FOR IMMEDIATE RELEASE

Drut Technologies Delivers Virtual Disaggregation for GPU Servers

Nashua, NH, – September 9, 2024 – Drut Technologies, today announced the launch of their DX 3.0 advanced software system, which offers virtual disaggregation for datacenter GPU servers.

Virtual Disaggregation modernizes GPU developments by making it possible to deploy faster, configure easier and finely allocate a precise number of GPUs for the most complex enterprise AI/ML applications.

This software deploys on GPU servers and enables the creation of isolated GPU resource pools within and across distributed physical servers. Servers that do not have enough GPUs can be augmented with GPUs from other servers. Servers that have too many GPUs can be divided into small GPU machines for ML/AI inferencing use cases and dedicated to specific users and workloads.

DX 3.0 is an advanced server software system that incorporates the unified network fabric to deliver what is called a vPOD. A vPOD is a cluster formed by grouping a number of CPU slices, memory slices, OS's, GPUs and NICs interconnected to form an isolated topology. Once deployed an AI/ML model is used to benchmark and verify the infrastructure. Once verified, it can then be used to deploy a user's AI/ML model of choice. vPODs can exist within the server and across servers over the network fabric, whether Ethernet, InfiniBand, or using Drut's photonic fabric.

Benefits and Markets:

- **Comprehensive Hardware and Software Integration**: DX 3.0 offers a complete solution that integrates seamlessly with new and existing server infrastructure, enabling quick and efficient AI model deployment. Customers can use traditional network designs or leverage the benefits of a photonic fabric.

- **GPU Isolation** - DX 3.0 allows users to run isolated GPU workloads in multi-GPU servers and across separate GPU servers. This capability allows users to dynamically isolate GPUs in multi-GPU servers to create a physically isolated resource group which optimizes utilization, reducing contention and resource stranding.

- **Resource Configurability** - In a multi-GPU server, DX 3.0 delivers efficient GPU resource utilization by providing the ability to isolate the minimum number of GPUs for a workload, while deploying the remaining GPUs for other workloads. Have an eight GPU server, but only need six GPUs, create three vPODs. One vPOD with six GPUs and two vPODs of one GPU each, to serve more application workloads simultaneously while driving resource utilization to maximum efficiency

- **Flexible Resource Groupings** - DX 3.0 provides the ability to carve up multi-GPU servers into smaller machines as well as combine GPUs from physically different machines into a vPOD.

- **Compatibility with Off the Shelf Servers:** DX 3.0 utilizes standard servers and RDMA-capable NICs, ensuring a high-fidelity GPU utilization without specialized hardware.

- **Diurnal Resource Groupings** - DX 3.0 delivers the ability to dynamically reconfigure the GPUs for diurnal workloads. GPUs can be configured into a vPOD for a user during the day and redeployed into a different vPOD for overnight workloads. No need to GPUs to be under-utilized at night when users are away.

- **Dynamic GPU Allocation**: Users can easily add GPUs to their defined vPODs, providing the flexibility to handle a variety of AI workloads.

- **Cost Efficiency**: By leveraging existing hardware and efficiently optimizing resource allocation, DX 3.0 delivers ML/AI utilization at a significantly lower cost than traditional solutions.

Focused on the rapidly expanding AI infrastructure market, DX 3.0 promises to bring easy to deploy, easy to configure and cost-effective utilization of GPUs for ML/AI workloads. Ideal users are enterprise organizations deploying increasing amounts of GPUs for internal users, GPU as a Service Providers as well as AI as a Service providers needing to deliver fine-tuned GPU instances for customers. Cloud Service Providers (CSPs) deploying GPUs for edge data services will find DX 3.0 useful in allocating GPUs for edge workload demand.

"DX 3.0 represents a significant leap forward in making cost effective use of some of the most expensive resources in a modern data center for enterprise customers," said Drut Technologies CEO, William Koss. "By providing a scalable, cost-effective solution, we are enabling businesses to harness the power of ML/AI infrastructure without the prohibitive costs associated with current deployment methods. GPU server virtualization and the associated tools to integrate GPUs with the GPU interconnect fabric are the foundations that users need to make GPU consumption ubiquitous and affordable."

Supporting Quotes:
“With Drut’s DX 3.0, data center operators can allocate the precise hardware resources required to service a specific AI workload, eliminating the need to overbuy GPUs. What’s more, Drut can interconnect vPODs over an existing data center fabric, although their photonic fabric is an intriguing upgrade to reduce job run times. While hyperscalers have the unique challenge of crunching numbers on trillions of tokens, Drut is offering an AI fabric for the rest of us.”
- Ethan Banks, Founder & CEO, Packet Pushers Interactive

"Drut Technologies with their DX 3.0 launch are revolutionizing GPU Server Virtualization with Advanced Virtual Disaggregation"
- Jai Kakkar, CTO & Chief Architect, SAN Data Systems Inc.

“The release of version DX 3.0 is excellent news. Drut demonstrates that innovation and customer focus are at the core of their DNA. Indeed, this solution, which complements DX 2.0, will allow us to address a broader range of companies by pooling their existing GPU servers through a 'software-only' solution. Kudos to Drut; we are delighted to be part of your ecosystem.”
- Raphael Maurice is Co-Founder and Director at Celeris Informatique

“Drut DX solution and its new 3.0 release fits in our strategy, which since 2002 has actively expanded our product portfolio with leading edge technologies and promoted the adoption of new and innovative computing and storage technologies" said Cosimo Damiano Gianfreda E4 Computer Engineering co-founder & CEO. "Thanks to the DX solution, E4 will have an additional technology to propose its HPC and Enterprises clients, based on an agile, scalable and affordable solution to meet their growing need to support AI outside of the Public Cloud. Drut DX provides the most efficient and cost-effective answers to what is becoming a critical problem: how to grow step by step my GPUs base while making it available potentially to all my servers."
- Cosimo Gianfreda E4 Engineering Co-Founder & CEO

About Drut Technologies
Drut Technologies is a leading provider of advanced data center solutions, specializing in photonic interconnects and GPU virtualization technologies. Based in Nashua, New Hampshire, the company is dedicated to pushing the boundaries of high-performance computing and AI infrastructure. Drut's industry-leading approach empowers clients to harness the power of AI computing by offering greater control, optimized performance, and significant cost savings. Drut Technologies operates globally with a strong presence in the United States, India, and Europe.

Availability
DX 3.0 is available Q3 2024. For more information or to schedule a demo, please visit www.drut.io or contact our sales team at info@drut.io

Media Contact:
Simon McCormack
simon@drut.io