Support more Kubernetes web app users with newer Amazon Web Services instances featuring 2nd Generation Intel Xeon Scalable Processors

Newer R5n series instances featuring Cascade Lake processors handled more users than older R4 series instances that used Broadwell processors

Your customers and employees rely on having consistent access to your website with predictable application performance—so if your company has decided to host its website on the cloud, you'll need cloud instances that can keep up with the pace of your business.

At Principled Technologies, we tested Weathervane 2.0, a multi-tiered web app workload for Kubernetes, on two general-purpose instance series for Amazon Elastic Cloud Compute (Amazon EC2): newer R5n series instances that featured 2nd Generation Intel Xeon Scalable processors (or, Cascade Lake), and older R4 series instances that featured processors from the Intel Xeon E5 v4 family (also known as Broadwell). We found that the R5n series instances handled up to 1.85 times as many Weathervane 2.0 users as the R4 series instances. Because R5n series instances cost just 1.10 times as much as R4 series instances, we determined that the Cascade Lake processor-enabled instances are a better value for businesses.

Handle up to 1.85x the Weathervane users

with small R5n series instances*

Handle up to 1.49x the Weathervane users

with large R5n series instances

*compared to older R4 series instances

How we tested

The instances we used

We purchased instances from two general-purpose Amazon EC2 series:

- Newer R5n series instances featuring Cascade Lake processors
- Older R4 series instances featuring Broadwell processors

We ran each instance in the US East 1 region. To create three-node Kubernetes clusters, we used the EKS Kubernetes-as-a-Service offering from Amazon.

Kubernetes is a platform for deploying and managing containerized applications. Application containers are packaged software units that encapsulate the entire runtime environment, including the application code, libraries, binaries, configuration files, and dependencies. One benefit of containers is their lightweight nature: a single server can often run many more containers than entire virtual machines.

The instance sizes we chose

Figure 1 shows the specifications for the instances we used with each cluster. To represent a couple of common general use cases, we tested two sizes from the two series: small instances with 8 vCPUs per node and large instances with 32 vCPUs per node.

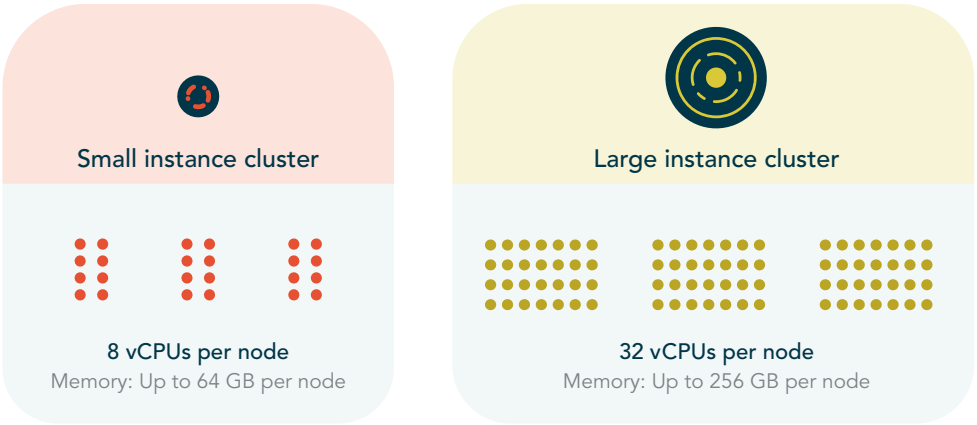


Figure 1: Key specifications for each instance. Source: Principled Technologies.

The Kubernetes workload we used

We tested each instance with the application-level Kubernetes benchmark Weathervane 2.0. According to the VMware website, Weathervane "measures the performance capabilities of a Kubernetes cluster by deploying one or more instances of a benchmark application on the cluster and then driving a load against those applications."[1]

The benchmark application in question is a real-time auction web app, where simulated users view items and make bids within a set time frame. The Weathervane benchmark test outputs results in terms of WvUsers. The WvUser metric "represents the maximum number of simulated users that could interact with the application instances without violating the QoS requirements." According to VMware, when using Weathervane to compare two Kubernetes clusters, the one that achieves a higher WvUsers score is considered the higher-performing cluster.[2]

How R5n series instances can help your business

Compared to the older R4 instances, R5n instances offer:

- Higher core frequency (3.1 GHz vs. 2.3 GHz)
- Intel Vector Neural Network Instructions support (AVX-512 VNNI)
- Higher network bandwidth at most sizes

Hands-on testing results

How large is your business? Is your main website internal or external-facing? We tested two sizes of instances to represent a variety of business needs. Being able to support a larger number of users means your business could better handle its current user base and potentially have room to grow in the future.

Small instances

A business seeking cloud-based instances to support a smaller environment—or one with a fairly consistent concurrent user count—may be content to use smaller instances with around eight vCPUs. We found that the cluster with newer R5n series instances powered by Intel Xeon Cascade Lake processors supported 1.85 times the number of Weathervane 2.0 users as the cluster with older instances powered by Intel Xeon Broadwell processors.

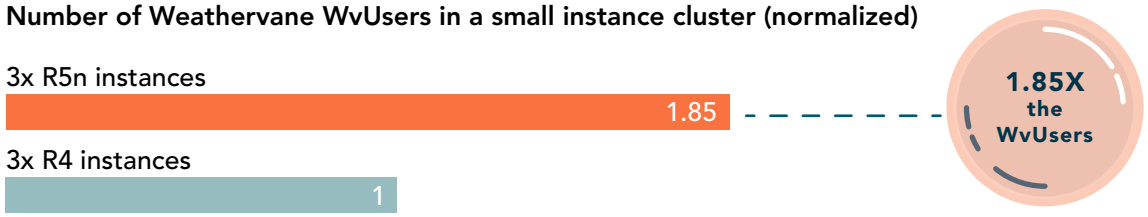


Figure 2: Relative number of Weathervane 2.0 users each cluster supported, normalized to the score achieved by the cluster with R4 series instances. More users is better. Source: Principled Technologies.

Large instances

Instances with around 32 vCPUs may be right for businesses that need to support a substantially larger volume of daily users, such as those with popular external-facing websites and lots of customers. Figure 3 shows that the cluster with newer R5n series instances supported 1.49 times the number of Weathervane 2.0 Kubernetes web server users as the cluster with older R4 series instances.

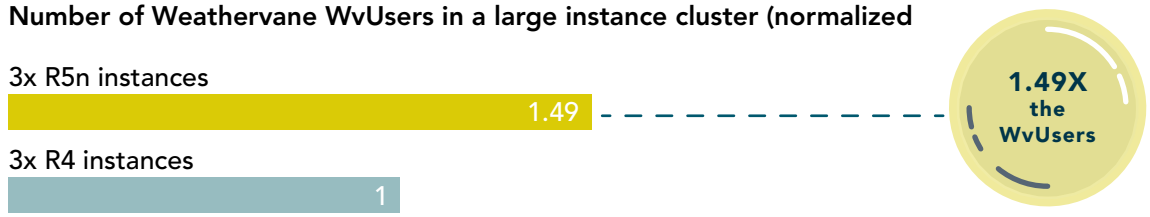
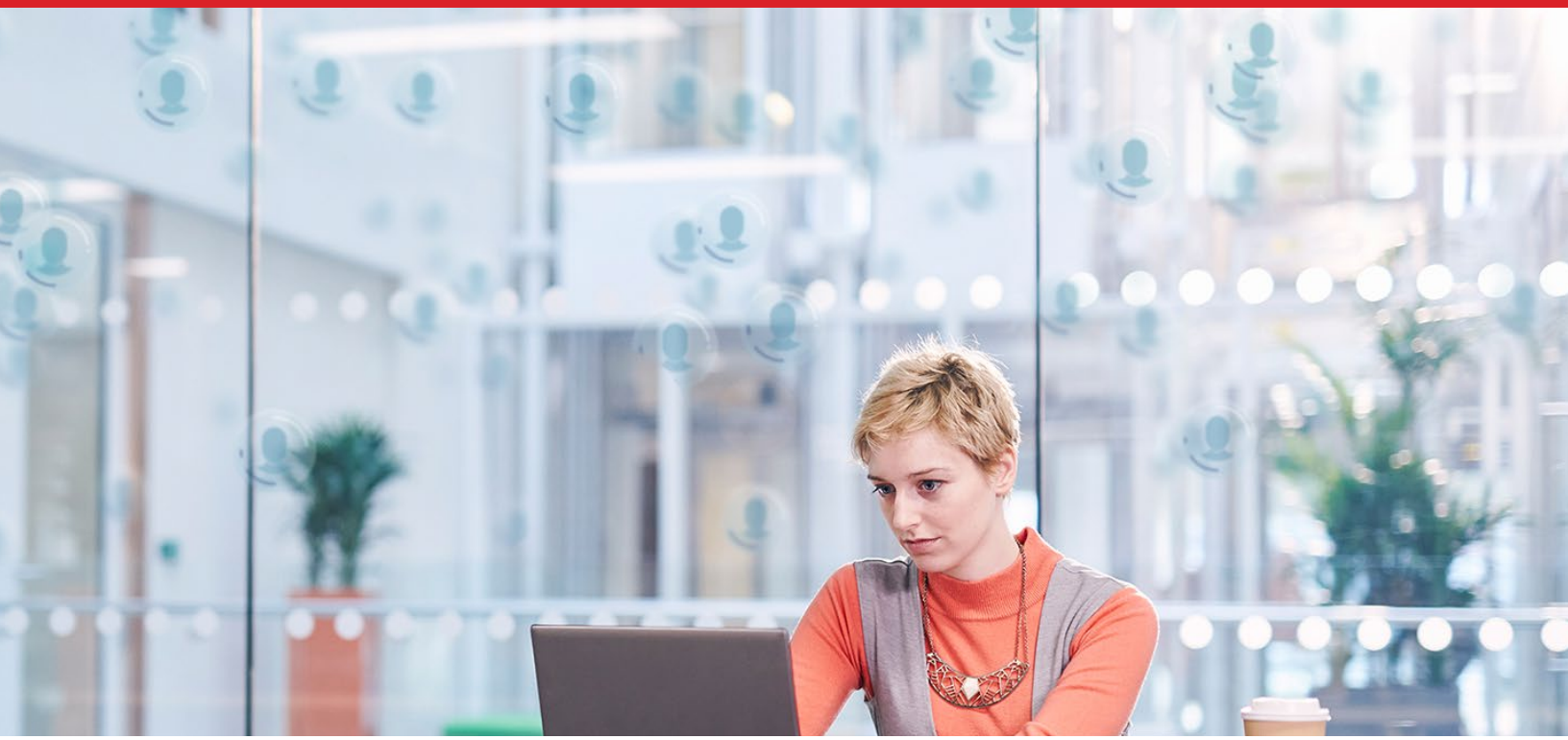


Figure 3: Relative number of Weathervane 2.0 users each cluster supported, normalized to the score achieved by the cluster with R4 series instances. More users is better. Source: Principled Technologies.

How much will it cost us?

IT decision-makers must weigh the benefits of newer, better-performing technology against the cost of said technology. Based on results of our hands-on testing, there's little question: R5n instances cost just 1.10 times as much as R4 instances despite delivering between 1.49 and 1.85 times the number of Weathervane users, making them a better value overall.[3]



Conclusion

In our Kubernetes web server tests on Amazon EC2 clustered cloud instances, we found that clusters of newer R5n series instances featuring Intel Xeon Cascade Lake processors had stronger Kubernetes performance than clusters of older R4 series instances that featured Intel Xeon Broadwell processors. The performance advantage ranged from 1.49 to 1.85 times that of the cluster of R4 series instances. Additionally, because there is only a 1.10x price difference between the two series of instances, the R5n series is the more cost effective for Kubernetes web server workloads.

Using R5n instances enabled by Cascade Lake processors could mean supporting more users—or ensuring a better experience for the ones you already have.

---

1 Harold Rosenberg, "Weathervane 2.0: An Application-Level Performance Benchmark for Kubernetes," accessed April 5, 2021, https://blogs.vmware.com/performance/2020/02/weathervane2-kubernetes.html

2 Harold Rosenberg, "Weathervane 2.0: An Application-Level Performance Benchmark for Kubernetes."

3 "Amazon EC2 On-Demand Pricing," accessed April 5, 2021, https://aws.amazon.com/ec2/pricing/on-demand/

Read the science behind this report at http://facts.pt/x61FwGh ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.