Solution Brief

Cybersecurity

Boost Webroot BrightCloud* CSI Inference Performance on Intel® Xeon® 6 SoC

AI/ML models optimize BrightCloud* CSI for efficient web policy management and cyber threat mitigation.

BrightCloud* Cloud Services Intelligence (CSI) is at the forefront of addressing the challenges posed by digital transformation and the shift from traditional websites to complex web applications. This evolution demands a new approach to web policy management, extending control beyond access to include user actions and permissions within web applications. Traditional rules-based methods of user function identification often struggle with the increasing complexity of web applications. BrightCloud CSI now leverages advanced AI/ML models during real-time traffic analysis to identify user actions while they use interactive web applications.

Leveraging AI is a cutting-edge approach to combating advanced cyber threats and applying policy management, but it requires significant computational resources. To overcome performance challenges in real-time traffic analysis, BrightCloud collaborated with Intel to optimize its CSI inference performance using Libtorch, and Intel® Extension for PyTorch* (IPEX), which provides performance optimizations for Intel hardware. These enhancements leverage Intel® Advanced Vector Extensions 512 (Intel® AVX-512) with Vector Neural Network Instructions (VNNI) and Intel® Advanced Matrix Extensions (Intel® AMX) on Intel processors, along with Intel® Xe Matrix Extensions (Intel® XMX) AI engines on Intel discrete GPUs. Through these optimizations, the collaboration has achieved up to an 8x reduction in model inference latency on Intel Xeon 6 SoC.

These enhancements enable BrightCloud CSI to perform real-time HTTP/HTTPS traffic analysis and behavioral analytics, enforcing precise policies. By integrating BrightCloud CSI into Secure Web Gateways, Next-Gen Firewalls, and SASE, it can offer a robust, scalable, and efficient cloud security solution.

Intel® Xeon® 6 Soc and Intel Optimizations

Intel® Xeon® 6 SoC enhances Intel AMX by introducing FP16 support alongside BP16 and INT8. It also increases memory speed and expands the cache size. These advancements significantly improve AI inference performance, enabling efficient inline inference with fewer CPU cores.

To provide a seamless experience for developers, Intel has developed IPEX and upstream its features into PyTorch to fully leverage Intel AVX-512 and Intel AMX capabilities. On CPUs, it dynamically assigns operators to the most efficient kernels based on the detected instruction set, utilizing Intel's vectorization and matrix acceleration units. The runtime extension also boosts efficiency with finer-grained thread control and weight sharing.

The inference performance comparison between the 3rd Gen Intel® Xeon® D processor (Intel® Xeon® D-2899NT) and the Intel Xeon 6 SoC shows substantial performance gains. This upgrade results in a 9.03x improvement in the CSI upload model (see Figure 1) and an 8.85x improvement in the CSI download model (see Figure 2).[1] These enhancements enable BrightCloud CSI to swiftly and effectively detect and respond to threats, ensuring organizations can operate securely in the cloud. This performance boost highlights the importance of combining advanced hardware with optimized software to meet the growing demands of modern cloud security.

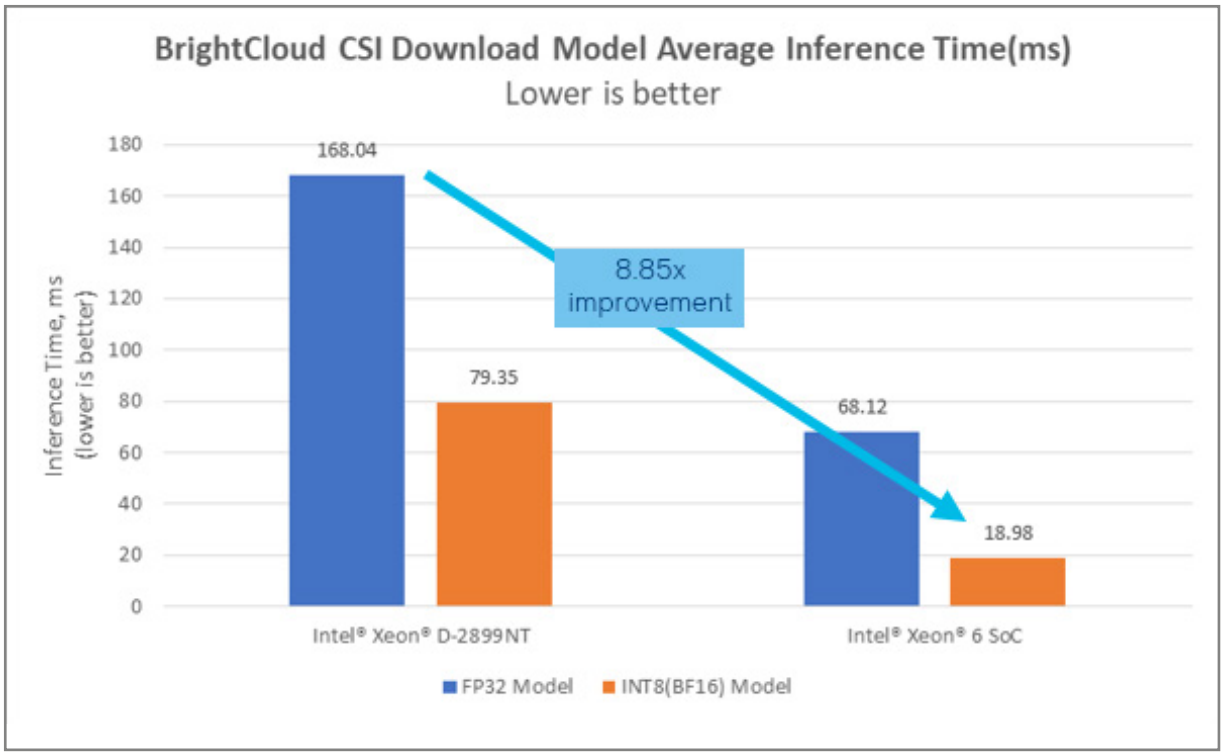

Figure 1. BrightCloud CSI Upload Model Average Inference

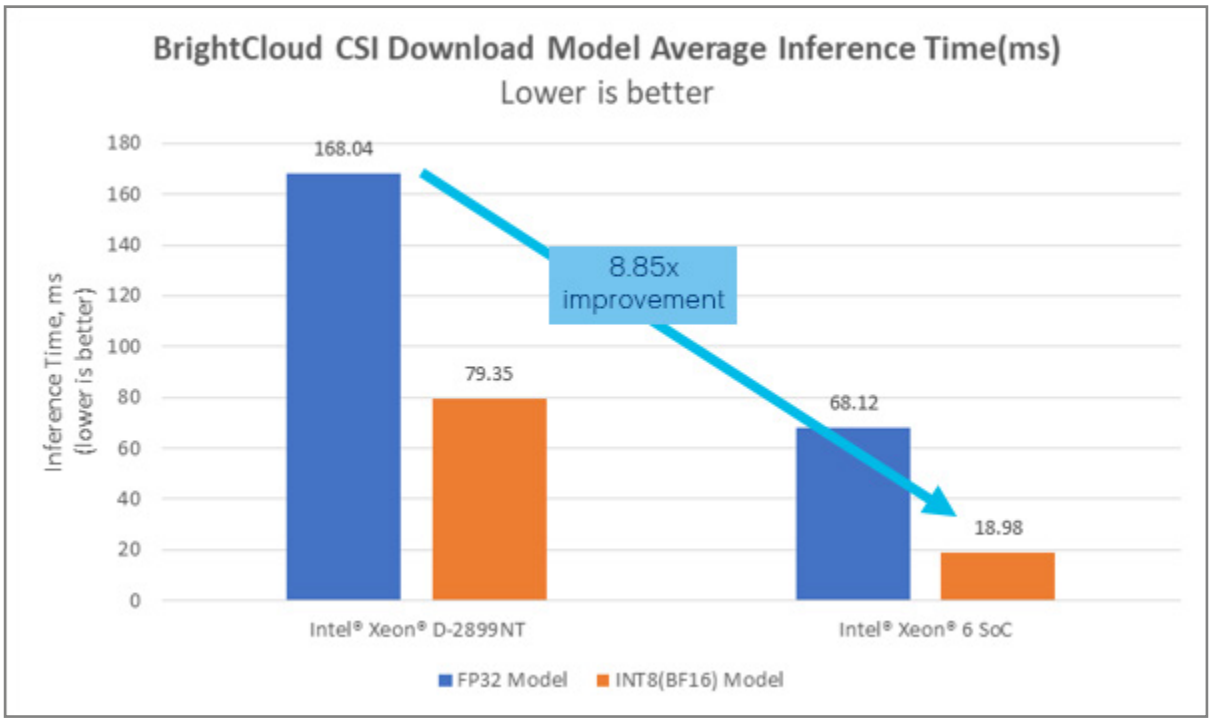

Figure 2. BrightCloud CSI Download Model Average Inference

Deploying BrightCloud CSI on an Intel Platform

Prepare the BrightCloud CSI AI running environment:

- **Install Libtorch 2.6.0 and Intel Extension for Pytorch (IPEX) 2.6.0**

Download libTorch package zip file and decompress the zip.

```
# wget https://download.pytorch.org/libtorch/cpu/libtorch-cxx11-abi-shared-with-deps-2.6.0%2Bcpu.zip
# unzip libtorch-cxx11-abi-shared-with-deps-2.6.0+cpu.zip
```

Download IPEX 2.6.0. IPEX extends PyTorch with the latest performance optimizations for Intel hardware. Placeholder <libtorch_path> represents the decompressed libTorch package path.

```
# wget https://intel-extension-for-pytorch.s3.amazonaws.com/libipex/cpu/libintel-ext-pt-cxx11-abi-2.6.0%2Bcpu.run
# bash <libintel-ext-pt-name>.run install <libtorch_path>
```

- **Install Memory Allocator (TCMalloc)**

TCMalloc introduces more efficient memory usage and reduces overhead on unnecessary memory allocations or destructions, and thus faster execution.

```
# sudo apt install libgoogle-perftools-dev
# export LD_PRELOAD=libtcmalloc.so:$LD_PRELOAD
```

- **Launch BCTI in daemon mode**

Now you may use the following commands to launch the BCTI

```
# ./bctid  -w  -m 4  -t 1 <port>
```

Conclusion

The collaboration between Webroot and Intel exemplifies a cutting-edge approach that utilizes Intel Xeon 6 SoC, delivering significant improvements in AI model inference performance and enhancing threat detection and response capabilities.

This same approach also can be applied to Intel-powered platform selections such as AWS, Google Cloud, and Azure, providing users with optimal performance and better TCO. For cloud instances, we recommend selecting Intel-based cloud instance types with Intel AMX capability to significantly increase the throughput and reduce the latency. Our recommended instances for major cloud providers include:

- **AWS:** R7i, M7i-flex or C7i-flex instance types
- **GCP:** C4, N4 or C3 instance types
- **Azure:** FXv2-Series, Dv6 series, Ev6-Series or DCesv5-Series instance types

For network security appliance designs, we suggest choosing Intel Xeon 6 SoC, 5th Gen Intel® Xeon® Scalable processor, or 4th Gen Intel Xeon Scalable processor. To further enhance performance, we recommend leveraging the Intel® Arc™ A770 Discrete Graphics.

Notices & Disclaimers

[1] Configuration Details: Intel® Xeon® D-2899NT: 1-node, 1x Intel® Xeon® D-2899NT CPU, 22 cores, 135W TDP, HT On, Turbo On, Total Memory 128GB (8x16GB DDR4 3200 MT/s [2933 MT/s]), BIOS IDVLCRB1.86B.0021.D41.2112031014, microcode 0x10002b0, 1x 223.6G KINGSTON SA400S37240G, Ubuntu 24.04.1 LTS, 6.8.0-44-generic.
Software: libtorch 2.5.1/IPEX 2.5.0, gcc 11.4, 2 Active Cores , Data Type int8, batch Size 1
Test conducted by Intel as of Mon Feb 3 07:36:04 PM MST 2025.

0725/DL/MK/PDF ♻ Please Recycle 365533-001US