Soul Machines stokes next-generation AI engines

As corporations seek to connect with customers in meaningful ways...

Soul Machines forges a new frontier in personalized service with its Human OS platform. The startup aims to help democratize the service industry by creating Digital People™ that run on AI-based simulation, providing seamless interactions with end users. To train machine learning and AI models, Soul Machines requires powerful machines with fast, efficient graphics capabilities that are not constrained by memory and compute capacity. The company used the HP Z8 G4 Workstation with NVIDIA RAPIDS™ software suite to set the course for large models such as an in-house version of Facebook's BlenderBot, a machine-learning chatbot on one workstation—key steps toward revolutionizing the customer experience.

soul machines™

Industry
Technology

Objective
Run large, memory-intensive artificial intelligence training models and virtualizations on the HP Z8 G4

Approach
Train complex machine-learning models such as an in-house version of Facebook's BlenderBot with 9 billion parameters

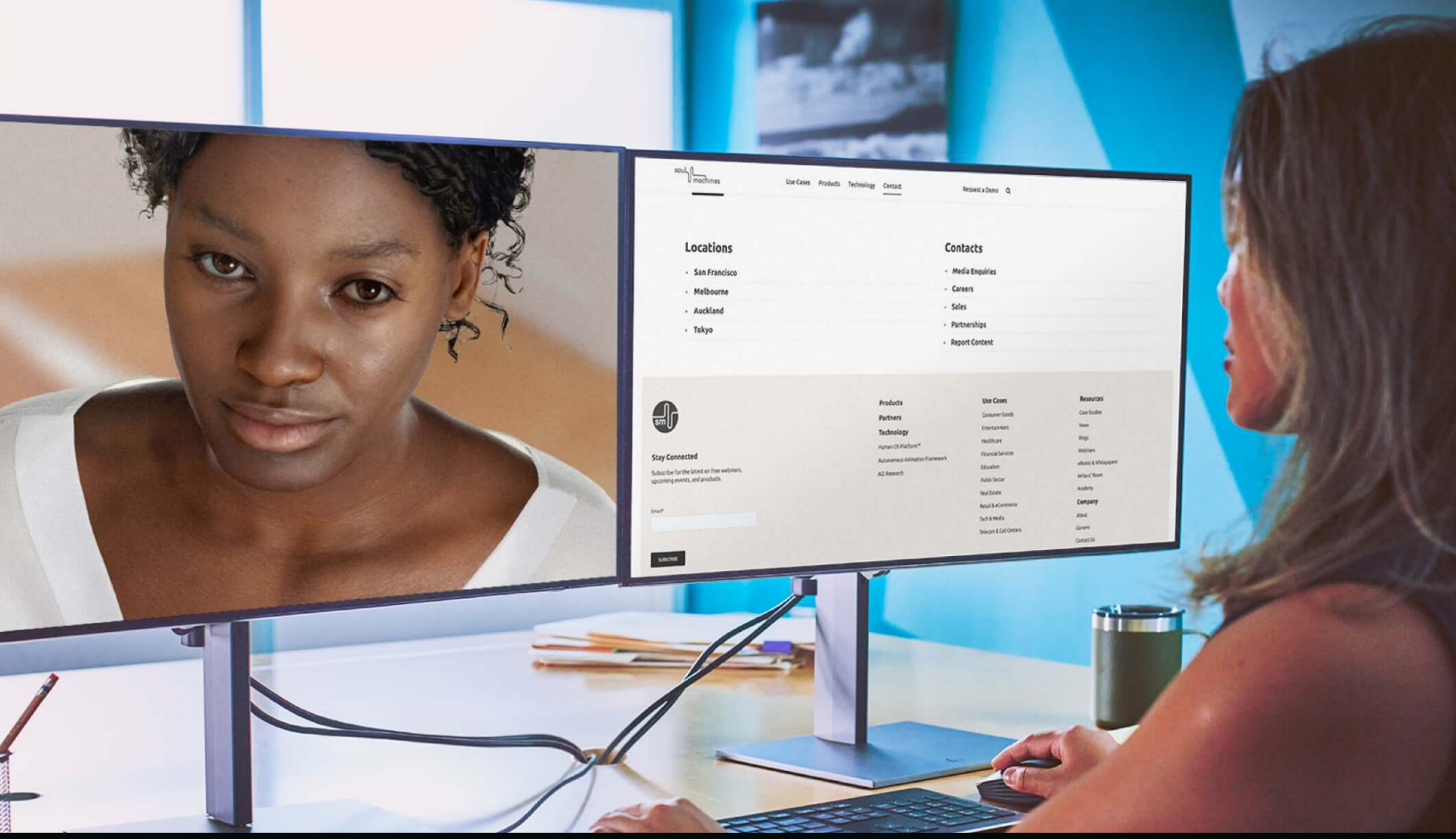
with intel

The goodness of human and machine collaboration

Soul Machines has pioneered the creation of Digital People™, combining hyper-realistic computer-generated imagery and autonomous animation resulting in human-like interactions for powerful brand experiences across industries worldwide. “Soul Machines’ ambition is to create Digital People to help democratize the service industry so we can help with education, medicine, and other fields where a personalized service is a key difference,” says Shane Blackett, Vice President of Software and Technology at Soul Machines.

Based in San Francisco with research and development operations in Auckland, New Zealand, Soul Machines works with visionary clients who want their end user to have a delightful customer experience.

> “Having a computer that could just facilitate it without having to go to that extra level of engineering was critical. Otherwise, we probably just wouldn’t have done it.”
>
> Shane Blackett, Chief Technology Officer, Soul Machines



"We want to go further than that to where the digital person and the user and the content on the webpage, for example, co-exist in a three-way interaction where the digital person might suggest something, the user might do something, or they might work together," Blackett says. "We're trying to provide a scalable platform for creating those Digital People and we're running an AI-based simulation to provide the fidelity and realism."

Soul Machines has developed a Human OS™ Platform, featuring a patented Digital Brain that helps deliver human and machine collaboration. The company works on a variety of models such as AI, machine and associative learning, and other complex solutions. Soul Machines' customers cross industry sectors, from financial institutions to consumer goods. Case in point: A digital person for a leading global CPG company in Japan sells cosmetics online in Japanese and English.

> "In terms of the machines we have available, the Z8 is probably the most powerful single machine."
>
> Alireza Nejati, Senior Principal R&D Software Engineer, Soul Machines

Models running under the digital persona are taught to recognize cognitive processes from learning and sensing to behavior. To train and run various AI-based and machine learning models and render these images, however, requires tremendous compute power, video memory, and graphics. Enter the HP Z8 Workstation.

“We have lots of training models that are memory constrained,” notes Alireza Nejati, Senior Principal R&D Software Engineer at Soul Machines, whose team members work on the HP Z8. “The big motivating factor for us was the amount of graphics (GPU) memory available on a single machine.”

IT matters

- ✓ Harness compute power of the HP Z8
- ✓ Utilize large GPU memory capacity
- ✓ Train significant machine learning and AI models on a single machine
- ✓ Use cluster management software to assign tasks
- ✓ Expand virtualization competencies

Business matters

- ✓ Explore new, available technologies to aid AI development
- ✓ Realize speed and efficiency
- ✓ Reduce need for manual intervention
- ✓ Make networks learn in real time
- ✓ Leverage a high-performing workstation for future projects

Wielding memory and power

"In terms of the machines we have available, the Z8 is probably the most powerful single machine," Nejati says. "It allowed us to run a lot of our models, which wouldn't be possible with just any machine because it takes a lot of memory."

He points to the Z8's configurability, dual Intel® Xeon® Gold processors, up to 3 TB DDR5 ECC memory, four NVMe M.2 slots for high-speed data storage, and up to 96 GB of GPU memory using two NVIDIA RTX™ 8000 graphics cards. The Z8's Intel® Optane™ DC Persistent Memory is also a benefit when it comes to working with very large, complex datasets.

Initially, Soul Machines deployed the Z8 as a server, running it manually. The company then turned to the NVIDIA RAPIDS™ suite of software libraries, which was bundled with the Z8 to manage resources more efficiently. The largest model Soul Machines had previously trained on the HP Z8 was an in-house version of Facebook's BlenderBot, an open domain chatbot trained in online dialog from multiple sources with curated datasets. While BlenderBot isn't as large as some other machine-learning models, it has 9 billion parameters.

"It's quite challenging to train it all on one machine and you really need all 96 gigabytes of video memory to train that model," Nejati says. "If it was 50 gigabytes, you wouldn't be able to do it."

Previously, when models were smaller, he says it was possible to fit a model on a single GPU for training.

"Modern models are so big that even just one instance of the model needs to be spread across multiple GPUs, and this is infinitely easier when all the GPUs are on the same machine," Nejati says. "When the GPUs are on different machines, it becomes another software engineering challenge altogether. You need to get it to properly distribute the whole model across multiple GPUs because you're not running it in parallel, it's running sequentially."

Customer at a glance

soul machines™

Application

Training machine-learning and AI-based models to render a personalized customer experience

Hardware

HP Z8 Workstation

The ability to do more

The HP Z8 also gave Soul Machines the opportunity to use resources more effectively. The BlenderBot project, for example, was an effort to learn how to push that technology in conjunction with Soul Machines' software. The power of the HP Z8 enabled a small company like Soul Machines to get it done.

"Having a computer that could just facilitate it without having to go to that extra level of engineering was critical," Blackett says. "Otherwise, we probably wouldn't have done it."

Aside from training models, Soul Machines has used the HP Z8 to do more with virtualization.

"One big challenge in developing these machine-learning models is that you've got to be able to run other people's code and build on it," Nejati says.

Running open-source code developed on various machines natively can be arduous, with an environment ripe for errors. Virtualization, Nejati says, offers a big benefit.

"You can just set up a virtual machine on your computer with any specifications you want, and it will run that code," he says. "When you're running machine-learning models, you want all the speed you can get. The HP Z8 gives you a lot of power and it allows you to run virtual machines as if they were native machines. That's actually saved us a lot of time."

> "It really took a more powerful computer like the Z8 to make the virtualization work well. There were certain things that I thought were really impossible and virtualization made them possible."
>
> Alireza Nejati, Senior Principal R&D Software Engineer, Soul Machines

There were models that Soul Machines tried to run locally on other workstations, but they did not work without virtualization.

"It really took a more powerful computer like the Z8 to make the virtualization work well," Nejati says. "There were certain things that I thought were really impossible and virtualization made them possible."

Making rapid and real-time machine learning—where networks are trained to be more learnable with only a few training samples—feasible has become necessary to create a seamless user experience. Having a system that serves immediate requests quickly is a big plus. The Soul Machines team learned that sending complicated machine-learning tasks to the HP Z8 was much faster than doing those jobs locally.

"Building that system on the hardware we were previously using was just not possible," Blackett says.

As the world moves toward more sophisticated intelligence and data, a workstation like the HP Z8 opens many doors for Soul Machines. Having the ability to build on models and investigate new possibilities is essential as the company grows.

"I'm sure there will be other challenges that we'll come across and we'll say we need some way to investigate, evaluate how that works with our business," Blackett says. "And if we're shut out of that because we just don't have gigantic computing power, then that doesn't help us as a startup. Knowing that we have tools like the Z8, to go into some of these really big models, really matters."

Learn more about Z by HP Workstations and solutions for data science

Learn more

Product may differ from images depicted.

4AA7-9918ENW, December 2023