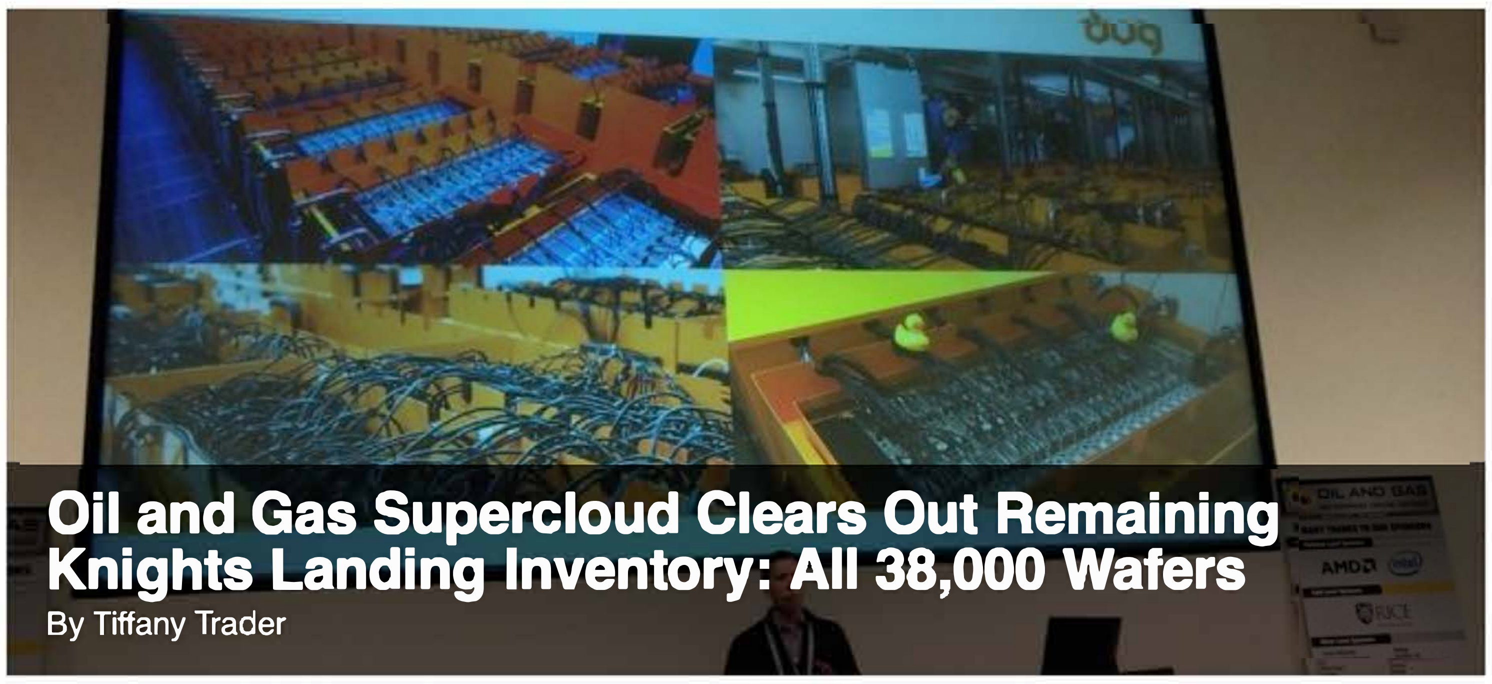
HPC wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them



Oil and Gas Supercloud Clears Out Remaining Knights Landing Inventory: All 38,000 Wafers

By Tiffany Trader

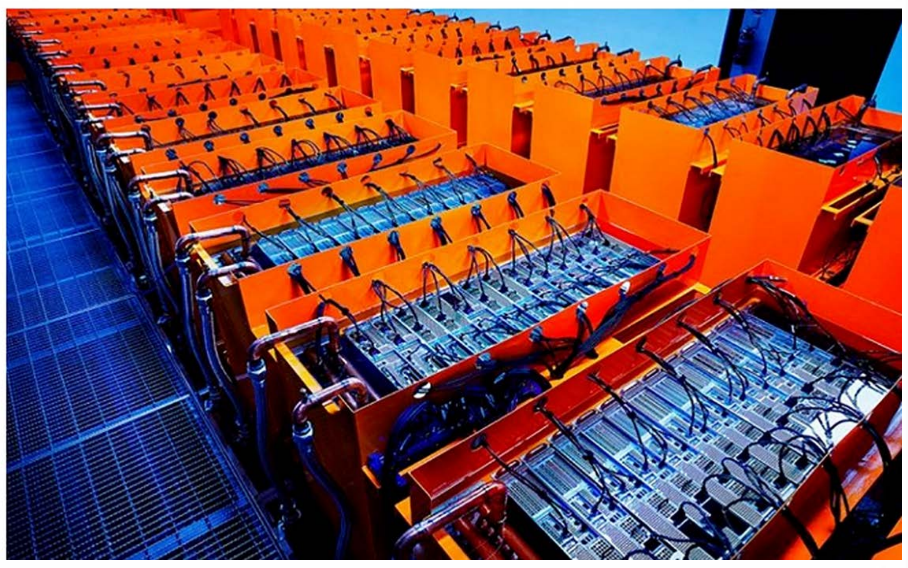
March 13, 2019

The McCloud HPC service being built by Australia's DownUnder GeoSolutions (DUG) outside Houston is set to become the largest oil and gas cloud in the world this year, providing 250 single-precision petaflops for DUG's geoscience services business and its expanding HPC-as-a-service client roster. Located on a 20-acre datacenter campus in Katy, Texas, the liquid-cooled infrastructure will comprise the largest installation of Intel Knights Landing (KNL) nodes in the world.

If you'd like to follow suit with your own KNL cluster and you don't have the hardware already, you're out of luck because not only has the product been discontinued (you knew this), but DUG has cleared out all the remaining inventory, snagging 38,000 wafers. We hear DUG similarly took care of Intel's leftover Knights Corner inventory back in 2014 (and those cards are still going strong processing DUG's workloads).

At the very well-attended Rice Oil & Gas conference in Houston last week, we spoke with Phil Schwan, CTO for DUG, who also delivered a presentation at the event. We chatted about DUG's success with Phi, their passion for immersion cooling, and some of the interesting decisions that went into the new facility, like the choice to run at 240 volts, as well as McCloud's custom network design.

DUG started off in oil services, in quantitative interpretation, before getting into processing and imaging, which has been the company's bread and butter for over a decade, but Schwan emphasized DUG is first and foremost an HPC company. "That's been our real focus in how we set ourselves apart – we have terrific geoscientists, but they are empowered to such a large degree by the hardware and the software," he shared.



"Bruce," DUG's Perth cluster, comprised of KNL nodes, totaling 20 single-precision petaflops. The "Bubba" tanks currently being installed at the Houston Skybox facility will look similar to these. Photo provided by DUG.

DUG currently enjoys somewhere in the neighborhood of 50 aggregate (single-precision) petaflops spread across the world (the company has processing centers in Perth, London, Kuala Lumpur, and Houston) but it is continually hitting its head on this ceiling. At the Skybox Datacenters campus, located in Katy, Texas, eight miles east of the company's U.S. headquarters in Houston, DUG will not only be adding to its internal resources for its geoscience services business, it will be priming the pump (significantly so) for its HPC as a Service business that it unveiled at SEG last year.

"Up until now it's been a purely service business – processing, imaging, FWI, and so on, but as soon as Skybox opens in early Q2, we'll have a lot more cycles to sell to third parties – and we have a few of those clients already beta testing the service both in Australia and here in the Americas."

Off The Wire

Industry Headlines

March 13, 2019

- University of Edinburgh to Host £79m ARCHER2 Supercomputer
- Excelero Awarded Another Patent for Heightened NVMe Storage Efficiency

March 12, 2019

- Global Fab Spending to See 2019 Decline, New Highs in 2020
- 4th IEEE International Conference on Rebooting Computing Issues Call for Papers
- Slovak Government Provides Capital Infusion of $17 Million to Tachyum
- Helping Keep Astronauts Safe with Advanced Simulations, Visualizations
- The Altair Partner Alliance Expands Offering to Include a Design-to-Cost Approach for Composite Part Design
- Mellanox Introduces NVMe SNAP Technology to Simplify Composable Storage

March 11, 2019

- Key Industry Players Converge to Advance Compute Express Link (CXL), a New High-Speed CPU Interconnect
- Purpose-Built Liquid Cooling Quick Disconnect Couplings Enter the HPC Market

To meet that demand, DUG has ordered the remaining Phi Knights Landing inventory from Intel, all 38,000 wafers. Once dies are cut and rolled into servers, the nodes will be combined with an infusion of KNLs transferred from DUG's other Houston site (the Houston capacity is collectively referred to as "Bubba") to provide around 40,000 total nodes with a peak output of about 250 (single precision) petaflops.

Schwan describes why DUG is so partial to the Phi products (the company is almost certainly Intel's largest customer of this line):

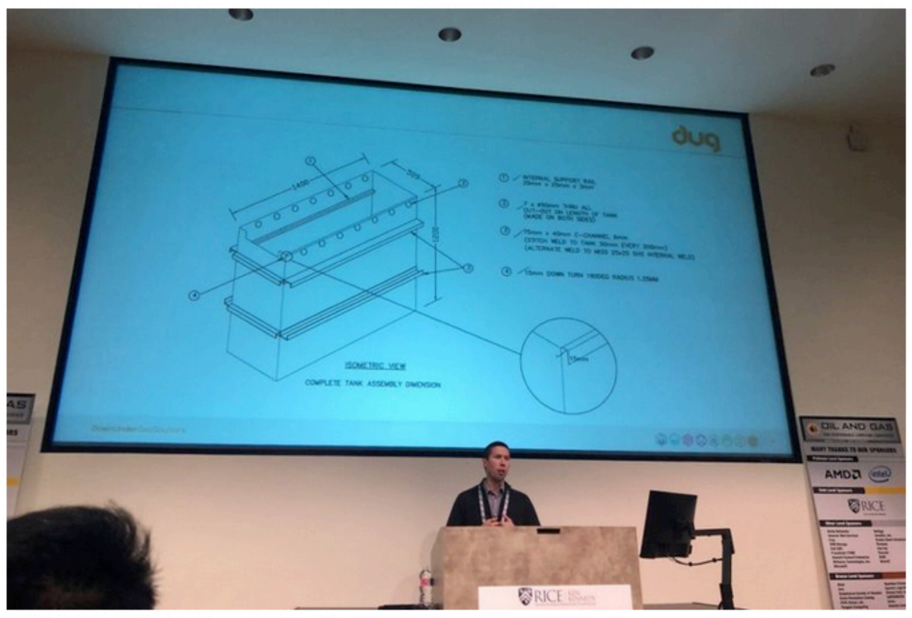
"There were a few reasons – number one, we came to the accelerator party fashionably late, and I think that worked well for us because if we had had to choose five years earlier, we would have chosen GPUs, and all of our codes would have gone in that direction and we'd be stuck there. Whereas our transition first to Knights Corner and then to Knights Landing – even if Intel did a bit of disservice by pretending that it's trivial and you just recompile and run – they are so much closer to the classic x86 architectures that we are already used to that we were able to take all of our existing skill sets, our existing toolchains and so on and make incremental improvements to make it run really well on the KNLs.

"The other thing is we run a bunch of things that are not necessarily hyper-optimized for the KNL – we run a lot of Java on the KNL and it runs great. And there's AVX512 vectorization in the JVM now as well – again if we write the code intelligently and it uses lots of threads and it's not terrible with how it addresses memory, the KNLs for us have been a huge win."

Memory was another plus in Phi's column, but DUG will provide other alternatives based on price-performance advantage and customer demand. "If you just look at the price of a high-end GPU, KNL comes with 16 Gigs of on-package memory, which is huge, to get a 16 Gig GPU you are talking many multiples of the price we pay for KNLs," he said. "So it's a no brainer for just a bang-for-buck perspective. But at the end of the day we are not really religious about it – if something else comes along that has better TCO, then we'll buy that instead. If we have McCloud clients as we already do who say we must have GPUs because we have this or that code that we don't want to rewrite, then we'll get the GPUs."

Although the new service is named "McCloud," Schwan himself is a little wary of the cloud terminology.

"Nowadays everybody is talking about cloud, but some people still hear cloud and they go, 'Yuck, I don't want to do anything with the cloud. The cloud is a pain in the ass. They don't understand my business.' We've tried to provide something that is geared for geoscience, for the market we know really well that provides as much or as little as they want to take advantage of. So it can be just hardware cycles, which is fine, but I think in a way is the less interesting part of the solution. We also provide our entire software stack for as much or as little as they want to use. Our own service business expertise, so for example if you're a major oil company who is really excited about FWI or wanting to focus on migration, but maybe you don't want to do all the preprocessing – you could absolutely get us to do the preprocess on DUG McCloud and then you go and do your special sauce."


DUG CTO Phil Schwan presenting at Rice Oil and Gas conference on March 5, 2019 – click to enlarge

Speaking of special sauce, a major one for DUG is immersive cooling.

The new immersion tanks at Skybox are DUG's 6th iteration design. The delta-T for the input water and output water is only about 4-5 degrees Celsius, and they can do that pretty much anywhere in the world, even in Perth and Houston summers, with evaporative chillers.

"If you only have to get it from 35 to 30 [degrees Celsius] that's pretty easy to do," said Schwan of the cooling technology. "But the design inside the room, actually getting everything lined up so you have the right amount of flow, and the right amount of pressure and you have the right valves and you make sure the tank at the very end of the row has the same cooling capacity as the tank at the very beginning of the row – all of these things just take detail and somebody who's willing to crunch the numbers and make sure from an engineering perspective that it's all done right. We haven't invented immersion cooling by any means, but I think we've just tried to look at every element and simplify as much as possible, whether it's something as obvious as not having lids on the tank to having a heat exchanger that's immersed in the tank so the fluid never leaves the tank – that was, as far as I know, the first time that's ever been done. And it just eliminates so much complexity, piping and manifolds, and pumping and computers to control all of it – we just don't need any of that."

DUG's initial ~40,000-node install at Skybox requires 500,000 litres of polyalphaolefin dielectric fluid (a standard synthetic oil – it's not 2-phase so it doesn't evaporate, hence no need for covers) to fill all the tanks, which are being brought in this week. At this stage, all of the plumbing in the room is complete, all of the underfloor, electrical and the panels are complete, and all the raised flooring is in. Outside they've started to put all the pumps down, started to put chillers in place, as well as the big switch gear, the big transformers. Energizing the facility is slated for the second half of April.

Inside Skybox data hall 2, future home of DUG's ~40,000 KNL nodes. 5,000 1,000-litre IBCs will be brought in to fill all the tanks with coolant. (Photo provided by DUG.)

Bringing Australian voltage to the U.S.

DUG's facility at Skybox will run at 240 volts. In the United States, nearly all datacenters run at 110 volts, of course. So how did this happen? It typically takes two transformations to get down to 110 volts from medium voltage, and each one of those transforms costs you about 5 percent, Schwan explained in his presentation. They were not willing to give up that 5 percent.

"We decided to run at more typical Australian voltage – 240 volts – and we do it with a single transform. None of the sparkies wanted to deal with it here," said Schwan. "We had to work very closely with Skybox and with the utility to make this happen but we were able to make it happen."

Photo showing 1 MW panel boards, presented by Phil Schwan at Rice Oil and Gas conference on March 5, 2019

The benefits didn't stop with the efficiencies gained by cutting out some of the transforms. "Because it's 240 volt, we're obviously running with lower current, and this means we can get away with fewer circuit boards, fewer circuit breakers, fewer PDUs, fewer everything. Each one of these boards is a 1 MW panel board – it operates an entire row of tanks – and we run 70 amp 3-phase all the way to the tank, which means we get away with about half as many of these as we otherwise would have. We were able to use this otherwise dead space – in a room this size, we have these posts to hold the ceiling up and by being able to fit a minimum number of panel boards in this otherwise dead space, we can fit an extra tank in the room – that's an extra 2.5 percent worth of gear."

I asked Schwan if he thinks other American datacenters should be following suit and where the tipping point is to gain an advantage by switching to 240 volt. “I think that most people doing HPC at the scale of the oil and gas industry are big enough to benefit from it,” he said. “But I think there are bigger wins even before that. Like immersion, I don’t know how anybody doing HPC at scale doesn’t see something like our immersion solution and not have it become a personal religion. The savings are just so compelling, the economics are just overwhelming.”

Outside the Box

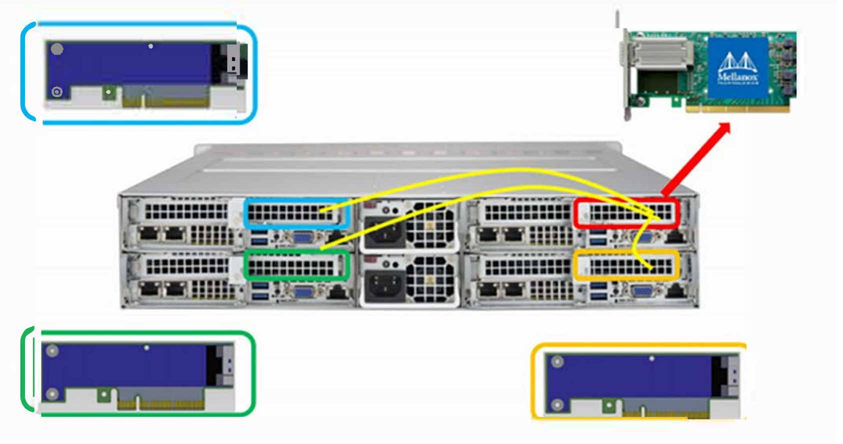


Diagram of 2RU chassis with Mellanox networking schema

Schwan also had some interesting things to report on DUG’s other hardware choices for its Skybox deployment, noting that while they’ve always done standard 10 Gigabit Ethernet, they went a different direction with this new facility, working closely with Mellanox. “We had everybody in the universe wanting to make a bid for that network and some very interesting proposals came out – but at the end of the day, the Mellanox solution really was a very outside of the box solution and provided some amazing advantages,” he said.

The technology, created exclusively for DUG, allows four servers to use a single 50 Gb/s network connection, with each node able to burst to over 30 Gb/s. The design relies on Mellanox’s SN2700 32 port 100 Gb/s Ethernet switches.

In each four-node chassis, DUG deployed a multi host network adapter with a NIC in node number one and the three other boards are connected by PCIe. Schwan said this delivered a slew of benefits. The most obvious is since they only had one external connection coming into every chassis that resulted in a quarter as many cables, a savings of ~30,000 cables. They were also able to get 200 nodes on a single switch, which reduced by about half the total number of switches in the entire room.

“But it also gives us some things that we weren’t really expecting,” Schwan said, “for example the standard networks we operate, they are all 10 Gigabit Ethernet and every node stands on its own so they can peak at 10 Gb/s. This one is a single 50 Gb/s connection coming into the node and any node on its own can burst at 30 Gb/s – and of course can do 12.5 Gb/s each, but especially within this four-node chassis, we have very low latency and extremely high bandwidth, which we take advantage of in applications.

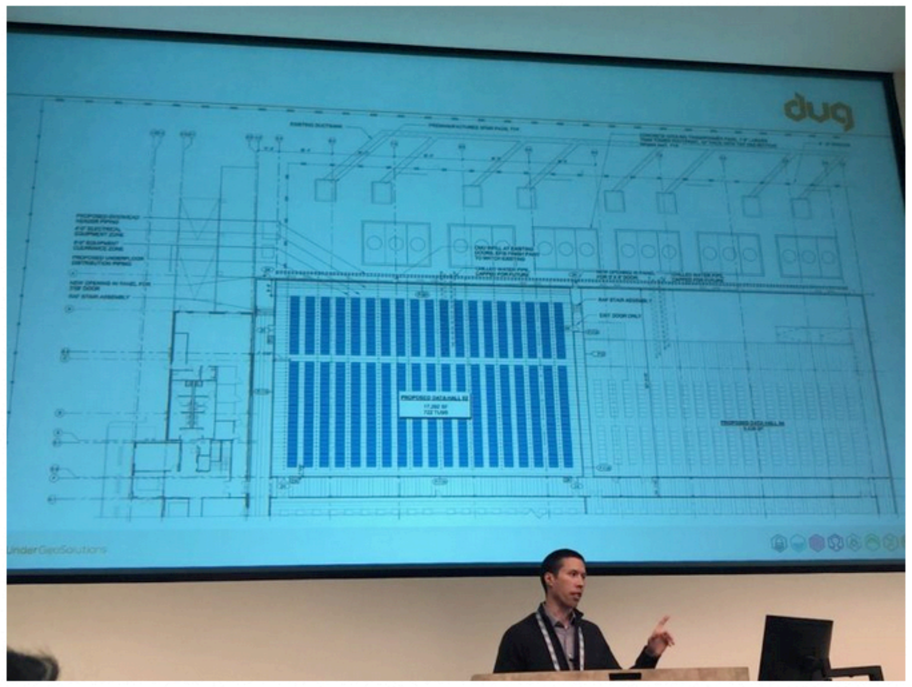
“The other thing worth emphasizing is none of that bridging in the master node goes through the OS it’s all handled by the card, so it’s not adding any extra load to that master node. The master node doesn’t even have to be on; as long as its plugged into the chassis it gets the power that it needs to power that master node and do the bridging. We haven’t created any new single points of failure in that chassis by having this single NIC.”

With all of these combined efficiencies, DUG’s datacenter achieves a power usage effectiveness (PUE) under 1.05.

We’re Number One?

We’ve previously reported that despite being on the bleeding edge of industrial compute capacity, DUG does not have Top500 aspirations. That has not changed. “Our application demands are nothing like what the Top500 demands and there is virtually no value for me in putting in an astronomically low latency interconnect, which is what I would need to do to do Top500 effectively on that number of processors,” said Schwan.

DUG has big plans to expand into the massive Skybox campus and talks confidently about reaching exascale.



Skybox Houston blueprint. DUG CTO Phil Schwan presenting at Rice Oil and Gas conference on March 5, 2019 – click to enlarge

“We’re going to put 40,000 compute nodes in this blue area, we have the data hall next door the same size and ready for us to start building as soon as this one’s finished, and then what you can’t see off the top of this diagram past the service yard, is 10 acres of land that we have architectural plans to build on when these two are full, so by the time this is done, this should easily be a multi-exaflop campus of compute centers,” Schwan said in his presentation, referencing his slide (see photo at right).

The data hall with all the small blue rectangles (each representing one tank) will draw about 15 MW. This comes out to about 8.5 kilowatts per square meter or about about 800 kilowatts per square foot. This is about three to four times the density of what you’d find in a typical, still fairly high density colo facility, by Schwan’s account.

Given how beneficial Phi has been for DUG, now that Phi is gone, what will they buy next? Schwan, who maintains a spreadsheet evaluating various processors’ TCO, likes the Intel Xeon roadmap, which he sees as having all of the best features of Phi.

“If you you look at the roadmap for Xeons going forward, it’s got AVX512, it’s got 40, 50, 60 cores, it’s got high-bandwidth memory as an option on package – well what is that – it’s a Xeon Phi. Assuming the TCO continues to be as excellent as it has been from Intel, I think the classic Xeon line will be a very natural fit for us. They almost have the accelerators built onto the chips now – but every six months, we look at it again, to see what’s interesting.”

Schwan reported that DUG has committed customers but it is not ready to announce them yet. “We’re not ready to make a big splash because I don’t have anything to sell them yet, that will be in Q2.”

Share this:

Tweet Share reddit this!