Chip startups come and go. Generally, we cover them because of novel architectures or potential for specific applications. But in some cases, like today, it is for those reasons and because of the people behind an effort to bring a new architecture into a crowded, and ultimately limited, landscape.

With $100 million in “patience money” from a few individual investors who believe in the future of sparse matrix-based computing on low–power and reprogrammable devices, Austin-based Knupath, has spent a decade in stealth mode designing and fabricating a custom digital signal processor (DSP) chip to target deep learning training, machine learning–based analytics workloads, and naturally, signal processing. The company, led by famed NASA administrator Dan Goldin, has already fulfilled a $20 million contract for its first generation DSP–based system and has an interesting roadmap ahead, which will include the integration of FPGAs, among other devices.

Goldin, who was the longest–acting administrator for NASA (from 1992 through 2001), oversaw numerous space missions and helped develop nuclear–powered ion engines for future use in Mars missions, among other achievements. He took a mid–career break to work for TRW where he developed the direct satellite communications technology that allows for satellite television, not to mention a range of other scientific uses. Goldin's career took a definite turn in the early 2000s, when his research interest flipped to neuroscience, sending him into a fellowship under Nobel Prize–winning neuroscientist Gerald Edelman, where he fleshed out connections that were forming in his mind about neuroscience and computing—and what those links might mean beyond neuromorphic computing devices.

As Goldin tells The Next Platform, the range of future applications he was watching were moving beyond Von Neumann architectures and beyond dense matrix approaches to problem–solving. His focus shifted to a well–established area in supercomputing applications—sparse matrix–based problems—and what might be needed to compute against those with lower latency via a more efficient memory model and lightweight cores. The result, which has taken over ten years to develop, was a chip delivered to Knupath's first customer in late 2015. Knupath's second generation product, which will emerge in late 2017 is the “Hermosa” chip, is an internally designed and fabricated custom DSP with the novel “Lambda” fabric connecting both multiple Hermosa chips together and, potentially, racks of systems employing them. All of which is designed to tackle an elusive future dominated by sparse matrix–based computations as Goldin predicts will be increasingly found in machine learning applications.

Such a system is scalable to 512,000 chips. Each chip has 256 tDSP cores (the “t” is for “tiny” with a single ARM management core). The latency story is a compelling one, with rack to rack latency of 400 nanoseconds (as good as the fastest Ethernet today)—all with the ability to handle sparse matrix computations efficiently and especially. So far, there have been some research forays into the possibility of shifting to a sparse matrix–fueled future for deep learning in particular, but there is, as of yet, no platform for doing so. Indeed, this would mean a completely shift in workflows, Goldin is betting his future on the fact that this shift will be well worth the effort.

The first generation product is PCIe based with multiple processors that can be built across many
Hermosa processors. The fabric is heterogeneous and can accept multiple types of processors (X86, GPUs, FPGAs) and the company intends that the second generation in 2017 will support all of these. It is also multi-program and multi-data, which means 256 different algorithms could, in theory, run on each of the Hermosa cores, which Goldin says will be increasingly important for future applications in both signal processing and machine learning as well as for some customers they are eyeing in financial services.

The Lamda fabric is the most interesting piece here. It can scale from a small number of processors on a board or a single processor in an edge device, all the way up to 512,000 chips. It is based on a distributed memory model wherein there is shared memory among the cores and that shared memory is also distributed in the system, which is why you see the DMA controllers (which move data around the system). There is 320 GB/s outbound from the processor with 16 10 GB/sec links going outbound bi-directional. All of this leads to aggregate memory bandwidth of 3.7 terabytes per second across the machine. On the scalability front then, each little “cluster” has the memory shared among the DSPs so the memory bandwidth numbers scale proportionally to the number of chips (adding more means adding more memory and memory bandwidth into the system).

Each Knupath processor core has built-in instructions for communications, sync and those for helping with sparse matrix problems (scatter/gather handling capabilities) and programmable DMA engines to support the memory footwork of the 256 tDSP cores. In terms of scalability, beyond the 512,000 addressability arrays that can be built, the system can provide up to 3.702 GB/sec memory bandwidth. The peak power consumption hits 34 watts per chip, providing a solid watt-to-watt performance comparison on FFT and other benchmarks the company has run (but will not share more about until after the launch in 2017).

We have talked quite a bit here about the use of GPUs for deep learning and related workloads, the potential for FPGAs there, and of course, on-off efforts from Google (the TPU is the best example) and others, so the question is, why use a DSP? There are, after all, some novel uses of the processors at scale on the horizon, including on the forthcoming Tianhe-2 supercomputer, but why aren't these in use beyond signal processing?

Goldin says that their major focus is still processing, but turning those capabilities into sparse matrix functionalities in a power-efficient envelope turns out to be a secondary use case. Although he could not share benchmarks, he noted that AlexNet and GoogleNet performance is between 2X to 6X—but of course, without detail we cannot comment much on that. On a more practical note, Goldin did mention two other things that might account for the choice. First comes cost, second comes programmability. DSPs are inexpensive (relatively, of course), but instead of licensing the technology from Texas Instruments or elsewhere, Knupath designed their own. On the programming side, Goldin says they wanted the flexibility of DSPs, especially for users on the signal processing side where finding and keeping FPGA programmers is an expensive, time-consuming challenge.

"We wanted to have the processing in immediate vicinity of the memory—a push model. You don't need the cache, you don't need to do fetch. We didn't design this just for processing, we balanced communications and processing in memory to keep balance. It's a communicator—there's a router right in the middle of it," Goldin explains. Unfortunately, the reason they signed a contract for the first chip, which came out in 2015, is because of that eDRAM feature, which put each of the tDSPs right adjacent to memory for immediate contact. While the next variant of their chip won’t be able to use it, they have found a suitable workaround, although they were not able to provide details as of yet.

The programming model for the PCIe-based accelerator version is much like CUDA/OpenCL when it
comes to communicating between the host system and chip with a more MPI-like model for communicating between the chips within the system. Unlike with a GPU, however, there is the capability to communicate between the cards and allow them all to talk to each other without going through PCIe or the CPU. Goldin says his team is working with Dr. Larry Smarr at CalIT2 to develop a future contest focused on working with sparse matrix operations in order to bolster the software side of the platform.

Goldin says the company is looking at partnerships for its next iteration, including FPGA and GPU makers. He also says there will be a bolstering of the ARM count for future generations.

“One thing to note is how this is different in terms of how data moves through the fabric,” Goldin explains. “Instead of fetching data and the application from memory, we are sending data through the architecture, and in that payload packet comes not only the data to be calculated but the pieces of programming required to do so, as well as the next destination for that data.” Ultimately, as other data flow architectures aim to do, it turns the Von Neumann architecture on its head.

“If you look at the human brain with its 200 billion neurons, each connected to maybe 10,000 to 100,000 neurons, it’s the most efficient approach to computation we know. That was our starting point, and that’s the reason the Lambda fabric is different than other architectures—it’s digital, not neuromorphic, but it is based off the same principles as the mammalian brain.”

To be clear, the Hermosa processor can still handle dense matrix well, even if it’s not the target. “We are still in the Wild West when it comes to machine learning,” says Goldin. “As we develop different algorithms, the platform has to be there and the new wave those are riding on is sparse matrix.”

The Next Platform made a few rounds of calls to the hyperscale companies we talk to regularly who are doing machine learning at scale and while none of these conversations were on the record, the takeaway is that while its still early days, there are indeed explorations about what performance, efficiency, and programmatic advantages there might be for sparse matrix approaches to deep learning training. However, it’s too early to tell, and the effort required to shift to such a model would require the promise of great returns.

And so, with that said, we’ve presented what we know about the architecture but to date, there is very little outside of dense research on the future of scatter–gather/sparse matrix for deep learning. Still, for $100 million in funding for a DSP–based approach, we have to think that someone saw an opportunity here, especially given the other deep learning–focused chips we’ve seen emerge over the last year. The other angle is that the initial customer was for some specific flavor of signal processing and playing up the machine learning side of the story was the best way to garner wider appeal, even if that’s a story that is still developing and to some degree unknown.