https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

"Blackwell Innovations to Fuel Accelerated Computing and Generative AI

Blackwell's six revolutionary technologies, which together enable AI training and real-time LLM inference for models scaling up to 10 trillion parameters, include:

- **World's Most Powerful Chip** — Packed with 208 billion transistors, Blackwell-architecture GPUs are manufactured using a custom-built 4NP TSMC process with two-reticle limit GPU dies connected by 10 TB/second chip-to-chip link into a single, unified GPU.
- **Second-Generation Transformer Engine** — Fueled by new micro-tensor scaling support and NVIDIA's advanced dynamic range management algorithms integrated into NVIDIA TensorRT™-LLM and NeMo Megatron frameworks, Blackwell will support double the compute and model sizes with new 4-bit floating point AI inference capabilities.
- **Fifth-Generation NVLink** — To accelerate performance for multitrillion-parameter and mixture-of-experts AI models, the latest iteration of NVIDIA NVLink® delivers groundbreaking 1.8TB/s bidirectional throughput per GPU, ensuring seamless high-speed communication among up to 576 GPUs for the most complex LLMs.
- **RAS Engine** — Blackwell-powered GPUs include a dedicated engine for reliability, availability and serviceability. Additionally, the Blackwell architecture adds capabilities at the chip level to utilize AI-based preventative maintenance to run diagnostics and forecast reliability issues. This maximizes system uptime and improves resiliency for massive-scale AI deployments to run uninterrupted for weeks or even months at a time and to reduce operating costs.
- **Secure AI** — Advanced confidential computing capabilities protect AI models and customer data without compromising performance, with support for new native interface encryption protocols, which are critical for privacy-sensitive industries like healthcare and financial services.
- **Decompression Engine** — A dedicated decompression engine supports the latest formats, accelerating database queries to deliver the highest performance in data analytics and data science. In the coming years, data processing, on which companies spend tens of billions of dollars annually, will be increasingly GPU-accelerated.

A Massive Superchip

The NVIDIA GB200 Grace Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU over a 900GB/s ultra-low-power NVLink chip-to-chip interconnect.

For the highest AI performance, GB200-powered systems can be connected with the NVIDIA Quantum-X800 InfiniBand and Spectrum™-X800 Ethernet platforms, also announced today, which deliver advanced networking at speeds up to 800Gb/s.

The GB200 is a key component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation

NVLink. Additionally, GB200 NVL72 includes NVIDIA BlueField®-3 data processing units to enable cloud network acceleration, composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds. The GB200 NVL72 provides up to a 30x performance increase compared to the same number of NVIDIA H100 Tensor Core GPUs for LLM inference workloads, and reduces cost and energy consumption by up to 25x.

The platform acts as a single GPU with 1.4 exaflops of AI performance and 30TB of fast memory, and is a building block for the newest DGX SuperPOD.

---

https://www.techradar.com/pro/this-is-what-nvidias-exaflop-supercomputer-in-a-rack-looks-like-the-dgx-gb200-nvl72-tower-most-likely-uses-48v-25ka-to-deliver-a-staggering-1440-petaflops-could-cost-millions#:~:text=DGX%20GB200%20NVL72%20weighs%201.36,stack%20system%20seems%20a%20possibility.

DGX GB200 NVL72 weighs 1.36 metric tons (3,000 lbs) and consumes a 120kW, a power load that Serve The Home points out, not all data centers will be able to handle. As many can only support a maximum of 60kW racks, a future half-stack system seems a possibility. The rack uses 2 miles (3.2 km) of copper cabling instead of optics to lower the system's power draw by 20kW.

--

https://www.datacenterknowledge.com/data-center-chips/nvidia-launches-next-generation-blackwell-gpus-amid-ai-arms-race-

"For example, organizations can train a GPT 1.8 trillion parameter model in 90 days using 8,000 Hopper GPUs while using 15 MW of energy. With Blackwell, organizations could train the same model in the same amount of time using just 2,000 Blackwell GPUs, while using only 4 MW of power, he said."

---

12 years ago 1 Exaflop was targeted by the government to consume no more than 20 megawatts with an expected completion by 2020, which has been more or less reached (the exascale part, at least- initially with three supercomputers instead of one). In 2024 the power figure has been reduced 80%.

". The research group concluded that "while an Exaflop per second system is possible (at around 67MW [megawatts]), one that is under 20MW is not. Projections from today's supercomputers... are off by up to three orders of magnitude.""

According to the NVL72 power consumption, each 120kW rack runs 20 Petaflops, which means that 50 racks is one exaflop, and would consume around 6 megawatts. Still a significant reduction at FP4. Seems like FP2 (if even possible) would use around 3 megawatts and FP1 would use 1.5 Megawatts (at 4nm). Most tensor units such as Google's TensorFlow and Groq use FP8, thus this figure appears to rely on low precision compute, which LLMs increasingly use. Thus once 18A nodes are produced, the power consumption will increasingly be lowered (less than a megawatt for a 1-bit LLM exaflop)