Upgrade to Google Cloud N4 instances featuring 5th Gen Intel Xeon Scalable processors and double Java server-side performance

Compared to older N1 instances with 1st Gen Intel Xeon processors, N4 instances with 5th Gen Intel Xeon Scalable processors handled more simultaneous Java work while maintaining acceptable response times, delivering twice the value

Whether your company is using Java to build enterprise-level apps, securely and reliably handle transactions at a high volume, or develop more efficient and scalable machine learning (ML) models, deploying the right Google Cloud™ instances can accelerate your workloads. Stronger Java performance can translate to improved user experiences, needing fewer instances, and an overall increase in the productivity and profitability of Java workloads.

In a study commissioned by Intel®, we used an industry-standard Java benchmark to test two types of Google Cloud instances: N4 standard instances featuring 5th Gen Intel Xeon® Scalable processors and older N1 standard instances with Intel 1st Gen Intel Xeon processors. The benchmark mirrors an international supermarket organization that handles a mix of work, including point-of-sale requests, online purchases, and data-mining operations.[1]

In our tests, the N4 instances achieved up to twice the performance of older N1 instances. These results indicate they could support more simultaneous operations on Java apps while maintaining acceptable response times. This performance improvement also meant that they delivered a better value, delivering up to twice the performance per dollar of N1 instances. Read on to learn more about our testing and what these results might mean for your organization.

Up to 2.0x the performance

on 16 vCPU and 64 vCPU instances

Up to 2.0x the performance per dollar

on 16 vCPU and 64 vCPU instances

This project was commissioned by Intel.

How we tested

We ran two types of Google Cloud instances in the us-east1 (South Carolina) region:

- N1 standard instances featuring 1st Gen Intel Xeon processors
- N4 standard instances featuring 5th Gen Intel Xeon Scalable processors

We tuned for best server-side Java throughput under response time, and we include both those results and server-side Java maximum transaction throughput results. The former results offer insight into the performance gains your business might expect from N4 standard instances in situations where there are latency constraints. The latter results are for situations where there are no latency constraints.

Because different organizations require different workloads, and those workloads may fluctuate in size, we ran tests on smaller, medium-size, and larger instances. We also calculated the performance per dollar of each instance. We based this data on the results of our testing and instance pricing we researched on May 20, 2024.

To learn more about our tests, configurations, calculations, and results, see the science behind the report.

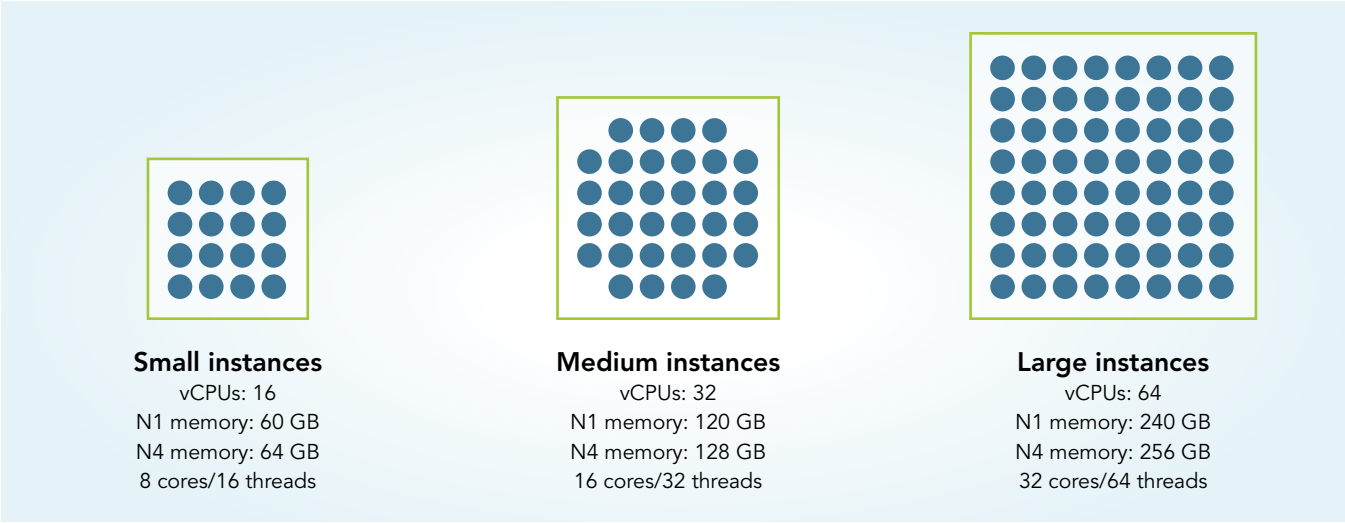

Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

What better server-side Java performance might mean for you

The benchmark provides two types of performance metrics:

Server-side Java throughput under response time: This metric reports the average maximum throughput, with response time constraints, of the system under test (SUT).[2] These results are pertinent to any company that has high-performance service level agreements (SLAs) in place. Failure to adhere to these high-performance standards may result in fines or other severe repercussions.

Server-side Java maximum transaction throughput:* The second metric reports the average maximum throughput, without response time constraints, of the SUT.[3] These results are pertinent to developers building enterprise-level apps, companies that handle transactions at a high volume, or data scientists building and deploying complex ML models to satisfy the specific requirements of multiple stakeholders.

**Please note we did not specifically tune for best maximum transaction throughput performance.*

Increase productivity and profitability with N4 instances

The Google Cloud Platform instances you choose can have a cascading effect on productivity and profitability. In Figures 1 and 2, we show the stark contrast between operations per second (OPS) N4 standard instances handled compared to N1 standard instances. It's important to note we tuned both instances for best response-constrained performance (Figure 1). These results are especially important for service providers that deal with surges during peak hours or seasons and companies with SLAs. However, non-restrained maximum transaction performance can also have a direct bearing on the amount of data you can process at one time (Figure 2). For full configuration and result details, read the science behind the report.

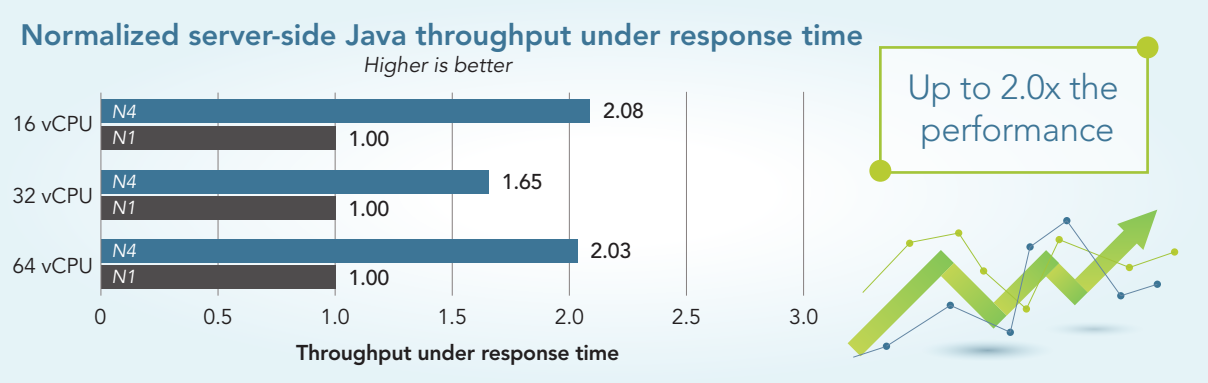

Figure 2: Relative server-side Java throughput under response time that N4 standard instances handled compared to N1 standard instances. Higher is better. Source: Principled Technologies.

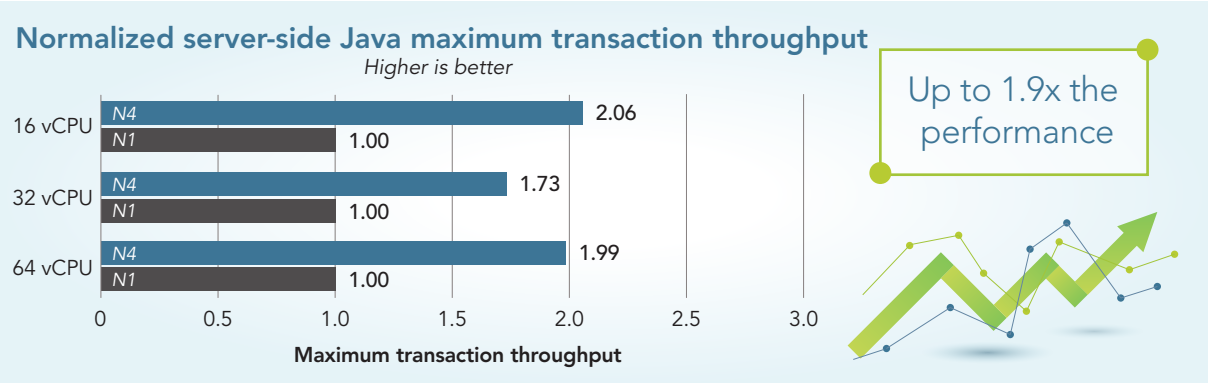

Figure 3: Relative server-side Java maximum transaction throughput that N4 standard instances handled compared to N1 standard instances. Higher is better. Source: Principled Technologies.

About Google Cloud N4 instances

Available in configurations ranging from 2 vCPUs to 80 vCPUs, N4 standard, high-CPU, and high-memory instances are "built from the ground up for flexibility and cost optimization through an efficient architecture of streamlined features, shapes, and next generation dynamic resource management, which makes better use of resources on host machines," according to Google Cloud.[4]

For more information, visit https://cloud.google.com/compute/docs/general-purpose-machines#n4_series.

Maximize value with N4 instances

Prioritizing customer satisfaction and minimizing the chance of costly SLA contract breaches are important for the success of your business. Still, there is also the bottom line to consider. In Figures 3 and 4, we show that the performance-to-cost ratios on N4 standard instances were significantly higher (better) than N1 standard instances—with up to twice the tuned Java throughput under response time performance per dollar on 16 vCPU and 64 vCPUs. For full configuration and pricing details, read the science behind the report.

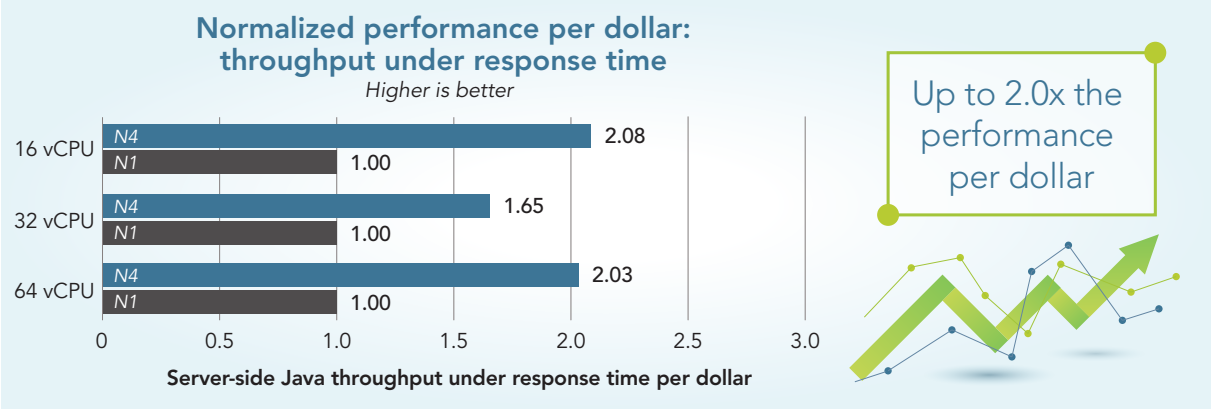

Figure 4: Relative server-side Java throughput under response time per VM cost of N4 standard instances compared to N1 standard instances. Higher is better. Source: Principled Technologies.

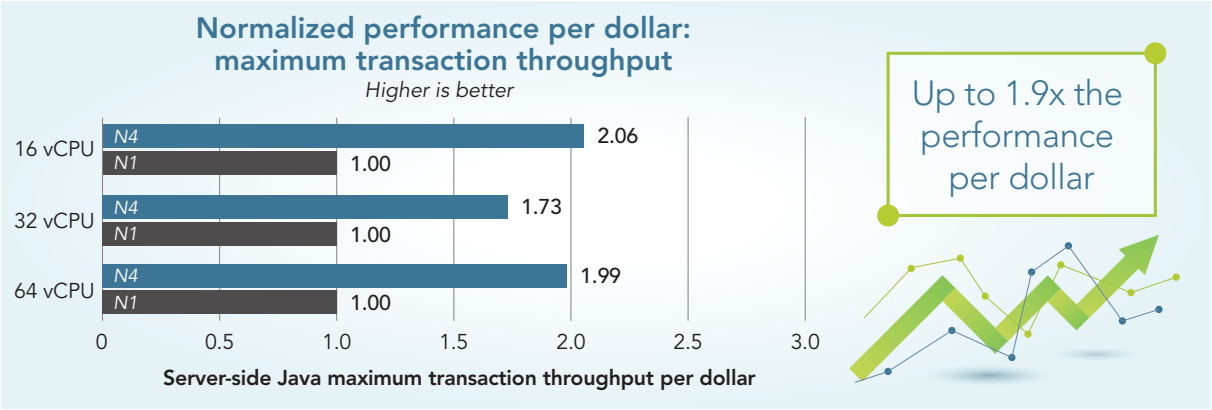

Figure 5: Relative server-side Java maximum transaction throughput per VM cost of N4 standard instances compared to N1 standard instances. Higher is better. Source: Principled Technologies.

About 5th Gen Intel Xeon Scalable processors

According to Intel, 5th Gen Intel Xeon Scalable processors deliver "more compute and faster memory at the same power envelope as our previous generation, plus outsized performance and TCO in AI, HPC, database, networking, and storage."[5] Along with PCIe Gen5 technology, DDR5 memory, and enhanced compute capabilities, 5th Gen Intel Xeon Scalable processors also offer Intel Accelerator Engines for a variety of workloads, including artificial intelligence (AI).[6]

For more information, visit https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable.html.

Conclusion

For Java workloads, the Google Cloud instance type you choose can affect productivity and profitability. Our Java server test results indicate that N4 instances enabled by 5th Gen Intel Xeon Scalable processors delivered up to twice Java throughput under response time at 16, 32, and 64 vCPU counts versus N1 instances with 1st Gen Intel Xeon processors. These performance gains lead to N4 instances delivering a better value, with up to double the performance per dollar on 16 and 64 vCPU instances. With an instance that can process more OPS at a better value, your company could provide a more satisfying experience for users, reduce Java workloads onto fewer instances, and grow your business.

1. SPEC, "The SPECjbb2015 benchmark," accessed June 7, 2024, https://www.spec.org/jbb2015/.
2. Standard Performance Evaluation Corporation (SPEC), "SPECjbb2015 Benchmark Run and Reporting Rules 1.02," accessed June 6, 2024, https://www.spec.org/jbb2015/docs/runrules.pdf.
3. Standard Performance Evaluation Corporation (SPEC), "SPECjbb2015 Benchmark Run and Reporting Rules 1.02."
4. Google Cloud, "General-purpose machine family for Compute Engine," accessed June 4, 2024, https://cloud.google.com/compute/docs/general-purpose-machines.
5. Intel, "5th Gen Intel® Xeon® Scalable processors," accessed June 12, 2024, https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable.html.
6. Intel, "5th Gen Intel® Xeon® Scalable processors."

Read the science behind this report at https://facts.pt/s9fbSlQ ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.