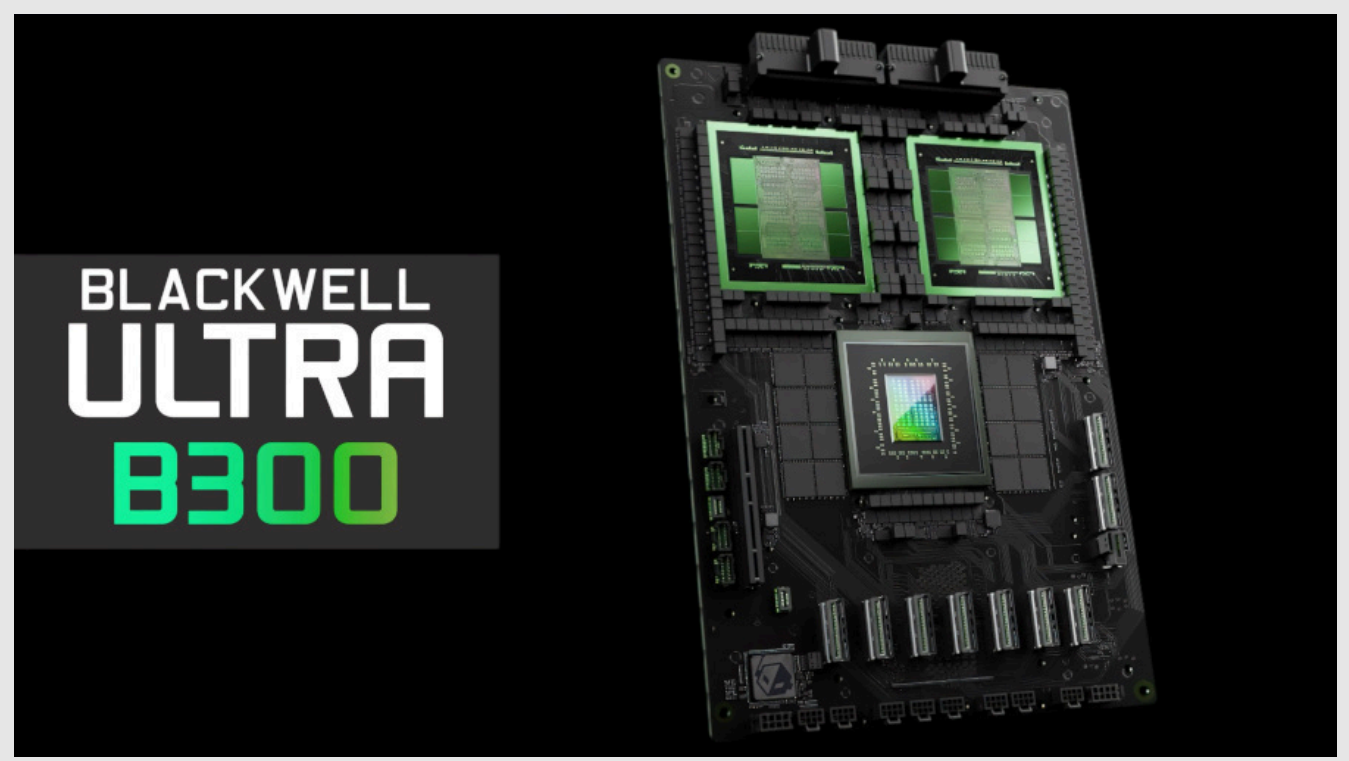
Nvidia B300或提前上市

点击数：2361

由于Nvidia的第一代Blackwell B200系列处理器在量产时出现问题不断跳票，同时还有未经证实的服务器过热问题。据外媒报道，Nvidia或将提前发布第二代Blackwell B300 GPU，性能将比B200提高50%，TDP随之增加200W。



据报道，B300系列处理器进行了大量的改进设计，但仍将使用台积电的4NP制造工艺（针对Nvidia进行优化的4nm节点），具有更高的FLOPS性能。同时在高带宽内存方面，将采用12-Hi HBM3E内存，每个GPU的将配备288GB内存，每GPU带宽仍为8TB/s。据称，B300的性能比B200提高了50%，代价是TDP也将提高200W，达到史无前例的1400W。有消息称，在B200交付后约半年时间，B300就将上市。

除了更高的性能和更大的内存外，基于B300 GPU的GB300平台将采用800G ConnectX-8 NIC网卡（收购自Mallenox），相比当前使用的400G ConnectX-7 NIC，带宽提高了一倍。并且，PCIe通道数量也从上一代的32条提升到了48条，这意味着将具有更好的横向扩展能力。

相比上一代GPU平台，B300与GB300进行了重大改进。据称，Nvidia为此重新设计了整个供应链。Nvidia将不再试图销售参考主板和整个服务器机柜，而是只销售搭载SXM Puck模块、Grace CPU和Axiada主机管理控制器（HMC）的B300。因此，这将允许更多公司参与到Blackwell B300的供应链中，同时也将Nvidia从不擅长的工作（比如设计服务器）中解放出来，或许就能避免像B200服务器那样的过热问题。

Nvidia只出售零部件，对于Google、微软及AWS等超级AI玩家而言，是一个不错的好消息。据称，从今年Q3开始，许多订单都转向了Nvidia这一更新但更贵的GPU产品上。截止到上周，几乎所有的超级玩家都已决定将采购GB300。一方面原因是GB300具有更高的性能和更大的内存（内存容量显著影响AI训练效率），另一个原因则是巨头们可以自行设计服务器。

由于上市时间的挑战，以及机架、冷却和电力方面的重大变化，巨头们很难在服务器级别对GB200进行太大的更改。这导致Meta放弃了能够从Broadcom和Nvidia多源头获得NIC的希望，转而完全依赖Nvidia。而类似的还有Google，其不得不放弃内部NIC，转而与Nvidia合作。

对于巨头们而言，他们早已习惯从CPU到网络、甚至小到对螺丝钉等所有组件进行成本优化，哪怕降低一毛钱的开销，对于采购量高达百万的巨头们，也是一笔不小的成本节省。

而现在，巨头们可以单独购买GB300，然后结合AI实际对主板、冷却系统等进行定制，从而发挥出理想的效率。虽然这需要进行大量的设计、验证和确认工作，但这对巨头们来说并不困难。比如Google、微软就聚集了一大批顶尖的硬件工程师，专门设计独有的服务器硬件，他们在这方面拥有比Nvidia多得多的经验。

但有时候，有经验、快速交付不一定是好事。有传闻称，微软将会是AI巨头中最晚部署GB300的企业，原因就是他们的设计团队太厉害了，已经根据GB200设计出了新的符合预期的硬件系统，并且在Q4还购买了相当数量的GB200。

‹ 上一页

下一页 ›

© 2026 VxWorks Club