Ingesting data for use with a large language model for AI:

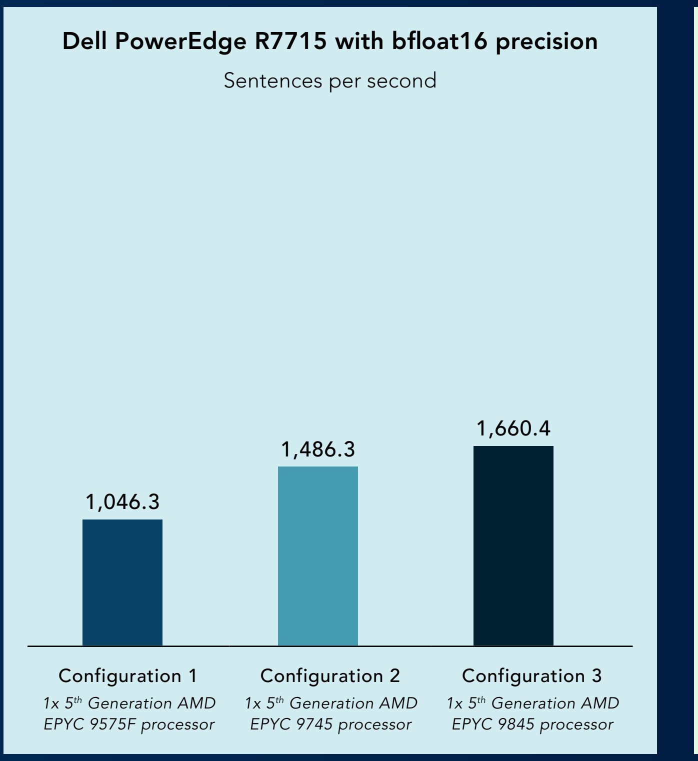
Latest-generation Dell™ PowerEdge™ servers powered by 5th Generation AMD EPYC™ processors offer a range of strong options

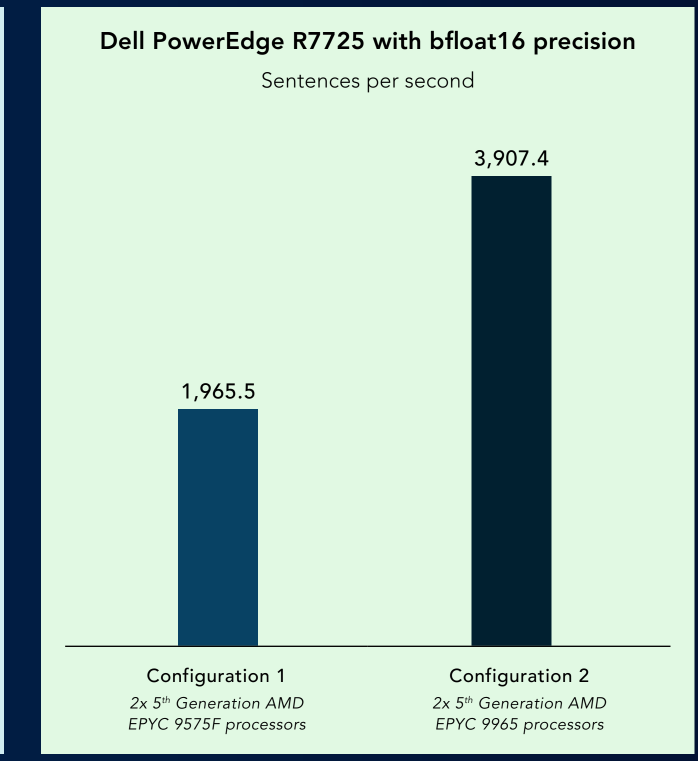
We measured the performance of multiple disaggregated infrastructure server configurations to help decision-makers choose the right one for their needs.

bfloat16 precision

Up to 1,660 sentences per second on a Dell PowerEdge R7715

Up to 3,907 sentences per second on a Dell PowerEdge R7725

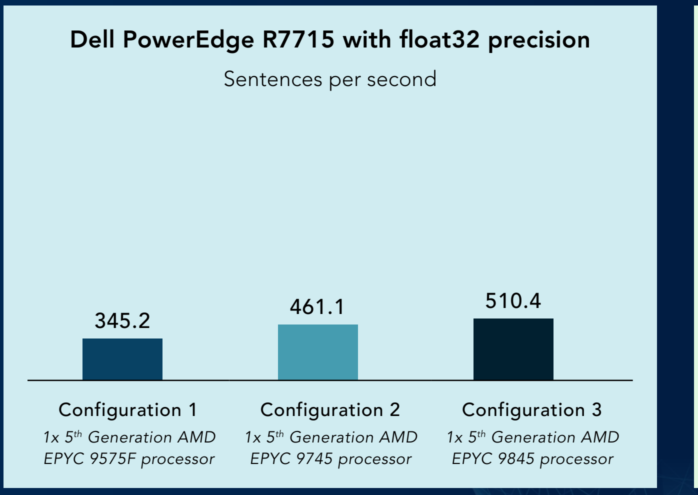

Dell PowerEdge R7715 with bfloat16 precision

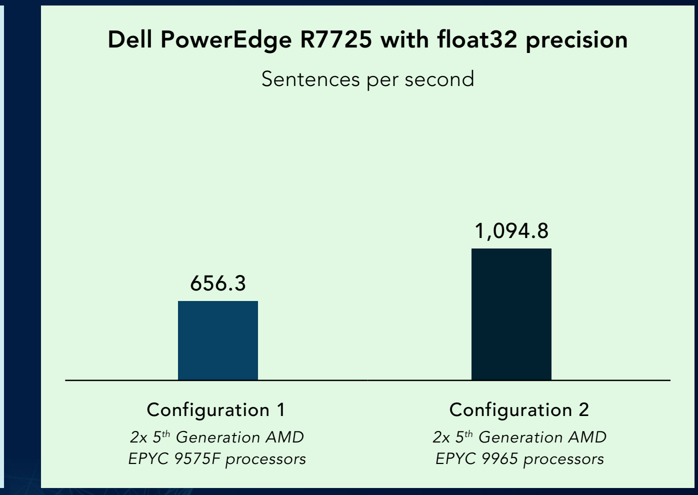

Dell PowerEdge R7725 with bfloat16 precision

float32 precision

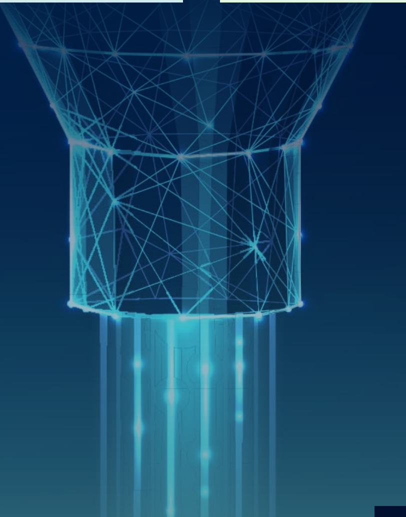
Up to 510 sentences per second on a Dell PowerEdge R7715

Up to 1,094 sentences per second on a Dell PowerEdge R7725


Dell PowerEdge R7715 with float32 precision


Dell PowerEdge R7725 with float32 precision

Configurations leveraging bfloat16 precision significantly boosted sentence processing rates, highlighting their suitability for demanding AI applications. A disaggregated architecture using these servers can help you independently scale compute and storage resources, optimizing efficiency and cost-effectiveness.



By carefully selecting the appropriate server model and processor configuration for your workload, you can achieve a balanced solution that accelerates AI ingestion while avoiding unnecessary overprovisioning, enabling faster deployment and expansion of internal AI platforms.

Read the report ▸

Principled Technologies®