

Microsoft® Azure® Ddv4 Virtual Machines Cut the Time and Cost of Complete Genomics Tasks Almost in Half

Genomics

Azure Ddv4 VM Clusters Completed a Set of Genomics Tasks in as Little as 52% of the Time of Other Clusters

Azure Ddv4 VM Clusters Completed a Set of Genomics Tasks at as Little as 52% of the Cost of Other Clusters

VMs Featuring 2nd Gen Intel® Xeon® Scalable Processors Executed a Set of Genomics Analytics Toolkit Tasks in as Little as 52% of the Time and as Little as 52% the Cost of VMs with Previous-Gen Processors

If your company is considering running your genomics workloads in the public cloud, keep in mind that Virtual Machines (VMs) can vary widely in terms of both their performance and costs. To benchmark the time and cost benefits, Intel tested Microsoft Azure VM clusters from three categories:

- **Ddv4 series VMs**, featuring exclusively 2nd Gen Intel Xeon Scalable processors
- **Dv2 series VMs**, with CPUs ranging from fourth-generation Intel Core™ i-series processors to 2nd Gen Intel Xeon Scalable processors
- **Default configuration VMs** from the following series: A, Av2, Dv2, Dv3, Ls, Fsv2, with CPUs ranging from second-generation Intel Core processors to 2nd Gen Intel Xeon Scalable processors

Testing used the Cromwell on Azure benchmark to measure performance of the Genomics Analytics Toolkit (GATK) application. The test workflow comprised 24 tasks. In this brief, we look at relative performance and costs both for the entire set of tasks and for one of the most resource-intensive tasks.

Ddv4 VM Clusters Featuring 2nd Gen Intel Xeon Scalable Processors Executed Genomics Tasks in Significantly Less Time Than Other VM Clusters

Figure 1 shows the relative time to complete genomic tasks in the GATK application. Compared to the default VMs using a range of older processors, the Ddv4 VM featuring 2nd Gen Intel Xeon Scalable processors completed the full set of tasks in just over half the time, a reduction of as much as 48%. To execute the resource-intensive HaplotypeCaller task, the Ddv4 VM needed just over one-quarter of the time the default VMs needed, a reduction of as much as 74%. These advantages can translate to carrying out your genomics analysis much more efficiently.

Relative Time to Complete Genomics Analytics Toolkit Tasks

Normalized Time | Less time is better

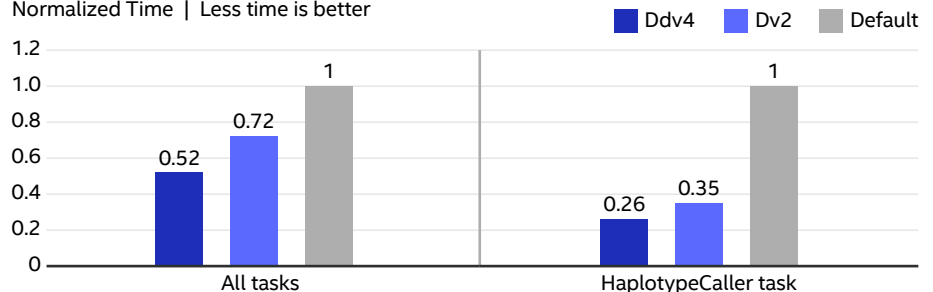


Figure 1. Relative time to complete genomics tasks. Less time is better.



Ddv4 VM Clusters Featuring 2nd Gen Intel® Xeon® Scalable Processors Executed Genomics Tasks at a Significantly Lower Cost Than Other VM Clusters

When a VM cluster can perform a set of tasks in less time, customers save by paying for less VM uptime. Figure 2 shows the relative cost to complete the same GATK tasks mentioned on the previous page. Compared to the default VMs using a range of older processors, the Ddv4 VM featuring newer processors completed the full set of tasks at slightly more than half the cost, a savings of as much as 48%. The cost of executing the resource-intensive HaplotypeCaller task on the Ddv4 VM cluster was slightly more than one-third of that for the default VM clusters, a savings of as much as 63%.

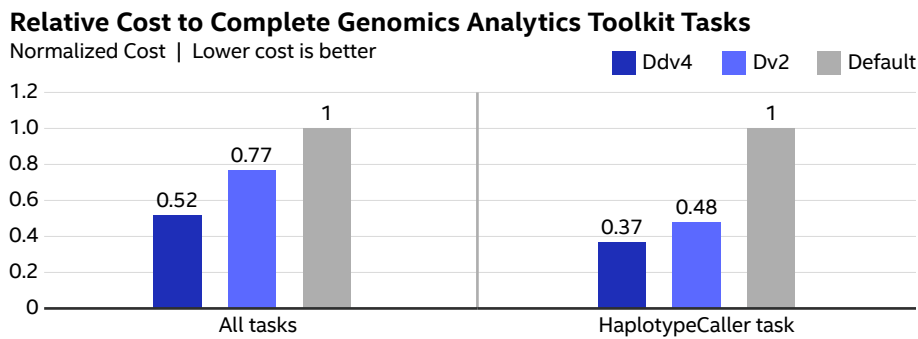


Figure 2. Relative cost to complete genomics tasks. Lower cost is better.

Conclusion

Genomics analysis applications are very compute-intensive, making it especially important to select a cloud VM with robust performance. Our testing showed that opting for Azure Ddv4 VMs featuring 2nd Gen Intel Xeon Scalable processors reduced the time to complete genomics tasks—and the cost to complete them—almost in half compared to using default VMs with older processors.

Learn More

To begin running your genomics workloads on Microsoft Azure Ddv4 virtual machines with 2nd Gen Intel Xeon Scalable processors, visit <https://docs.microsoft.com/en-us/azure/virtual-machines/ddv4-ddsv4-series>.

All tests by Intel on Azure/uswest2. All tests: Linux, Input Data Set 30X Coverage Human Whole Genome Sequence (WGS); NA12878, Workload GATK Best Practices Pipeline for Germline Variant Calling with pre-processing, GATK 4.0.10.1, Genomics Kernel Library (GKL) 0.8.6, Cromwell 52, Picard 2.20, BWA 0.7.15-r1140, Samtools 1.3.1. Tools in <https://hub.docker.com/r/broadinstitute/genomes-in-the-cloud/>; us.gcri.io/broad-gotc-prod/genomes-in-the-cloud:2.4.3-1564508330, us.gcri.io/broad-gatk/gatk:4.0.10.1. Workflow defined: <https://github.com/microsoft/gatk4-genome-processing-pipeline-azure>. Run Iterations:3. VM details: Ddv4 series: 8272CL: Standard_D16d_v4: 16 vCPUs, 64GiB RAM, 600GiB SSD; Standard_D8d_v4: 8 vCPUs, 32GiB RAM, 300GB SSD; Standard_D4d_v4: 4 vCPUs, 16GiB RAM, 150GiB SSD; Standard_D2d_v4: 2 vCPUs, 8GiB, 75GiB SSD; Dv2 series: 8272CL, 8171M, E5-2673 v4, or E5-2673 v3; Standard_D3_v2: 4 vCPUs, 14GiB RAM, 200GiB SSD; Standard_D4_v2: 8 vCPUs, 28GiB RAM, 400GiB SSD; Standard_D5_v2: 16 vCPUs, 56 GiB RAM, 800GiB SSD; Standard_D2_v2: 2 vCPUs, 7GiB RAM, 100GiB SSD; Standard_D1_v2: 1 vCPU, 3.5GiB RAM, 50GiB SSD; Default Config: E5-2660 (A); E5-2660, E5-2673 v4 (Av2); 8272CL, 8171M, E5-2673v4, E5-2673v3 (Dv2, Dv3); E5-2673 (Ls); 8168, 8272CL (Fsv2); Standard_A2: 2 vCPUs, 3.5GiB RAM, 135GiB SSD; Standard_A3: 4 vCPUs, 7GiB RAM, 285GiB SSD; Standard_A1_v2: 1 vCPU, 2GiB RAM, 10GiB SSD; Standard_D2_v3: 2 vCPUs, 8GiB RAM, 50GiB SSD; Standard_F16s_v2: 16 vCPUs, 32GiB RAM, 128GiB SSD; Standard_L4s: 4 vCPUs, 32GiB RAM, 678GiB SSD



Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See above for configuration details. No product or component can be absolutely secure. Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Printed in USA 0822/JO/PT/PDF US001

