# SHOOTING THE GAP BETWEEN CLOUDS AND ON PREMISES FOR HPC

Written by Timothy Prickett Morgan.

VERNE GLOBAL
THENEXTPLATFORM

**If only high-performance computing (HPC) were as simple as buying the fastest processors, the densest memory, and the zippiest interconnects, slapping it all together and watching the applications hum. But HPC takes a lot more than this, and designing an HPC system takes a lot of skill and effort, as it does to make the applications that ride atop these clustered systems take full advantage of the hardware at their disposal.**

HPC is also not as simple as taking the vanity free, minimalist system designs of the hyperscalers and cloud builders and lashing them together with some fast networks, maybe adding in some accelerators like GPUs and FPGAs as well as sophisticated middleware to manage traditional simulation and modelling workloads in a style familiar to the HPC centres of the world. This will work, up to a point, but it is not optimal along many different vectors.

Getting optimal is, in a nutshell, why Verne Global took over a former NATO airbase and an Allied strategic forces command centre outside of Keflavik, Iceland, back in 2012 and converted it into a super-secure datacentre, and why the company took the wraps off its hpcDIRECT service back in late 2017. The idea, as encapsulated behind its TrueHPC blueprints and methodology, essentially boils down to providing a cloud experience for HPC applications without any of the drawbacks that come with running HPC on the public cloud.

Those drawbacks? They include unpredictability of performance, unpredictable capacity at a certain configuration, and cost. While the public clouds such as Amazon Web Services, Microsoft Azure and Google Cloud Platform offer what appears to be infinite capacity, that capacity is certainly not infinite and is actually not inexpensive compared to running HPC applications atop premises clusters or on bare metal clusters like the hpcDIRECT service from Verne Global. If you do that math – and that is what HPC centres do for a living – then hpcDIRECT can provide a premium service, but not at the premium cost of similar capacity on these public clouds. These big public clouds do have some advantages, such as being able to roll out capacity on a global basis - and to do so fairly quickly. But these turn out to be of little advantage to HPC centres, which tend to be more methodical and take their time architecting systems and which are more interested in security and repeatability of results than they are quick consumption of compute.

“We believe that with true HPC, you need to be able to have an HPC cluster that you are able to deploy in a repeatable fashion and in a way that is documented and automated as it changes over time,” explains Tate Cantrell, chief technology officer at Verne Global. “This is something relatively new to HPC, but it is common at the hyperscale companies out on the Internet. Traditionally, HPC clusters have been acquired for between $10m and $100m dollars, and you put them in one place, you put a job scheduler on them, and once they are optimised you don’t really change them. We believe that the advent of HPC in the cloud is enabling real competition and the opportunity to get an environment that is most optimised for the applications. But sometimes that is going to mean moving a cluster, or upgrading it, so mobility is as important as the performance or the Total Cost of Ownership of the initial cluster that gets installed. The customers understand the requirements of the application, but we are the HPC specialists that can help them find that optimised environment and keep it optimised as conditions and applications change.”

As far as Cantrell is concerned, the jury is still out as to whether Ethernet will find its place at the upper echelons of HPC, where Verne Global is positioning hpcDIRECT, and that means picking low-latency InfiniBand to cluster machines together.
The company has worked with the OpenStack community – in particular on the OpenStack Ironic bare metal extensions to the open-source cloud controller – to make it work well with InfiniBand. The hpcDIRECT service uses the OpenStack Keystone service for authentication, and using Ceph (owned by Red Hat) as its block and object storage, with the focus being on low cost. Customers can build storage clusters on the bare metal servers if they need parallel file systems commonly used in HPC environments, such as the open-source Lustre or BeeGFS file systems or IBM's Spectrum Scale (formerly GPFS). If customers have a preference for a particular supplier of HPC storage appliances – DataDirect Networks is the popular one among current hpcDIRECT customers who take this approach, according to Cantrell – then Verne Global is happy to stand up a private storage cluster in its own datacentre for customers to use.

In terms of compute, the vast majority of hpcDIRECT customers want X86 servers, and for the most part that means Intel Xeon processors – although there is increasing interest in AMD and especially the future release of AMD's Rome. And Verne Global also has testbed machines which allow customers to run benchmark tests on ARM-based Marvell ThunderX2 processors. In some cases, the server nodes that are aimed at production workloads contain Nvidia Tesla GPU accelerators. The table below shows the standard feeds and speeds of the current hpcDIRECT systems:

| OPTION | PROCESSOR | PHYSICAL CORES | RAM | PRIMARY STORAGE | SECONDARY STORAGE | ETHERNET | INTER-CONNECT | GPU |
|---|---|---|---|---|---|---|---|---|
| gpc.s01.arm | ARMv8 TX2 (cn9960-2000) | 16 Cores @ 2.0GHz | 64 GB | 2x240GB SSD | – | 2x10 Gbit | – | – |
| gpc.m01.arm | ARMv8 TX2 (cn9975-2400) | 28 Cores @ 2.4GHz | 128 GB | 2x240GB SSD | – | 2x10 Gbit | – | – |
| gpc.b01.arm | ARMv8 TX2 (cn9980-2500) | 32 Cores @ 2.5GHz | 192 GB | 2x240GB SSD | – | 2x10 Gbit | – | – |
| | | | | | | | | |
| gpc.t01.x86 | Single Intel Xeon Skylake Silver 4110 | 8 Cores @ 2.1GHz | 32 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.s01.x86 | Dual Intel Xeon Skylake Silver 4110 | 16 Cores @ 2.1GHz | 64 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.m01.x86 | Dual Intel Xeon Skylake Gold 5115 | 20 Cores @ 2.4GHz | 96 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.b01.x86 | Dual Intel Xeon Skylake Gold 6132 | 28 Cores @ 2.6GHz | 128 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.x01.x86 | Dual Intel Xeon Skylake Gold 6154 | 36 Cores @ 3.0GHz | 192 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpu.s01.x86 | Dual Intel Xeon Skylake Silver 4110 | 16 Cores @ 2.1GHz | 192 GB | 2x240GB SSD | 2x480GB SSD | 2x10Gbps | – | 2xT4s |
| gpu.m01.x86 | Dual Intel Xeon Skylake Gold 5115 | 20 Cores @ 2.4GHz | 256 GB | 2x240GB SSD | 2x480GB SSD | 2x10Gbps | – | 4xT4s |
| gpu.b01.x86 | Dual Intel Xeon Skylake Gold 6132 | 28 Cores @ 2.6GHz | 384 GB | 2x240GB SSD | 1x1.6TB NVMe | 2x10Gbps | – | 4xV100s |
| gpu.b02.x86 | Dual Intel Xeon Skylake Gold 6132 | 28 Cores @ 2.6GHz | 384 GB | 2x240GB SSD | 1x1.6TB NVMe | 2x10Gbps | IB EDR | 4xV100s |
| gpu.x01.x86 | Dual Intel Xeon Skylake Gold 6154 | 36 Cores @ 3.0GHz | 768 GB | 2x240GB SSD | 1x3.2TB NVMe | 2x10Gbps | 2 x IB EDR | 8xV100s |
| | | | | | | | | |
| gpc.t02.x86 | Single Intel Xeon Cascade Silver 4208 | 8 Cores @ 2.1GHz | 32 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.s02.x86 | Dual Intel Xeon Cascade Silver 4210 | 20 Cores @ 2.2GHz | 64 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.m02.x86 | Dual Intel Xeon Cascade Gold 5218 | 32 Cores @ 2.3GHz | 128 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.b02.x86 | Dual Intel Xeon Cascade Gold 6254 | 36 Cores @ 3.1GHz | 192 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpc.x02.x86 | Dual Intel Xeon Cascade Platinum 8260 | 48 Cores @ 2.6GHz | 384 GB | 2x240GB SSD | – | 2x10Gbps | – | – |
| gpu.s02.x86 | Dual Intel Xeon Cascade Silver 4210 | 20 Cores @ 2.2GHz | 192 GB | 2x240GB SSD | – | 2x10Gbps | – | 2xT4s |
| gpu.m02.x86 | Dual Intel Xeon Cascade Gold 5218 | 32 Cores @ 2.3GHz | 256 GB | 2x240GB SSD | – | 2x10Gbps | – | 4xT4s |
| gpu.b02.x86 | Dual Intel Xeon Cascade Gold 5218 | 32 Cores @ 2.3GHz | 384 GB | 2x240GB SSD | 1x1.6TB NVMe | 2x10Gbps | – | 4xV100s |
| gpu.b03.x86 | Dual Intel Xeon Cascade Gold 5218 | 32 Cores @ 2.3GHz | 384 GB | 2x240GB SSD | 1x1.6TB NVMe | 2x10Gbps | IB EDR | 4xV100s |
| gpu.x02.x86 | Dual Intel Xeon Cascade Gold 6254 | 36 Cores @ 3.1GHz | 768 GB | 2x240GB SSD | 1x3.2TB NVMe | 2x10Gbps | 2 x IB EDR | 8xV100s |

The InfiniBand network is used to link compute and storage nodes together, while the Ethernet network is used to interface with the outside world or, in the rare cases where the HPC applications are not latency-sensitive, to link nodes together. One example is the Monte Carlo simulations that underpin the risk analysis systems at financial institutions, where heavy compute is needed but the backplane network does not have to be high bandwidth or low latency to the same degree as traditional HPC technical applications in simulation and modelling.

The hardware is the easy part. Creating the TrueHPC blueprint for a running cluster that provides the mechanism to be moved from an on-premises location to Verne Global datacentre – that is the secret sauce. The blueprint makes the job of moving possible and preferable to running HPC applications on the big cloud providers.

"When we created the TrueHPC blueprints, we wanted to make them human readable," says Cantrell. "In 80 lines of code, we can describe a working HPC cluster. It's not just about porting the application any more, it is porting the cluster. We come in and provide HPC specialist support from the beginning, get the requirements, put them into a human readable blueprint, put in their applications, and make it so that they have this repeatable cluster deployment that they can move around our infrastructure without having to hire infrastructure experts. Our tools and our experts handle that for them."

hpcDIRECT blueprints are human-readable text files in YAML format. The blueprints provide all of the details required for bringing an HPC cluster to life, including compute, networking and storage. Once the cluster is deployed, software packages and libraries are brought to life automatically with hpcDIRECT elements that are built using Ansible technology. The end user can roll their own elements, or pick from a variety of solutions that are pre-built by the Verne Global dev-ops team. It also calls up the systems software – perhaps Linux and a Kubernetes or Singularity container environment – and the specialized libraries that the HPC application might need and loads them on the instances in the hpcDIRECT cloud. But perhaps the most important thing is that Verne Global has a team of HPC experts that can help customers set all of this up as well as help them configure the right systems for the applications. And once it's done, the cluster is portable to the standard infrastructure options Verne Global provides, or even to bespoke machines should a customer need something that is not standard. Such hand-holding and the option of custom configurations is definitely not available on the big public clouds. In fact, such customisation is against the very principles of hyperscalers, who must have a high degree of homogeneity to make it possible to operate at scale and extract profits from their infrastructure.

**Ultimately, it is about the bang for the buck**

While the TrueHPC approach with the hpcDIRECT service is a key differentiator for Verne Global as it competes against the big public clouds and some niche HPC cloud players, HPC first and foremost is about getting the most compute for the least cost. This is table stakes.

To help demonstrate the performance of the hpcDIRECT cloud, Verne Global commissioned HPCNow! to run independent benchmarks on similar configurations on the hpcDIRECT and Amazon Web Services. Microsoft's Azure cloud was also tested, but its InfiniBand-based instances used older iron and software and were therefore not representative. To stress test these clouds, HPCNow! took the OpenFOAM computational fluid dynamics tool and the motorcycle airflow simulation that is part of the tool. The number of elements in this base test is 200,000, and this was increased to 41.6 million elements to scale up to a reasonable, medium-sized HPC workload. The set-up used OpenFOAM v1812, with the code compiled using the GCC 7.3.0-2.30 compilers and set up with the OpenMP 3.11, FFTW 3.38, OpenBLAS 0.31, ScaLAPACK 2.0.2, METIS 5.1.0, SCOTCH 6.0.6, and HDF5 1.10.2 libraries. The files were stored in NFS, just to keep it simple, but a BeeGFS file system with RDMA capability could have been used to remove some storage overhead.

Five runs were attempted for each of the HPC cloud configurations, and wall time to complete the simulation of airflow around the motorcycle was the performance metric.

On the Verne Global hpcDIRECT set-up, 20 machines were networked using 40 Gb/sec FDR InfiniBand; these machines had a pair of "Skylake" Xeon SP-6130 Gold processors, which have 18 cores running at 2.1 GHz each, plus 384 GB of main memory and a pair of 960 GB flash SSDs. If you look at the table below, you can see what performance of the five runs looked like across the four different cluster sizes:

| RUN | RUNTIME 64 CORES | RUNTIME 128 CORES | RUNTIME 256 CORES | RUNTIME 512 CORES |
|---|---|---|---|---|
| 1 | 535 | 282 | 152 | 89 |
| 2 | 537 | 281 | 151 | 89 |
| 3 | 538 | 280 | 151 | 89 |
| 4 | 543 | 280 | 152 | 89 |
| 5 | 541 | 280 | 151 | 89 |
| MIN | 535 | 280 | 151 | 89 |
| MAX | 543 | 282 | 152 | 89 |

On the Amazon Web Services cloud, the system was configured with the c5n.9xlarge instance on the EC2 service, which has two custom Skylake Xeon SP-8000 Platinum series processors and exposes 36 virtual cores, each running at a much higher 3 GHz clock speed. These instances are configured with 96 GB of main memory, they used Elastic Block Service with 7 Gb/sec of bandwidth to host the NFS file system with a pair of standard Nitro 25 Gb/sec links employed for the fabric interconnect - something AWS provides with a little more bandwidth than the Verne Global hpcDIRECT systems but minus the RDMA.

| RUN | RUNTIME 64 CORES | RUNTIME 128 CORES | RUNTIME 256 CORES | RUNTIME 512 CORES |
|---|---|---|---|---|
| 1 | 702 | 324 | 180 | 130 |
| 2 | 694 | 324 | 178 | 133 |
| 3 | 706 | 341 | 181 | 125 |
| 4 | 702 | 342 | 180 | 119 |
| 5 | 704 | 322 | 179 | 121 |
| MIN | 694 | 322 | 178 | 119 |
| MAX | 706 | 342 | 181 | 133 |

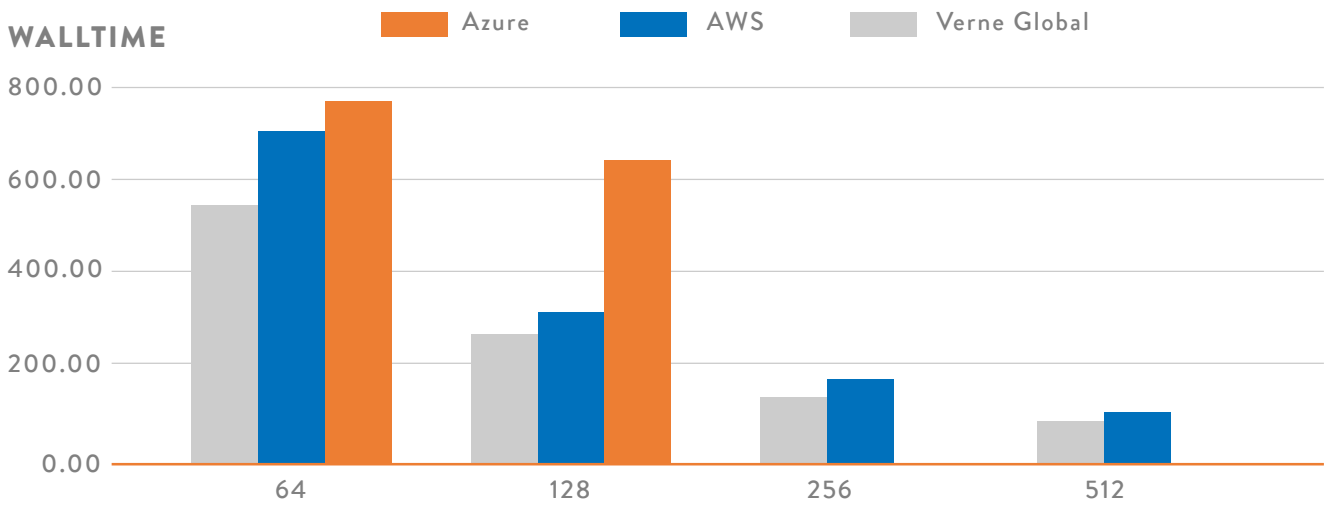
As you can see above, the reproducibility of results was pretty good on AWS when a small number of cores was used, but as the core counts grew, so did the variability in the runtime. This, no doubt, is due to network effects and probably to noisy neighbour issues that don't happen on bare metal clouds like hpcDIRECT.

There were some issues scaling up the OpenFOAM simulation on the Microsoft Azure cloud, as you can see in the table below:

| RUN | RUNTIME 64 CORES | RUNTIME 128 CORES | RUNTIME 256 CORES | RUNTIME 512 CORES |
|---|---|---|---|---|
| 1 | 762 | 1129 | 1892 | – |
| 2 | 754 | 1115 | 1572 | – |
| 3 | 796 | 1092 | – | – |
| 4 | 804 | 633 | – | – |
| 5 | 868 | 1123 | – | – |
| MIN | 754 | 633 | 1572 | – |
| MAX | 868 | 1129 | 1892 | – |

The Azure H16mr instances date from the end of 2014, and are the most recent ones that offer InfiniBand support. These instances use "Haswell" Xeon E5 v3 processors, in this case with 16 virtual cores, 224 GB of memory, 2 TB of disk, and 40 Gb/sec FDR InfiniBand interconnects. There were all kinds of issues in running OpenFOAM on this older instance. We provide this information for completeness, but again, do not think it is representative of the performance that might be on the Azure cloud with a modern InfiniBand set-up.

If you take the best runs for each cloud and stack them up against each other, hpcDIRECT has a significant advantage compared to AWS running the OpenFOAM Test, below:




OpenFOAM is particularly sensitive to network latency and not as much to clock frequency, which is one of the reasons why the hpcDIRECT set-up with slower machines is beating the AWS systems with faster ones. What is also important to note is that the scalability on the bare metal hpcDIRECT instances as the core counts rise is much better on this application compared to the scalability on the AWS cloud. The hpcDIRECT set-up scales as you might expect, with 128 cores doing 1.9X that of 64 cores, and 256 cores doing 3.5X and 512 cores doing 6X. The AWS set-up, using a 64 virtual core cmp.m04.x86 instance on the hpcDIRECT cloud as a reference, delivers 0.77X, 128 cores delivers 1.7X, 256 cores yields 3X and 512 cores yields 4.5X.

Now here is the kicker. Depending on the configuration, instance types, and method of reservation chosen, on a like-for-like term, hpcDIRECT will cost half for a given amount of capacity compared to AWS for a six-month contract, which is typical among Verne Global's HPC customers.

If it performs better and it costs less, that sure makes the decision a lot easier. But there is more to it still.

"We have approached this differently from the public clouds," says Cantrell. "If customers are concerned about how they achieve optimal application performance, they need to be concerned about whether the next change that happens at the big public clouds is going to affect the performance of their applications."

Growth of public cloud is certain. What's more uncertain is the performance of the calibre required by HPC. One thing, however, is sure: if hpcDIRECT performs better and costs less than public cloud, then that sure makes the decision of eschewing the big, hyperscale providers for Verne Global a lot easier.

## ABOUT THE NEXT PLATFORM

*The Next Platform* delivers in-depth coverage of high-end computing at large enterprises, supercomputing centers, hyperscale data centers and public clouds. It goes behind the headlines to help readers understand the technologies being employed to solve some of our most pressing computational challenges, discover how they integrate and understand purchasing choices. *The Next Platform* is published by Stackhouse Publishing and Situation Publishing with coverage lead by co-editors Timothy Prickett Morgan and Nicole Hemsoth.