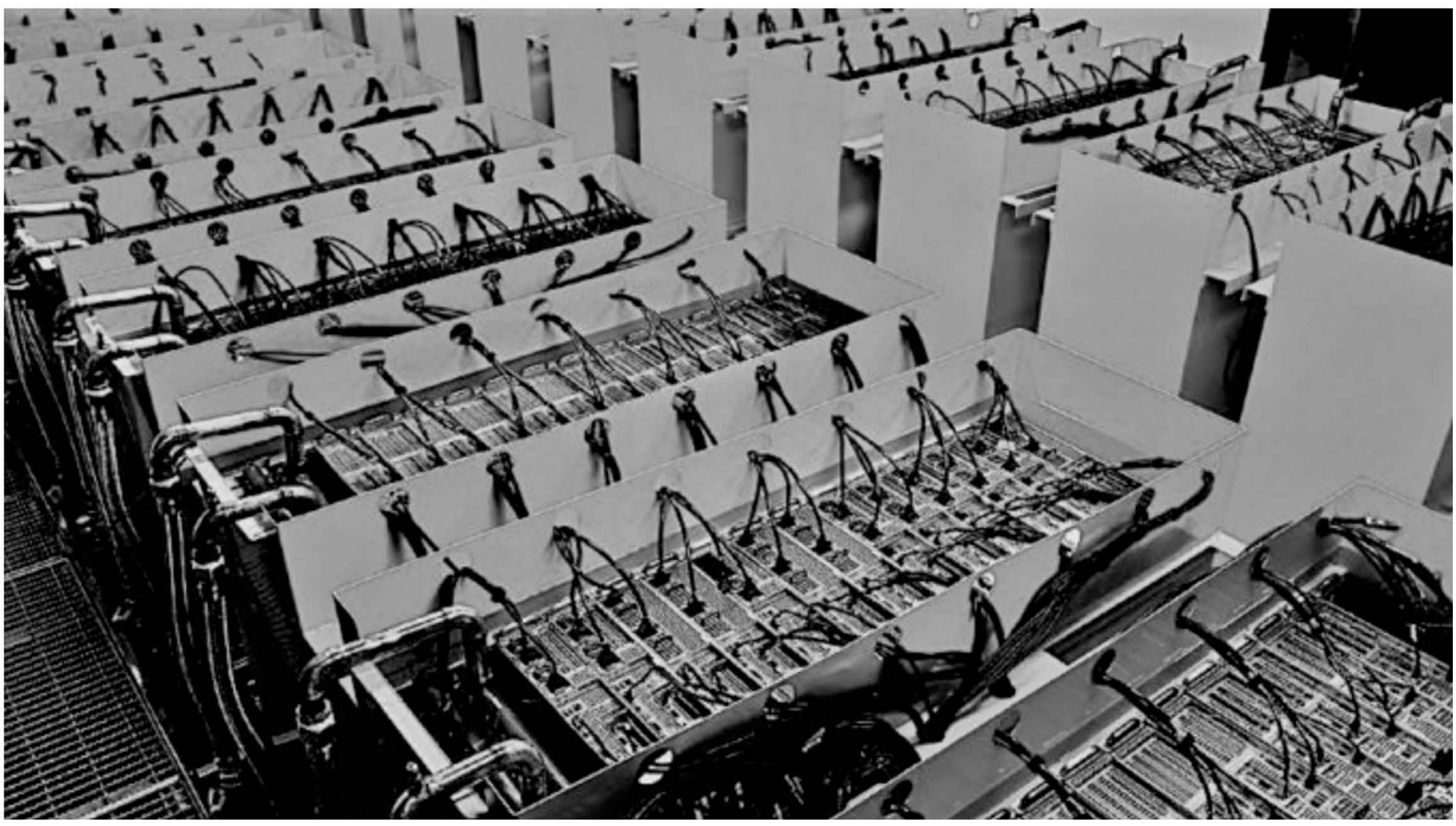
DUG SETS FOUNDATION FOR EXASCALE HPC UTILITY WITH XEON PHI

May 16, 2019 Timothy Prickett Morgan



While exascale systems, even at the single precision computational capability commonly used in the oil and gas industry, will cost on the order of $250 million, that cost pales in comparison to the capital outlay of drilling exploratory deep water wells, which can cost $100 million a pop. The trick is to be pretty sure where petroleum has been trapped and is waiting patiently to be extracted before the drilling starts. Software, and a scalable system that makes efficient use of it, can make the difference between striking black gold or drowning in a sea of red ink.

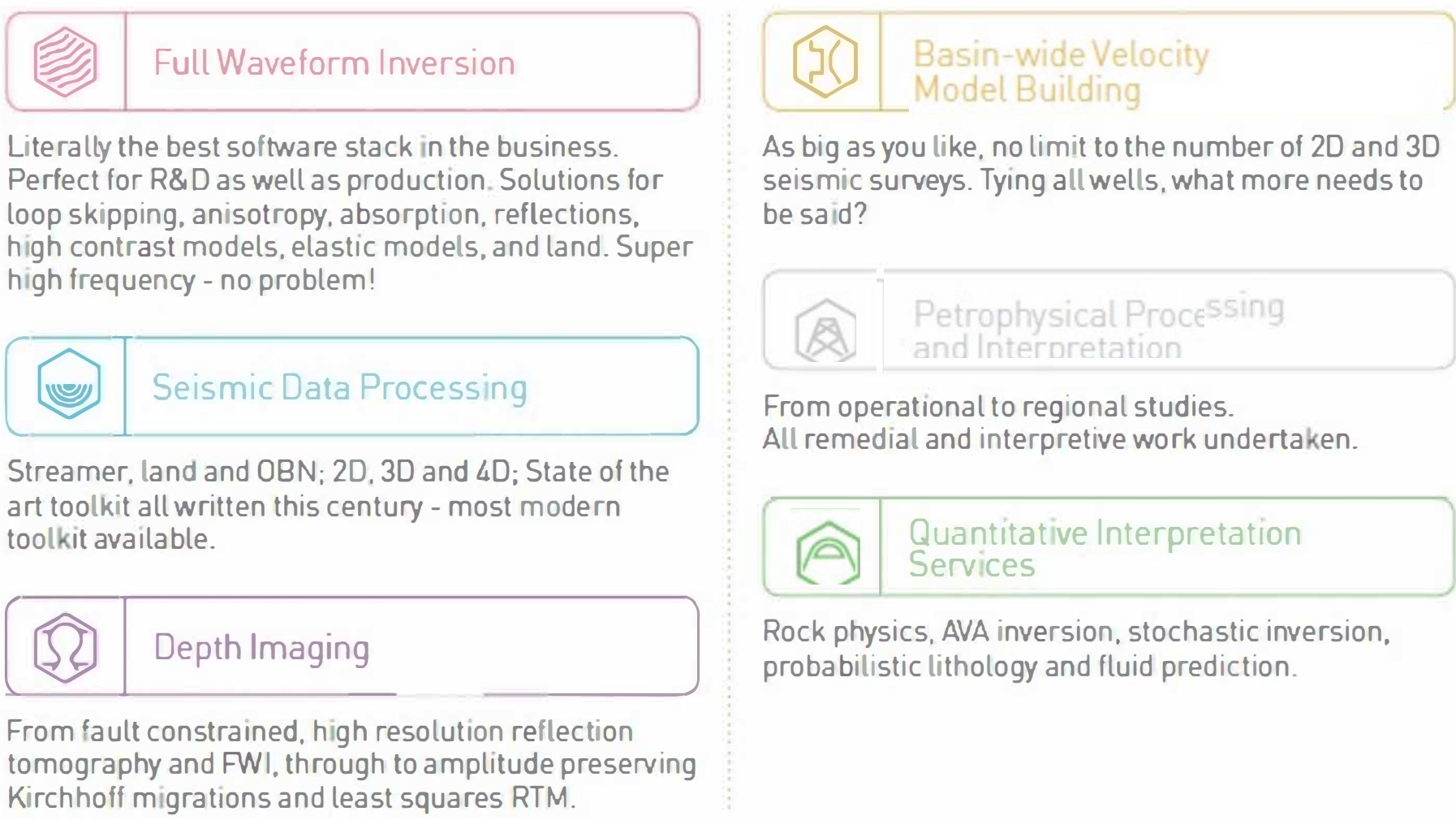
That is why DownUnder GeoSolutions, an upstart software maker and compute cloud provider specializing in the oil and gas industry but with aspirations to build a compute utility that can run all manner of HPC workloads, has invested so heavily in the past few years and is on the path to exascale computing like the famous national labs of the world are. The difference here is that DUG is a private company and is using its own money, fueled by the processing that it is doing on behalf of those scouring the world for oil and gas reserves.

We did a profile of DUG a year and a half ago, when the company was beginning its journey to exascale and had tapped Intel's "Knights" Xeon Phi family of co-processors and then processors as the main engines to run its software, which is called DUG Insight. That software has been in development for more than a decade, and it proves that the seismic processing and imaging work that lets oil and gas companies see miles below the Earth's surface where petroleum is lurking can run on massively parallel engines like the Xeon Phi. The Knights family of processors was quietly sunsetted last year, of course. So this may seem like an odd choice at first.

Phil Schwan, who was hired by DUG founders Matt Lamont and Troy Thompson back in 2008 to lead the DUG Insight software development team and who is now chief technology officer at the company, is unabashed about the use of Xeon Phi chips from Intel in the buildout of the "Bubba" supercomputer, which is being unveiled in a new datacenter called the Skybox Houston today and which is the backbone of the HPC compute utility service that the company is now calling DUG McCloud. It looks like DUG will not only be the largest commercial user of these processors, but the last user of any note as well.

"For my money and for my own internal workloads, which are still substantial, I love the Knights Landing in terms of bang for buck and overall return on investment," Schwan tells *The Next Platform*. "Right now I don't think it can be beat. But you know we examine that every six months or so. There will come a point when there are no more Xeon Phi processors to be had or when there is something else better. We have got a commitment from Intel to continue to be able to buy Xeon Phi chips, which have been earmarked for us. And so we are able to fill this data hall at the Skybox Houston datacenter with the Knights Landing variants."

Depending on the yields that Intel gets in its Xeon Phi manufacturing, which should be pretty good given that the chips are etched using mature 14 nanometer technologies, DUG will eventually fire up on the order of 40,000 Xeon Phi processors to finish building the first chunk of Bubba, which will weigh in at around 250 petaflops at single precision. Signal processing applications, of which seismic processing is but one example, usually depend on 32-bit single precision processing, which means DUG can cram a lot of work through a vector pipeline that is 512 bits wide, as the AVX-512 units on the Knights Landing chips are. At the moment, the DUG Insight software does not make use of 64-bit full precision or 16-bit half precision, so these features are basically useless to the company although could be important in the future as new workloads come onto the McCloud service.



That DUG Insight software stack is the secret sauce of the McCloud service, and has been rearchitected by DUG with the help of Intel, including a geophysical layer that was written symbolically and, significantly, using Python, which mixes and matches well with other codes and allows for users of the McCloud service to add their own modules and goodies to the software stack, which does reverse time migration (RTM), a key part of the imaging after seismic analysis is done, and full waveform inversion (FWI), a very computationally intensive application that takes all of the seismic waveform data generated basically by smacking the ground hard and listening to echoes. The DUG Insight stack includes modules for loading and managing data, seismic preprocessing, regularization, time and depth imaging, seismic inversion, rock physics, and high frequency FWI, all integrated and all ready to run in a cloudy fashion on the four supercomputers that DUG has located in Australia, Kuala Lampur, England, and the United States. (More on that hardware and the Skybox Houston facility in a moment.)

The thing to remember about the oil and gas industry, says Schwan, is that it really is still just a lot of very hairy batch processing. There are big workflows, data is run through them with many variations on the theme, it runs all night or all day, and people see the output and try to interpret the imagery to see where the oil and gas deposits might be hiding under the ground, often with ocean atop it to make it challenging to extract.

"It really is classic batch processing," Schwan explains. "But what has really lit the industry on fire is called Full Waveform Inversion, particularly now that computing is finally catching up to what the theory said was possible. We are building this exascale compute center, first and foremost, for FWI because it is a process that grows in compute with frequency to the factor of four. Historically, people have run 10 Hz or 12 Hz FWI, meaning that the output of that process is a model that has 10 Hz or 12 Hz frequency content in it. But with today's compute, we are able to run 50 Hz, 80 Hz, even 100 Hz models where you are getting out of the model basically everything that you recorded in that wavefield. And this takes 5,000X to 10,000X as much compute as running that 10 Hz or 12 Hz model. You need at exascale computer to do that. The neat bit is that with this a big FWI job, we don't need 40,000 nodes to all communicate with each other. We can run it in parts of a few dozen nodes that communicate closely with each other using MPI, and then the rest of it just scales out, with those parts are acting quite independently."

The McCloud service aimed at the oil and gas industry is more than just running the DUG Insight code on some hefty iron. DUG has to be better than the big public clouds if it wants to expand its HPC utility, and in some cases, that is not that difficult.

"If you're an oil and gas company today that wants to use a public cloud, you kind of have to build all the HPC infrastructure yourself," Schwan elaborates. "You have to get a network file system running if you want to have a sort of classic HPC file system. You need to set up your own job queuing system for your batch jobs, which as I explained oil and gas is still a batch processing centric. You've got to manage your own security because it's a public cloud. What we provide is HPC as a service that looks pretty much exactly like what they're used to having on premises. We give them a ready to go Lustre cluster file system, job queuing system, and security – all the access is by VPN, so the attack surface isn't open to the entire broad Internet. That's the real distinguishing characteristics of McCloud. And for the oil and gas industry specifically, we are providing all of our software and expertise, which we have used to provide oil and gas processing services last fifteen years. It is a software stack is all optimized for this hardware. And because we've been using it to deliver services to everybody in the oil and gas industry, they know that it is geophysically sound and will deliver results that they already trust because they are probably already one of our service customers."

And if not, one of their competitors probably is.

The interesting parallel here is that just like machine learning theory was very old but there wasn't enough compute and memory bandwidth – or large datasets – to prove that it worked, and worked well, until around 2011 or so, FWI is finally getting enough compute and memory bandwidth at a price that can be tolerated to allow it to transform the oil and gas industry.

The DUG computing facilities currently have an aggregate of 66 petaflops of single precision oomph, and the Houston facility in the United States will see the Bubba system grow by a factor of 24X compared to where it was last year, at 12 petaflops, to reach that target of 250 petaflops single precision. The Bubba system that is moving into the Skybox Houston datacenter currently weighs in at 31 petaflops, and is based on a mix of Xeon Phi 7210 and 7250 processors as well as some standard Xeon nodes. The Xeon Phi 7210 has 64 cores running at 1.3 GHz with 16 GB of MCDRAM high bandwidth memory, as you can see from **our detailed analysis of the Knights Landing chips when they debuted three years ago**, while the Xeon Phi 7250 has 68 cores running at 1.4 GHz for a little more performance. The former is rated at 5.32 teraflops SP, while the latter is rated at 6.1 teraflops SP.

The first data hall being used by DUG in the Skybox Houston facility has 22,000 square feet of space and 15 megawatts of power, and will be able to house systems that will use up all of those 40,000 Knights Landing processors plus some Xeon systems. The current configuration on day one has around 5,000 Xeon Phi processors (one per node) rated at about 30 petaflops SP and around 1,000 Xeon processors (two per node) rated at just under 1 petaflops SP.

At this point, DUG has been buying finished Xeon Phi and Xeon systems directly from Intel, including Intel processors, Intel motherboards, and Intel enclosures, plus 50 Gb/sec Ethernet network interfaces from Mellanox Technologies which are shared by multiple nodes in the enclosure and memory sticks that come from some vendor (our guess is Samsung, given the tight relationship Intel's server folks have with that memory maker in the server biz). The Intel system crams four Xeon Phi nodes into a 2U enclosure, which has a single 50 Gb/sec ConnectX-4 Ethernet NIC shared across those nodes, hyperscale style, and that allows burst mode from any node up to 30 Gb/sec if the other virtual slices on the NIC are not too busy. The leaf nodes in the Bubba system that link nodes to each other in the racks are based on 100 Gb/sec Spectrum ASICs from Mellanox, and so are the spine switches that lash racks and rows together are also based on the same chips. The neat thing is that with the multihost NIC approach, Bubba can have 200 nodes hanging off of a single leaf switch, which radically cuts back on networking costs. And given that the DUG Insight software is not a heavy user of MPI, the nodes do not need lower latency or higher bandwidth than the stock Ethernet is providing.

"We have designed our workloads to really be extremely parallel and almost the only communication that occurs is between the individual compute nodes and the Lustre storage servers," says Schwan. "That's where the big I/O communication goes. But it is even better than this. One of our workhorse applications is migration, and to do a standard, production, compute-intensive migration, an individual compute node might only consume 10 MB/sec, 20 MB/sec, or 30 MB/sec over the course of its runtime. That's not super unusual for our applications. So individual compute nodes really don't demand a lot and they can they can cache that I/O stream really effectively. They know what they're going to need next well in advance of needing it, so latency isn't really an issue either."

That sounds like magic, but it is just good software co-design with hardware, which is the path forward for all large scale computing, really.

One of the neat things about DUG is that it has not only adopted immersion cooling to drive up the efficiency of its datacenters, but it has also invented its own dieletric fluid, called DUG Cool, obviously, and its own immersion tanks for servers and storage, which are shown below:

DUG has a power usage effectiveness in its datacenters of around 1.03, which means just about all of the 15 megawatts of the Bubba facility will be used to actually do compute and storage. By way of comparison, the average datacenter in the early 2010s was around 1.9 PUE, a well run datacenter today is around 1.45 PUE, a so-called green datacenter is around 1.25 PUE. **Google is averaging around 1.11 PUE on a trailing twelve-month basis right now**, and the Facebook datacenter in Prineville, Oregon – the social network's first datacenter, **is weighing in at 1.08 PUE**, although its more modern datacenter in Lulea, Sweden **is at 1.16 PUE** (not all that great) and ditto for the Altoona, Iowa center, **which comes in at 1.18 PUE**.

Obviously, with this being the end of the line with the Knights family of processors, Schwan has a long-range plan involving any number of possible compute engines to reach the exascale heights at single precision for the McCloud service. But Schwan also makes a cogent observation about the Knights chips.

"My perspective on Knights Landing is actually quite different from what most people believe," says Schwan. "Instead of Intel killing off Xeon Phi because it didn't sell as well as it had hoped, it should have raised the flag and said mission accomplished, that the Xeon Phi was an incredible success. If you look at the Xeons that everyone is going to be buying in two years or three years, what do they have? They have lots of cores, 48 in some cases just this year. They've got a mesh network between the cores, and they have got AVX-512 vector units, and in some cases they are going to have high bandwidth memory. Well, *what's that?* It's a Xeon Phi."

DUG has enough space in that second computing hall to get to around 650 petaflops SP using processors and accelerators that are available today – so it can add an incremental 400 petaflops SP in that second room. To get to exaflops is not that much of a stretch, with another 350 petaflops that can be added when the Skybox Houston facility is expanded again. The plan is to get to 650 petaflops SP by the end of this year, probably using a mix of CPUs and GPUs, depending on customer needs, and then to reach 1 exaflops SP by 2021.

DUG is not betting so much on the cloud as it is being pulled into the cloud, and it is fulfilling a need, not building a field of dreams, according to Schwan. And hence the aggressive buildout for Bubba.

"If we look at all of the demand that people are expressing just from oil and gas, we sell that whole first room of Bubba this year," Schwan says. "It is astonishing how quickly the industry has turned from wanting to do it all on premises to absolutely embracing third party computer services – and with FWI the demand is tremendous. But until the room is actually there and they can walk up to it and touch the machines, you know it's hard to determine how quickly that that demand will turn into signed agreements. We could consume all of this compute with our own internal service business and FWI. I have no doubt, it's just a question of what the timeline is, but demand for those services is growing exponentially."

And so the compute to support it will have to as well, and Bubba will keep on embiggening.