Huawei launches Atlas 950 SuperCluster — ambitious to achieve 1 ZettaFLOPS FP4 performance, scale of hundreds of thousands of APUs

Huawei launches Atlas 950 SuperCluster with the goal of achieving 1 ZettaFLOPS FP4. The system scales to hundreds of thousands of APUs, promising to open the era of super powerful AI.

At the Huawei Connect 2025 event, the Chinese technology giant officially launched **the Atlas 950 SuperCluster** - a new generation AI solution at the data center level. This system is advertised to be able to achieve **1 ZettaFLOPS FP4 performance for AI inference** and **524 ExaFLOPS FP8 for AI training, thanks to the power of hundreds of thousands of Ascend 950DT** neural processing units (APU) . If true as announced, this will be one of the most powerful AI supercomputers on the planet, capable of challenging Nvidia's Rubin systems expected to launch by the end of 2026.

Terrible performance

The Atlas 950 SuperCluster consists of 64 Atlas 950 SuperPoD clusters, similar to how Nvidia designed the GB300 NVL72 or Vera Rubin NVL144 systems. In total, the cluster uses 524,288 Ascend 950DT accelerators, distributed in more than 10,240 optically connected racks.

Theoretically, the SuperCluster can provide **524 ExaFLOPS FP8** for AI training and **1 ZettaFLOPS FP4** for AI inference. However, this is still a 'theoretical peak' number and may not reflect actual performance, which depends on many factors such as software optimization, network infrastructure, and power consumption.

Notably, Huawei combines both the RoCE (Remote Direct Memory Access over Converged Ethernet) and UBoE (UnifiedBus over Ethernet) protocols that it developed itself. According to Huawei, UBoE can reduce idle latency, increase hardware reliability, and save network infrastructure compared to RoCE.

With this massive configuration, the Atlas 950 SuperCluster is designed to handle 'super-giant' AI models with hundreds of billions to tens of trillions of parameters – the next generation of LLMs (large language models) and sparse models are expected.

The price of 'brute force'

Huawei admits it can't make a single chip more powerful than Nvidia's GPU. Instead, its strategy is to use a massive number of accelerators to 'balance the power'. An Atlas 950 SuperPoD cluster contains 8,192 Ascend 950DT chips, 20 times more than the previous generation Atlas 900 A3.

Compared to the competition, Huawei claims that the Atlas 950 SuperCluster outperforms the Nvidia Vera Rubin NVL144 (1.2 ExaFLOPS FP8, 3.6 ExaFLOPS FP4). But the issue is physical size: Huawei's system takes up **64,000 square meters** (about 150 basketball courts or 9 soccer fields), while Nvidia's Rubin NVL144 only requires a few square meters. And that's just the 'core' part, not including space for power, cooling, and supporting infrastructure.

Not only that, Huawei also revealed plans to launch **the Atlas 960 SuperCluster** in Q4/2027. This system will surpass **1 million Ascend 960 NPUs** , achieve **2 ZettaFLOPS FP8** and **4 ZettaFLOPS FP4** , and continue to maintain both UBoE and RoCE in parallel.