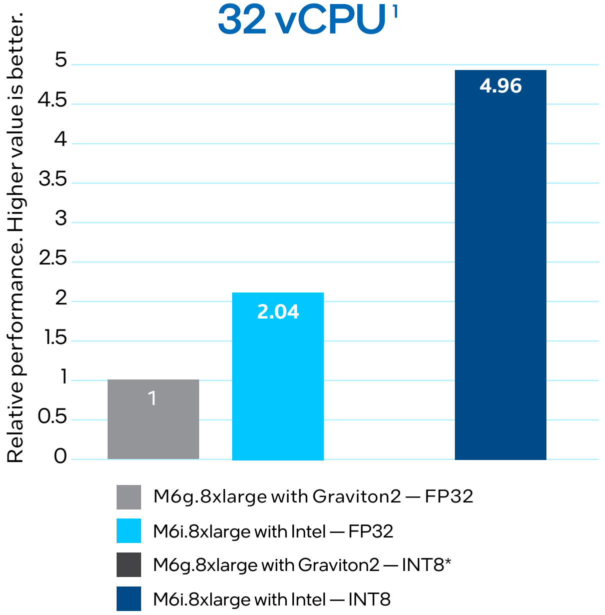
Faster Natural Language Processing with Intel

Natural language processing (NLP) delivers seamless integration with customers and propels your business forward. Take it to the next level on Intel.

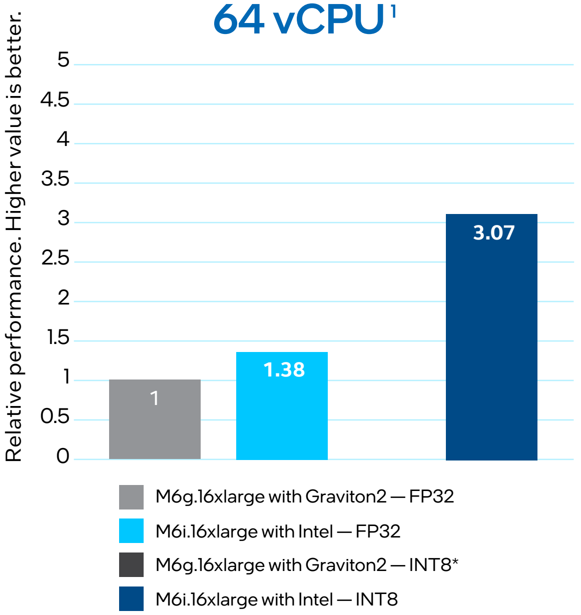
M6i with Intel Outperfoms M6g with Graviton2

BERT-Large Inference Sentences Per Second 32 vCPU[1]



* M6g INT8 support was not available when tested for Tensorflow in the BERT-large model used.

BERT-Large Inference Sentences Per Second 64 vCPU[1]



* M6g INT8 support was not available when tested for Tensorflow in the BERT-large model used.

Solve Common Problems

- Analyze user-typed text more quickly
- Process results faster for predictive text
- Deliver meaningful automated responses to human users
- Remove communication barriers

Ideal Uses

- Smart assistants
- Search
- Language translation
- Sentiment analysis

Demanding Workloads

Processing input from customers and other users puts a heavy demand on compute resources. Deliver results fast by landing NLP applications on cloud instances with Intel processors.

Cloud Performance Advantages

Up to 3.97x

better price performance on NLP in M6i with INT8 precision than in M6g with FP32 precision (32vCPU)[2]

Up to 2.46x

better price performance on NLP in M6i with INT8 precision than in M6g with FP32 precision (64vCPU)[3]

Intentional Workload Placement

Choose from **AWS EC2 M6i**, **C6i instances**, and **DL1 instances** for your NLP workloads. Discover the benefits **Intel leadership in AI technology** brings to your business.

Development Tools & Resources

Prepare, build, deploy, and scale AI solutions for your workloads. Visit **developer.intel.com/ai** for help getting the most out of your cloud.

Let's work together. Contact us.

[1] See intel.com/3gen-xeon-config [85], [86] for configuration and test information.

[2] See intel.com/3gen-xeon-config [86] for configuration and test information. AWS M6i.8xlarge instance delivers up to 3.97x better price performance and up to 4.96x better performance on BERT-Large inference work with INT8 precision than M6g.8xlarge instance with AWS Graviton2 processors with FP32 precision. Pricing information from US-West-2 region as of January 9, 2023, https://aws.amazon.com/ec2/pricing/on-demand/

[3] See intel.com/3gen-xeon-config [86] for configuration and test information. AWS M6i.16xlarge instance delivers up to 2.46x better price performance and up to 3.07x better performance on BERT-Large inference work with INT8 precision than M6g.16xlarge instance with AWS Graviton2 processors with FP32 precision. Pricing information from US-West-2 region as of January 9, 2023, https://aws.amazon.com/ec2/pricing/on-demand/

Note: One benchmark does not fit all workloads. Select meaningful measurements to inform your purchasing decisions. For more information, see intel.com/3gen-xeon-config [101] and [102] and https://www.intel.com/content/www/us/en/content-details/752333/active-benchmarking-for-better-performance-predictions.html

See our complete Legal Notices and Disclaimers.