Test report: Strong performance for AI image classification workloads on Stratus ztC Endurance 7100 compute platforms

Artificial intelligence (AI) for image recognition and classification is growing fast, with one study estimating a 2024 market share of USD 2.55 billion with expectations of growth to USD 4.44 billion over the next five years.[1] With a wide range of use cases that includes everything from expediting medical diagnoses to enabling quality assurance for manufacturing and making online shopping suggestions, using AI to classify images can help organizations analyze visual data quickly to deliver the timely answers they need. For businesses seeking to improve key processes with AI-based image classification, one of the first steps is selecting server hardware that can adequately handle this computationally demanding work.

To determine its suitability for running AI inference workloads such as image classification, Principled Technologies tested a Stratus ztC Endurance™ 7100 server using a ResNet-50 image classification workload at various levels of precision. Across all three precision levels, we found that the Stratus ztC Endurance 7100 offered strong throughput and low latency for CPU-based inference, showing that it's a viable platform for AI image classification for various use cases—from those that prioritize accuracy to those that prioritize speed.

High throughput
up to 5,043 images per second at INT8 precision level*

Low latency
as low as 5.8 milliseconds at INT8 precision level*

Fault tolerant
99.99999% uptime according to Stratus[2]

**in PT hands-on testing*

Running AI inference workloads on the Stratus ztC Endurance 7100

There are countless critical real-world applications for classifying images using machine learning. AI image classification can help speed up quality assurance in industrial manufacturing by providing an automated method to prevent subpar products from making it to market. Accelerating this portion of the manufacturing process can ultimately get products on the shelf faster to meet customer demands. Quick inference for image classification can also lead to quick suggestions for customers shopping at online retailers, recommending other products based on the items they've shown interest in. No matter your specific image classification use case, choosing servers with strong throughput can help you get answers from your datasets more quickly.

ResNet-50 is a convolutional neural network that runs 50 layers deep to quickly perform image classification. Using ResNet-50 models from the TensorFlow framework as well as Intel Reference Models, we ran image classification performance tests at three different precision levels: FP32, bfloat16, and INT8. The benchmark reports throughput in the number of images per second that the system could classify, as well as latency (wait times) during analysis.

Figure 1 shows the throughput that the Stratus ztC Endurance 7100 achieved while running the ResNet-50 image classification workload at three precision levels using TensorFlow's and Intel Reference Models' default suggested batch sizes. These throughput numbers indicate that the Stratus ztC Endurance 7100 is a platform capable of supporting AI inference work for image classification workloads. See page 4 to learn more about precision levels and batch sizes.

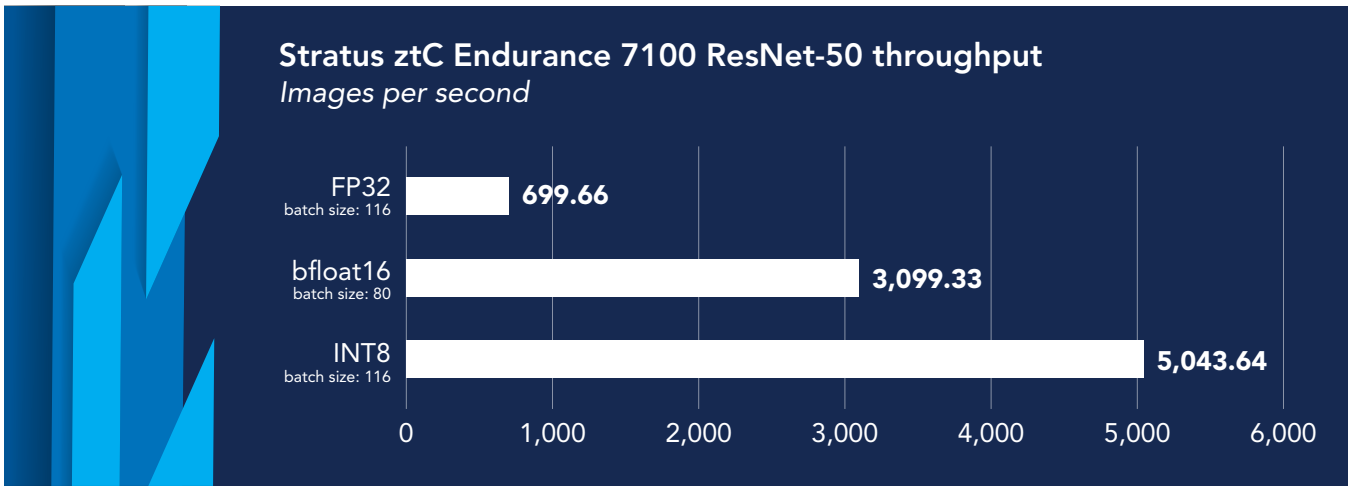

Figure 1: Throughput, in images per second, that the Stratus ztC Endurance 7100 achieved on ResNet-50 tests at three different precision levels with default batch sizes. Source: Principled Technologies.

Figure 2 shows the latency that the Stratus ztC Endurance 7100 achieved while running the ResNet-50 image classification workload at three precision levels, with a batch size of 1. Lower latencies mean quicker image classification, so as expected, the most precise model had the highest latency. Across the three precision levels, the Stratus ztC Endurance 7100 server had acceptable latencies that show its suitability for running AI inference workloads for image classification.

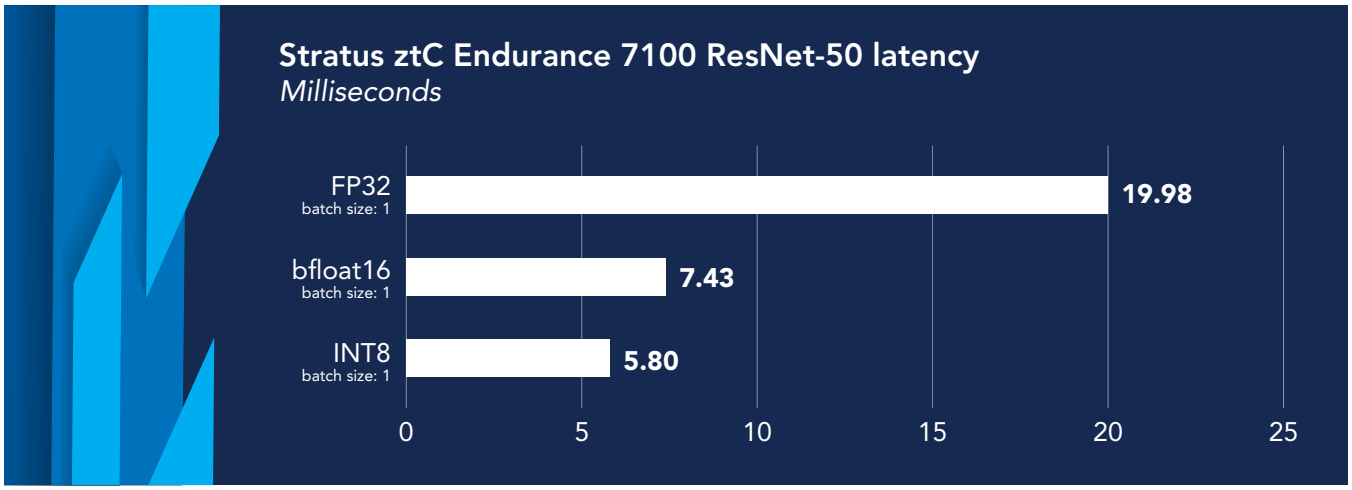

Figure 2: Latency, in milliseconds, that the Stratus ztC Endurance 7100 achieved on ResNet-50 tests at three different precision levels with a batch size of 1. Source: Principled Technologies.

About the Stratus ztC Endurance 7100 server

The Stratus ztC Endurance platform delivers intelligent, predictive fault tolerant (FT) computing for next generation, sustainable operations. The ztC Endurance family offers three configurations to provide varying levels of processing power and memory to meet a variety of data center and edge computing needs. We tested the top-of-the-line Stratus ztC Endurance 7100 platform, which Stratus markets toward those who need "high performance for high volume data- or transaction-intensive applications in large remote plants or corporate data centers."[3]

The Stratus ztC Endurance 7100 features two 4th Generation Intel® Xeon® Gold processors with 48 total cores, up to 1,024 GB DDR5 RDIMMs, and up to 38.4 TB native NVMe® storage. The platform also boasts a fully redundant hardware architecture, with customer-replaceable unit (CRU) modules that enhance serviceability and availability. (The configuration we tested differed from the max configuration; see the science behind the report for more details.)

A cornerstone of the ztC Endurance platform is its reliability, with "built-in computing fault tolerance [that] delivers 99.99999% availability to run critical applications" thanks to an "Automated Uptime Layer with Smart Exchange [that] provides continual proactive health monitoring and automatically takes action to maintain system availability and protect against data loss when needed."[4]

To learn more about the Stratus ztC Endurance 7100 server, visit https://www.stratus.com/solutions/platforms/ztc-endurance/.

What do precision levels mean?

When you're running AI workloads, you can choose the level of accuracy, or precision, that you need the computer to return, allowing you to prioritize the speed of results, image classification accuracy, and/or resource utilization.

As a Principled Technologies engineer put it in the AIXPRT benchmark blog, "Higher levels of precision for inference tasks help decrease the number of false positives and false negatives, but they can increase the amount of time, memory bandwidth, and computational power necessary to achieve accurate results. Lower levels of precision typically (but not always) enable the model to process inputs more quickly while using less memory and processing power, but they can allow a degree of inaccuracy that is unacceptable for certain real-world applications."[5]

We tested with three different precisions:

- FP32, or single-precision (32-bit) floating point format, offers a high degree of mathematical precision. This is the highest precision format we used in testing, and would provide the level of precision required for use cases such as die inspection on a semiconductor line or medical imaging.
- Bfloat16, or half-precision (16-bit) brain floating point format, uses half the number of bits as FP32 to represent a model's parameters. It "decreases time to convergence without losing accuracy" and offers the same range as FP32 but uses half of the memory space.[6] Use cases for this precision level might include weather forecasting and climate modeling.
- INT8 is the 8-bit integer data type that has a lower precision level than FP32. INT8 precision can significantly improve latency and throughput, but this increase in speed is often (but not always) at the cost of accuracy. Use cases that might use this level of precision include identification of valve settings in a manufacturing plant or cameras recognizing vehicle license plates to match security records.

Adjusting batch sizes for different windows into image classification results

For ResNet-50, batch size refers to the number of images you want the framework to process simultaneously. Smaller batch sizes tend to deliver lower latency, meaning a batch size of 1 "can be a good indicator of how a system handles near-real-time inference demands from client devices."[7] This is why we used a batch size of 1 for measuring latency on the Stratus ztC Endurance 7100.

For our maximum throughput tests, we used larger batch sizes—the model frameworks' default batch sizes of 116 or 80—because increasing the number of simultaneous images to classify shows another aspect of what a system is capable of when it comes to running inference work.

Conclusion

The use of AI to classify bulk image data is climbing across many types of organizations—from healthcare to transportation to retail—all of which can benefit from strong servers that can derive answers from image data quickly. Our test results show that the Stratus ztC Endurance 7100 offers a platform suitable for running these kinds of AI inference workloads, offering strong throughput and low latency at multiple precision levels running ResNet-50 image classification. This means that whether your organization's work requires extreme precision (such as in medical imaging) or values speed (such as in online shopping suggestions), the Stratus ztC Endurance 7100 is a strong platform that can meet your AI image classification needs.

1. Mordor Intelligence, "AI Image Recognition Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029).
2. Stratus, "Stratus ztC Endurance Datasheet," accessed August 13, 2024, https://resource.stratus.com/datasheet/stratus-ztc-endurance/?_gl=1*rh5pn0*_gcl_au*MjA4NzE4NjY1MC4xNzIzODEzNjk1.
3. Stratus, "Stratus ztC Endurance Datasheet."
4. Stratus, "Stratus ztC Endurance Datasheet."
5. Justin Greene, "Understanding the basics of AIXPRT precision settings," accessed August 14, 2024, https://www.principledtechnologies.com/benchmarkxprt/blog/2019/09/05/understanding-the-basics-of-aixprt-precision-settings/.
6. Google, "Improve your model's performance with bfloat16," accessed August 16, 2024, https://cloud.google.com/tpu/docs/bfloat16.
7. Justin Greene, "Understanding AIXPRT batch size," accessed August 14, 2024, https://www.principledtechnologies.com/benchmarkxprt/blog/2019/08/08/understanding-aixprt-batch-size/.

Read the science behind this report at https://facts.pt/FaTo3uL ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Penguin Solutions.