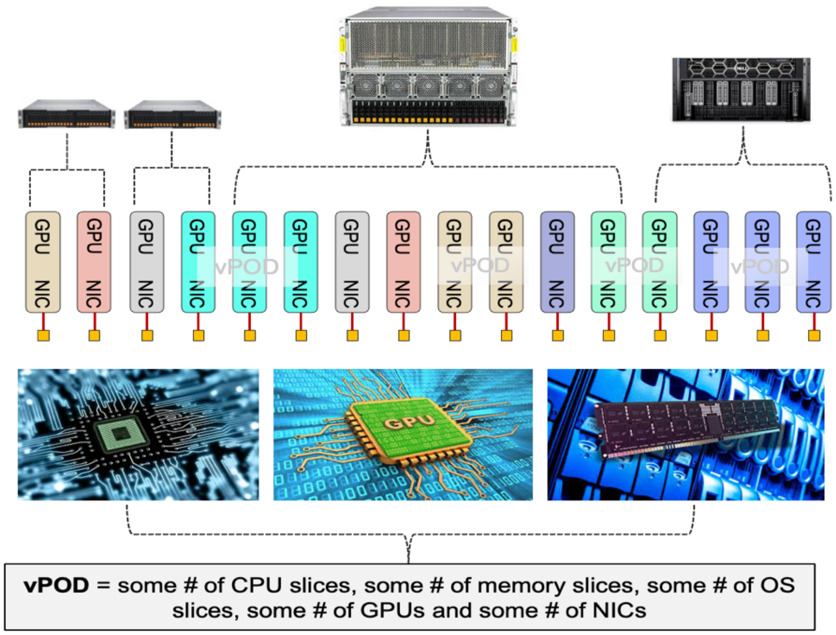
DX 3.0 Advanced Software System

Key Benefits

- **GPU Isolation**: Allow users to run isolated GPU workloads in multi-GPU servers
- **Resource Configurability**: Efficiently allocate GPUs in multi-GPU servers
- **Flexible Resource Groups**: Divide multi-GPU servers and combine GPUs from different machines
- **Existing Server Augmentation:** Combines old and new GPUs into virtual pods, optimizes GPU usage
- **Diurnal Resource Grouping:** Lets you adjust GPU configurations for time-of-day workloads
- **Cost Efficiency**: AI/ML utilization at a significantly lower cost points by improving the utilization of GPUs for workloads

DX 3.0 is an advanced server software stack that provides virtual disaggregation capability for GPU servers. The base layer includes a server software operating system and an isolated resource partitioning environment, wrapped into a machine deployment mechanism driven by the Drut Software Platform (DSP). GPU servers are created by the Drut Fabric Manager (DFM), building vPODs out of the available machine partitions.

A vPOD is constructed out of a number of CPU slices, memory slices, operating systems, GPUs and NICs deployed with 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. For the user, we glue all this together with our AI/ML workbench software and automate the configuration process. It is the easy button for AI/ML.



vPOD = some # of CPU slices, some # of memory slices, some # of OS slices, some # of GPUs and some # of NICs

Features

- **Hardware and Software:** Bring your own hardware, with optional Photonic interconnect
- **System Deployment:** Base OS, PXE, DHCP and DNS
- **GPU Server Isolation:** Dynamically isolate GPUs in multi-GPU servers to create physically isolated resource groupings (vPODs)
- **User Isolation**: Partitions GPU machines into specific user/workload instances
- **GPU Allocations:** Add GPUs to existing vPODs, or re-allocate to new vPODs
- **Off the shelf servers** : Compatibility with industry standards server, GPUs and NICs
- **Interconnect**: Fully automated photonic interconnect option, for minimum latency direct connect

Enterprise Customers: Enterprises looking to improve GPU utilization will find that DX 3.0 provides a scalable and cost-effective way to integrate AI/ML into their operations. Whether users are financial traders, digital artists, data scientists, machine learning engineers, business analysts or AI researchers, DX 3.0 provides the ability to efficiently deploy GPUs in an isolated vPOD to provide guaranteed resource pool.

Cloud Providers: GPU as a Service Providers can empower businesses by offering precise GPU isolation and allocation through virtualization, enabling efficient resource management and tailored solutions for diverse workloads.

AI as a Service: Enterprises with internal users and Cloud Service Providers with external users will find that DX 3.0 will be an important software component for AI as a Service. DX 3.0 will create the user vPOD around the required GPUs, load the LLM of choice and deliver the vPOD to the user as if it was their personal bare metal GPU server.

Complete Solution – GPU Server/Interconnect Integration

DX 3.0 is about combining valuable resources into resource groupings called vPODs. vPODs can span across servers, across network fabrics (e.g. InfiniBand, Ethernet, Photonic) and different generations of GPUs and suppliers.

Efficiency

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 (8) GPU server, but only need five GPUs, create three vPODs. One vPOD with five GPUs, one with two GPUs, and one with one GPU.

Isolated and Performant

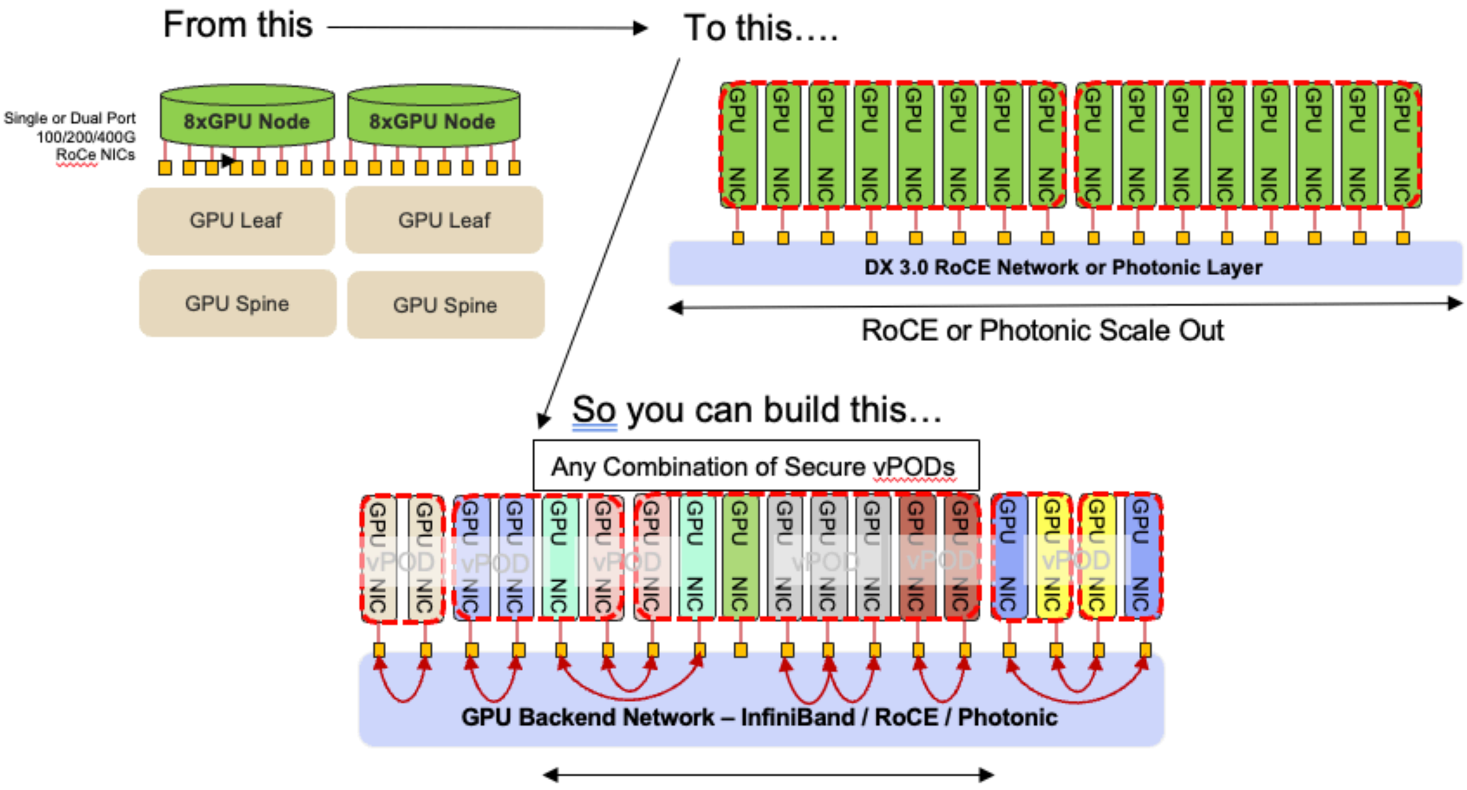
Using tried and tested industry standard server technology, DX 3.0 provides resource passthrough and isolation, offering dedicated and performant peripherals attached to your vPODs machines.

Why Choose DX 3.0?

We view DX 3.0 as the beginning of our journey to help customers fully exploit one of the most valuable resources in a modern data center. A DX 3.0 standout feature is its ability to slice multi-GPU servers, allowing multiple users to share GPU resources efficiently. This optimization allows enterprises to ensure that GPU resources are utilized to their fullest potential.

Users can dynamically allocate additional GPUs to their defined vPODs, providing the flexibility to scale AI capabilities according to their specific needs. This adaptability is crucial for handling varying AI workloads and maximizing performance.

Utilizing off-the-shelf hardware DX 3.0 leverages standard servers with GPUs and RDMA capable NICs, which means businesses can avoid the high costs associated with specialized hardware. This approach not only reduces initial capital expenditure but also ensures high-fidelity GPU utilization. Even older systems can be combined with newer systems in a vPOD.


From this
To this….
Single or Dual Port
100/200/400G
RoCe NICs
8xGPU Node
8xGPU Node
GPU Leaf
GPU Leaf
GPU Spine
GPU Spine
GPU
NIC
DX 3.0 RoCE Network or Photonic Layer
RoCE or Photonic Scale Out
So you can build this…
Any Combination of Secure vPODs
vPOD
GPU Backend Network – InfiniBand / RoCE / Photonic

DX 3.0 System Requirements

DX 3.0 is a server and interconnect fabric product. It is designed to operate as a system allowing for virtual server disaggregation via the creation of vPODs. DX 3.0 is deployed on a server as operating system and hypervisor. Two additional appliances with ~80 cores each is required to run Drut's Software Platform that includes management software, fabric manager and AI/ML workbench software. Backend GPU network can be InfiniBand, RoCE or an all-photonic fabric from Drut Technologies.

Conclusion

We think that DX 3.0 is set to revolutionize the AI/ML landscape by making powerful AI/ML processing accessible and affordable for enterprises. With its innovative features, cost-efficient design, and seamless integration capabilities, DX 3.0 empowers businesses to harness the full potential of AI/ML while driving innovation and competitive advantage in their respective fields.

For more information on the Drut DynamicXcelerator please review our website https://drut.io

For more detailed information contact info@drut.io