”

Long said alternative systems such as workstations with multiple GPUs or running a group of standalone servers with GPUs may provide more GPU per dollar, but require more support with limited flexibility and lower utilization.

“If you have a single type of workload, then filling servers with GPUs is great,” he said. “But if you have a lot of workload diversity, then those systems can get bottlenecked. So sure, you might get more GPUs, but your utilization would be lower and administration costs would be a lot higher.”

### **Case Study No. 3 -- Genomics Lab: “Hands-down different than anybody else”**

A genomics lab that studies human and plant genomics purchased a Liquid enclosure to use with several high performance file storage systems in 2020. The lab has petabytes of data and data sets that can include 200 GB files.

The lab acquired an eight-slot Liquid expansion chassis with four A100 GPUs, two RTX 8300 GPUs and two Intel Optane SSD cards. One of the lab’s system architects said Liquid’s GPU On-Demand capability provides flexibility he couldn’t find from any other system.

### **Maximum Resource Utilization**

“Liquid’s most promising feature is composability,” he said. “We have GPUs in an enclosure, cable them to a number of servers, and then have software that can move resources around the servers on the fly.”

That is hands-down different than anybody else is doing. One you might have an application that needs eight GPUs and the next day you have four applications that need two each, and when you don't have to go to the data center and pull something out of a rack and move a card around ... that's really nice."

He said his lab frequently needs to make changes like that and has a limited supply of GPUs because of their cost. "There are use cases where I want to pile all four A100s in one server for an application because it scales that way," the system architect said. "Other times, I have an application that will only use one GPU. So, it's nice to spread them out and let them run multiples out a time."

He said options included pricey GPU-centric chassis, or smaller rackmount servers. The first option cost too much – "We're a non-profit, so every dollar counts" – and the second wouldn't provide enough GPUs. The system admin said there were several ways to get performance on par with his Liquid system but were more difficult to manage than Liquid's GPU On-Demand technology.

"I come from an older school, where if I can get 90% to 95% of possible performance and it's easy, that's better than getting 98% of performance and it takes weeks to set up and maintain," he said. He said reduced power is an "intangible benefit" he gets from Liquid's GPU On-Demand. "When you move GPUs to an external enclosure instead of a server they cool better, they don't require as much power, and they don't get as hot. That can be a big consideration, depending on where you are located."

## Evaluator Group Opinion

Outside of minor critiques around things such as an interface described as a "little immature, but will get there," according to one user, Evaluator Group found Liquid GPU On-Demand technology filled a need for HPC that few vendors can address.

Liquid's sweet spot is high-performance applications that use multiple GPUs. By composing GPUs, Liquid can overcome the problem of limited server slots for GPUs. Customers we spoke to found it easy to shift resources among workloads. It also does so while using industry-standard GPUs, storage, compute and networking.

Customers found Liquid meets the vendor's goals of maximizing performance, increasing agility, and improving efficiency and sustainability while also lowering total cost of ownership in some cases. These factors will become even more important as GPUs continue to play a larger role in HPC and data center environments to cope with demanding AI and ML applications. Accomplishing this in a vendor-agnostic manner without requiring drivers, agents or software modules on compute nodes enable customers to run these applications on a wide variety of hardware and networking fabric. Liquid's GPU On-Demand approach can be especially appealing for companies who run performance-based applications such as AI and ML.

## About Evaluator Group

Evaluator Group Inc., an Information management and data storage analyst firm, has been covering systems for over 20 years. Executives and IT Managers rely upon us to help make informed decisions to architect and purchase systems supporting their data management objectives. We surpass the current technology landscape by defining requirements and providing an in-depth knowledge of the products as well as the intricacies that dictate long-term successful strategies.

### ***Copyright 2023 Evaluator Group, Inc. All rights reserved.***

*No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or stored in a database or retrieval system for any purpose without the express written consent of Evaluator Group, Inc. The information contained in this document is subject to change without notice. Evaluator Group assumes no responsibility for errors or omissions. Evaluator Group makes no expressed or implied warranties in this document relating to the use or operation of the products described herein. In no event shall Evaluator Group be liable for any indirect, special, consequential, or incidental damages arising out of or associated with any aspect of this publication, even if advised of the possibility of such damages. The Evaluator Series is a trademark of Evaluator Group, Inc. All other trademarks are the property of their respective companies.*