Meta Is Making a Monster AI Supercomputer for the Metaverse

By **Jason Dorrier** - Jan 26, 2022

Meta is building a new supercomputer to train enormous machine learning algorithms. Though only partially complete, the AI Research Supercluster (RSC) already ranks among the most powerful machines on the planet. When it's finished, the company formerly known as Facebook says it will be the fastest AI supercomputer anywhere.

Meta hopes RSC can improve their products by training algorithms that better surface harmful content. Further out, the company says advances might enable real-time language translation between tens of thousands of people online and multitasking algorithms that can learn from and generalize across different inputs, including text, images, and video.

All this, the company said, will help advance real-world applications like robotics and, of course, build the foundations of the (as yet primordial) metaverse. "In the metaverse, it's one hundred percent of the time a 3D multi-sensorial experience, and you need to create artificial-intelligence agents in that environment that are relevant to you," Jerome Pesenti, Meta's VP of AI, told the *Wall Street Journal* this week.

Whatever the ultimate applications, the investment shows tech's biggest players—from Meta to Alphabet and Microsoft—deem it increasingly crucial to be competitive in cutting-edge AI.

Big AI Is in Vogue

The announcement is part of a trend towards ever-bigger machine learning algorithms requiring greater computing resources and bigger data sets.

In 2020, OpenAI's natural language algorithm GPT-3 showed big gains could be realized by growing the number of internal connections in algorithms, known as parameters, and the amount of training data piped through them. With 175 billion parameters, GPT-3 was 17 times larger than its predecessor GPT-2. Encouraged by GPT-3's success, Microsoft unveiled its Megatron AI last year, an algorithm three times bigger than GPT-3, and Google and Chinese researchers each built algorithms with over a trillion parameters. Anticipating the next step, Meta said they plan to use RSC to train algorithms with trillions of parameters.

Increasingly, these sprawling algorithms require supercomputers, the room-sized machines scientists use to simulate physical systems, from elementary particles to Earth's climate to the universe at large. Last year, for example, OpenAI announced its partner Microsoft had built a dedicated supercomputer to train its models. According to the companies, the new machine ranked in the top five fastest supercomputers in the world (at the time).

Though Meta didn't give numbers on RSC's current top speed, in terms of raw processing power it appears comparable to the Perlmutter supercomputer, ranked fifth fastest in the world. At the moment, RSC runs on 6,800 NVIDIA A100 graphics processing units (GPUs), a specialized chip once limited to gaming but now used more widely, especially in AI. Already, the machine is processing computer vision workflows 20 times faster and large language models (like, GPT-3) 3 times faster. The more quickly a company can train models, the more it can complete and further improve in any given year.

In addition to pure speed, RSC will give Meta the ability to train algorithms on its massive hoard of user data. In a blog post, the company said that they previously trained AI on public, open-source datasets, but RSC will use real-world, user-

generated data from Meta's production servers. This detail may make more than a few people blanch, given the numerous privacy and security controversies Meta has faced in recent years. In the post, the company took pains to note the data will be carefully anonymized and encrypted end-to-end. And, they said, RSC won't have any direct connection to the larger internet.

To accommodate Meta's enormous training data sets and further increase training speed, the installation will grow to include 16,000 GPUs and an exabyte of storage—equivalent to 36,000 years of high-quality video—later this year. Once complete, Meta says RSC will serve training data at 16 terabytes per second and operate at a top speed of 5 exaflops.

If completed today, that would make RSC the fastest AI supercomputer in the world. But it's worth digging into what exactly that means for a moment.

Apples to Apples?

Supercomputers vary widely in how they're built. Common configurations include both central processing units (CPUs) and GPUs, but the makers of the chips differ, as does the infrastructure wiring them all together. To compare supercomputers, the industry uses a benchmark called floating-point operations per second—or more colloquially, flops—which measures the number of simple equations a supercomputer solves each second.

According to the most recent Top500 list, the world's fastest all-around supercomputer, Fugaku, hails from Japan.

Fugaku, which doesn't actually use any GPUs, recorded a blistering top speed of 442 petaflops (or 442 thousand trillion operations per second). That's fast. But systems like Fugaku are increasingly built to train AI too. So, Top500 began reporting a new benchmark for AI applications specifically. Since machine learning algorithms don't require the same precision as scientific applications, the new AI benchmark uses a less precise measure. By that measure, Fugaku hits peak speeds above an exaflop—or a million trillion operations per second. This is what's meant by an AI supercomputer.

Now, back to Meta.

Most machines on the Top500 list are operated by governments and universities. Private supercomputers, like RSC and the machine built by OpenAI and Microsoft, don't appear on the list. For performance, we have to take the companies at their word. Assuming RSC hits peak speeds of 5 exaflops for AI applications, it would beat Fugaku by a decent margin. But whether that will still be best in the world later this year isn't as clear. The upcoming Frontier supercomputer is expected to be three times faster than Fugaku for high-precision applications. Also built for AI, Frontier will be stiff competition for top AI supercomputer.

It's also worth noting peak performance on a benchmark is not equivalent to actual performance on real-world workloads. According to high-performance computing analyst Bob Sorensen, "The real measure of a good system design is one that can run fast on the jobs they are designed to do. Indeed it is not uncommon for some HPCs to achieve less than 25 percent of their so-called peak performance when running real-world applications."

An emerging AI benchmark, called MLPerf, is closer to measuring performance on real-world tasks. It doesn't yet measure how fast systems train very large models, but it's still a helpful comparison. In the most recent MLPerf results, systems using NVIDIA A100 chips, the same as those used to build RSC, dominated the field. And the biggest system tested, NVIDIA's own Selene AI supercomputer, trained the (now-diminutive) BERT language processor in just 16 seconds, compared to 20 minutes for smaller systems.

So any way you slice it, RSC will be (and already is) a formidable machine for AI research.

Is Bigger AI Always Better?

To date, building bigger and bigger algorithms does seem to yield gains. But not all researchers believe those gains will continue forever or always be worth the spiraling energy and financial resources needed to train algorithms. Large language models, in particular, also tend to pick up all manner of unsavory habits and biases during training.

Luckily, there's also work afoot to make algorithms more efficient and accountable.

Last year, AI research organization DeepMind released a 280-billion-parameter large language model called Gopher that could outperform other large language models. More interestingly, however, they also developed a much smaller 7-billion-parameter model called RETRO. Given the ability to consult an external database of examples to inform its predictions—a memory, of sorts—RETRO punched well above its weight class by matching or outperforming algorithms 25 times its size. DeepMind said it's also easier to trace the algorithm's reasoning, making it more transparent and potentially easier to eliminate bias.

So, while making enormous algorithms on supercomputers is eye-catching, RETRO shows innovation in *how* those models are built is equally important. Research on both ends of the spectrum will likely continue apace, one hopefully feeding into and improving the other.

Image Credit: Erick Butler / Unsplash

JASON DORRIER

Jason is editorial director of Singularity Hub. He researched and wrote about finance and economics before moving on to science and technology. He's curious about pretty much everything, but especially loves learning about and sharing big ideas and advances in artificial intelligence, computing, robotics, biotech, neuroscience, and space.