

# Creative Consultants Taps InfiniBand to Deliver Efficient GPU-based HPC Cluster

*Takes Multi-GPU Parallel Computing Out of the Box, Rivaling Performance of Host-based Systems and Helping Researchers Get More Work Done in Less Time*



## CASE STUDY

### Summary

Creative Consultants LLC, is an Albuquerque-based computer consulting firm that produces efficient performance computing for scientists and engineers. The company designs and supports high performance systems and specializes in heterogeneous and stereoscopic solutions.

The team at Creative Consultants set out to find the optimal architecture to achieve parallel efficiency with GPUs. They used InfiniBand interconnect technology to build a GPU cluster. The resulting system provided the same computational power per GPU across a network of nodes as could be gained by joining multiple GPUs together via a PCIe bus in a single chassis.

Taking parallel computing out of the box without compromising performance expands the possibilities in HPC and can result in systems that help researchers get more work done in less time

### Challenge

Creative Consultants' mission is to build HPC systems for its research customers that deliver the most compute power and efficiency for the investment.

The company sought to push the envelope in the field by using GPUs in new and innovative ways. GPUs can theoretically deliver an order of magnitude more FLOPS than CPUs. However, to achieve these benefits, Creative Consultants needed to develop and tune an appropriate architecture and port application code to work on the new configuration.

They started out by using NVIDIA

GPUs and CUDA (compute unified device architecture), a parallel computing architecture developed by NVIDIA, and experimented with various software and hardware combinations to optimize the system performance.

When working with one compute node and one GPU, it is a relatively simple challenge to gain more power: just add more GPUs on the motherboard. Growing the system beyond four GPUs requires going outside of the box and linking multiple GPU nodes together into a cluster. Gary Scantlen of Creative Consultants discovered that bandwidth and latencies can become a problem with GPU clusters; these factors can inhibit computational efficiency.

One challenge is that the GPU can be like an island; you need an efficient way to get data in and out. Gary started writing benchmarks, and experimenting with various configurations to see where the bottlenecks were. In his quest to move as much data as quickly as possible across the network and keep up with the GPUs, Gary used QDR InfiniBand adapters from Mellanox Technologies and remote direct memory access (RDMA) to optimize the distribution of data between GPUs in different nodes.

"We experimented with other interconnect technologies, but there was just no comparison with what we could achieve using InfiniBand," said Greg Scantlen of Creative Consultants. "The combination trumped everything else in terms of more bandwidth and reduced latency benefits."

### The InfiniBand Solution

Creative Consultants provided a breakthrough by developing a configuration called Stella. The name is short for Constellation, which refers to a Beowulf cluster that provides the performance advantages of vector supercomputers on low cost clusters powered by GPUs.

Stella is a 9-GPU, 3-node cluster with switchless QDR InfiniBand. It features GPU-based systems boxes with multiple subnet managers. Each node includes a dual-port InfiniBand adapter, where each of the ports is connected to the other two nodes, to form a ring topology between multiple GPUs.

The compute is done through parallel GPU communications over native InfiniBand. The CPUs manage the GPUs and stay out of the computational path. InfiniBand offers efficient transport and offloading; it lets GPUs communicate directly via RDMA without CPU intervention; this saves time in the data path between GPUs. It takes the overhead away so that QDR link can run more iterations in less time.

Stella proved that inter-node QDR InfiniBand connections can be equivalent in data transfer speed to on-board PCIe (i.e., intra-Node PCIe).

### Putting the System to the Test

The team at Creative Consultants put the system to the test in the following projects.

They modified NBODY (software which is distributed as part of CUDA SDK that simulates a dynamic system of particles) to run on Stella; this produced 4.5 single precision

teraflops of compute power.

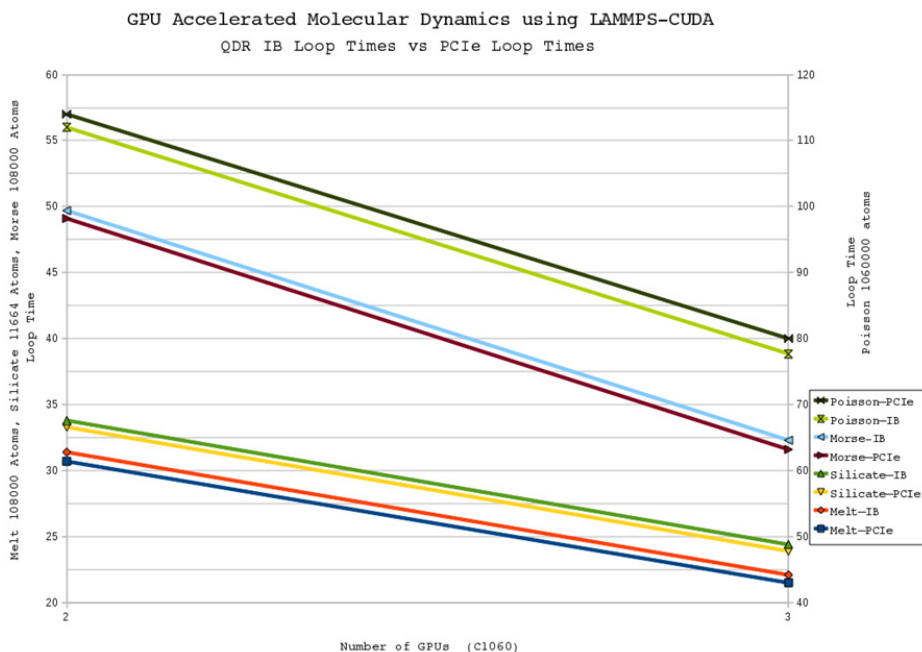
A second project involving LAMPPS (an open source molecular dynamics environment from Sandia National Laboratories) was done using eStella, which is an online version of Stella.

Creative Consultants worked with Paul Crozier and Christian Trott, both members of the LAMPPS-CUDA development team at Sandia National Lab. EStella, an agile online system, gave them a way to test their latest work.

They ran different tests using various combinations of GPUs interconnected by GPUs and InfiniBand. Creative Consultants did an efficiency test to see how well LAMMPS code was parallelizing (see graph and chart below).

They determined that one GPU per node in three nodes performed as well as three GPUs in one node: a testament to InfiniBand's ability to efficiently and cost-effectively connect GPU nodes for parallel HPC applications.

INPUT File	Number of Atoms	LOOP TIMES				
		G2N1	G2N2	G3N1	G3N3	
in.flow.poislarge	10600000		114.0	112.0	80.0	77.7
in.melt	108000		30.7	31.4	21.5	22.1
in.silicate	11664		33.3	33.8	23.9	24.4
in.morse	108000		49.1	49.7	31.6	32.3



The results show that InfiniBand provide same or lower latency compared to internal PCIe latency, which means that GPUs connected on separate servers see the same latency (or even lower) between them as if they were located in the same server.

### A Boon for Researchers

Low cost GPU clusters can have dramatic implications for how groups of researchers get work done.

Much research work is currently being accomplished with single box systems. As more and more work goes to systems built from GPUs, you could have one supercomputer for small groups that can produce five teraflops of performance

The systems can give researchers more control over their work, by democratizing supercomputing power and reducing the need to go to a large facility to run applications. Systems like Stella can offer an economical solution that makes it practical to take the computing power in house and do the work locally.

“It is about getting the same work done in less time, and increasing researcher productivity,” said Gary Scantlen.