Date: 03/02/2020

71-75 Shelton Street, London, WC2H 9JQ.

Powerful Distributed Computer Systems, over 10 – 40Gb/s Interconnect (Ver. 1.0)

SMART Cities 'Big Data' applications demand the highest levels of processing power over a wide area network, connected by the lowest latency, peak throughput, and maximum availability interconnect. High performance personal computer hardware is relatively cheap, however with the latest interfaces such as Thunderbolt 4 and USB 4 interface technology throughput is outstripping main-stream computer hardware performance. To illustrate this, If we take for example the Thunderbolt 3 interface maximum theoretical throughput of 40Gb/s (5GB/s), then this would require the next generation PCIe 4.0 x4 NVMe solid state storage which tops out at around 4.95GB/s (as PCIe 3.0 x4 NVMe solid state storage speed is typically 3.4GB/s). and even then this would not saturate the physical link bandwidth. We would need to jump another generation PCIe 5.0 x4 at 16GB/s to saturate Thunderbolt 3 I/O and meet the expected future demands of Thunderbolt 4. These latest physical layer interfaces place extreme demands on the interconnect, essentially eliminating copper as a possible solution other than for the very shortest of links – which in turn dramatically reduces the use cases for the technology. In order for powerful distributed computer system clusters to not loose a significant percentage of the matrix-network performance fibre optic links need to be employed as the interconnect. However optical fibre interconnect is still fairly scarce and expensive, especially for Thunderbolt 3. Corning Inc. are the main manufacture of Thunderbolt 3 optical fibre interconnect, however the physical layer connector used is mini display port to support thunderbolt 1 and 2, not USB-C (used for Thunderbolt 3), they are for self-powered peripherals only i.e. no copper power lines are included with the optical fibre, there have been reports of early failure of these optical fibre interconnects which are possibly due to the heat generated at the peripheral or computer that it is connected to, and these interconnects are not rated for hazardous environments requiring fail-safe performance. However for this brief we can demonstrate practically what is possible from these powerful distributed computer clusters using Thunderbolt 3 copper and optical fibre interconnect and accept some sub-optimal solution performance loss. 10 to 40Gb/s Ethernet over optical fibre using routers DNS and QOS settings, is the obvious choice for real-world mission critical applications.

Distributed computing involves splitting up a task across several computers in a local cluster, parallelising the tasks with each machine being scheduled these tasks from a master. These computer clusters work best if the machines that are to be liked together are relatively close in specification. Cluster computing is dependant on each machine

having access to the same data. Each node in the cluster will operate independently of the rest, so that if one machine slows down or crashes, the other computers in the cluster can compensate for this whilst still making use of the data supplied by the slowing or crashed node. The faulty node can then be rebooted and execute a batch file automatically to effectively “self-heal’ the network cluster.

For storing the project data large very fast drives are required to supply the maximum bandwidth of the physical layer Interface. For this brief all nodes use PCIe 3.0 x4 NVMe SSDs with serial read speeds of: 3.4GB/s and serial write speeds of 2GB/s(Corsair Force MP510), Hynix HFS-GD9’s with a serial read speeds of 3.4GB/s and serial write speeds of 2.2GB/s and Samsung SSDs with 2.8GB/s serial read along with a 1.6GB/s serial write speeds. The performance of these PCIe 3.0 x 4 NVMe SSDs was verified using Crystal Disk Mark 7, to allow us to determine the level of performance loss over the theoretical 10Gb/s(1.25GB/s) full-duplex Thunderbolt 3 interconnect.

In order to enable communication between different hosts in the Thunderbolt chain, network properties should be configured as Bridging Mode or Routing Mode. Bridging mode is easier to set up however the performance when using multiple computers will decrease with each computer added. Routing mode adds a higher level of complexity in configuration while generally achieving a higher point-to-point throughput. All computers on the Thunderbolt chain need Thunderbolt software installed, regardless of their position in the chain. Thunderbolt networking should be first authorised in the Thunderbolt software before setting up the network configuration. In Bridging mode Dual Thunderbolt port computers in the middle of the Thunderbolt chain are configured as bridges, which share the Same subnet to allow communication between them. Whereas in Routing Mode one or more computers in the Thunderbolt chain are configured as routers, routing traffic on different subnets of the thunderbolt network.

Once the network is set up by connecting all of the computers together via their Thunderbolt 3 USB-C ports using either active copper interconnect (incorporating a retiming integrated circuit), or a optical fibre connection, we used Crystal Disk Mark 7 again this time to test each computers solid state drives over the Thunderbolt 3 connection, to ascertain the effect of the Thunderbolt 3 10Gb/s (1.25GB/s) physical layer interface in peer-to-peer mode (result shown in fig. 1a). Crystal Disk Mark 7 proves to be a very useful tool when used over Thunderbolt 3 interconnect to determine the performance loss due to the Thunderbolt 3 physical layer interface (USB-C), in peer-to-peer mode.

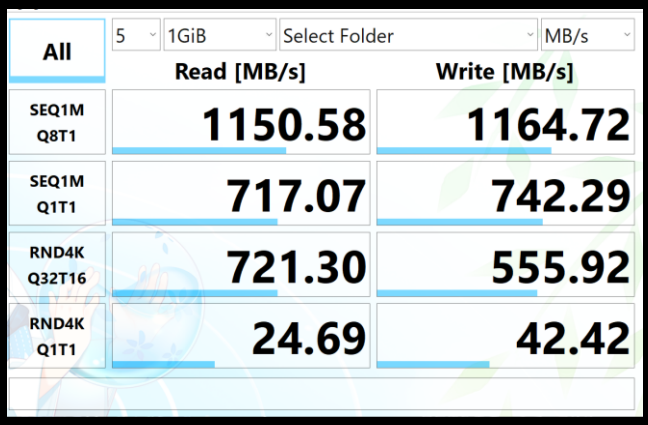

Figure 1a) Corsair Force MP510 to Hynix HFS-GD9 SSD (1150MB/s = 9.2Gb/s), via Thunderbolt 3.

The results in figure 1a) were also confirmed by copying a 20GB video file from one computer to another over the Thunderbolt 3 interface, which completed in about 18 seconds.

The next stage is to install a queue manager on the master computer, which is determined by the cluster application. For simplicity Blender (graphical rendering program) was installed on all computers and Dr Queue (render farm management) Desktop version (not Server version) was installed on the master, as the configuration is simpler on the Desktop version.

Blender is particularly mathematically intensive and is designed to work with more than one instance running on different computers, that could also be for example a 'big data' application such as analysing commuters travel patterns, types of transport, travel times, and changes in behaviour due to incentives and penalties. By opening an instance of Blender on each machine reading from the same '.blend' animation file, then switching to the output page and ensuring every computer is writing its output to the same directory. 'Touch' and 'No Overwrite' options were then selected, so when the rendering is started, each Blender instance will grab the next available frame that hasn't been created by another node. This is an effective and quick way of spreading the load amongst the cluster, with only a small master to client overhead.

One of the best tools for render farm queue management is 'Distriblend', that automatically distributes Blender jobs between the various nodes of the cluster. Using 'Distiblend' the benchmark results obtained indicated almost a quadruple increase in processing speed across four computers, with the master computer remaining unburdened essentially acting as a supervisory control and data acquisition (SCADA) unit. If the master computer was then combined with appropriate SCADA software, it could for example run a distributed computer environmental monitoring and control system.