GPI – SUPPORT FOR GPUs

- GPI: Programming Interface to create a partitioned global address space
- Global Memory Areas can be located in host memory or in GPU memory
- Direct access to remote GPU memory
- Uses NVIDIA GPU-Direct technology for direct data transfer between two GPUs
- No copies to host memory are required – direct read/write of GPU remote memory
- Minimal overhead for transferring data between GPUs
- Low latency for small data transfers
- Weak (point to point synchronization) is allowed by using read with remote notification

Micro-Benchmarks

Himeno

- Stencil Code compared against optimized communication hiding MPI version using two Kepler K10
- Decomposition in slowest direction
- Better performance and better scalability with GPI