ワークロードに最適化されたインテル ソリューションのご紹介

インテル株式会社 営業本部
佐藤　義和

2022年10月20日

通知および免責事項

性能は使用状況や構成などにより異なります。詳しくは、www.Intel.com/PerformanceIndex。

パフォーマンス結果は、構成に示された日付でのテストに基づいており、すべての公開されたアップデートが反映されていない場合があります。

構成の詳細については、バックアップを参照してください。

製品やコンポーネントには、絶対的な安全性を保証するものはありません。

インテルは、Principled Technologiesが運営するBenchmarkXPRT Development Communityを含むさまざまなベンチマーク・グループへの参加、スポンサー、技術サポートを通じて、ベンチマークの開発に貢献しています。

お客様の費用と結果は異なる場合があります。

インテルのテクノロジーは、ハードウェア、ソフトウェア、サービスのアクティベーションが必要な場合があります。

一部の結果は推定またはシミュレートされたものである可能性があります。

インテルは、サードパーティーのデータを管理または監査していません。

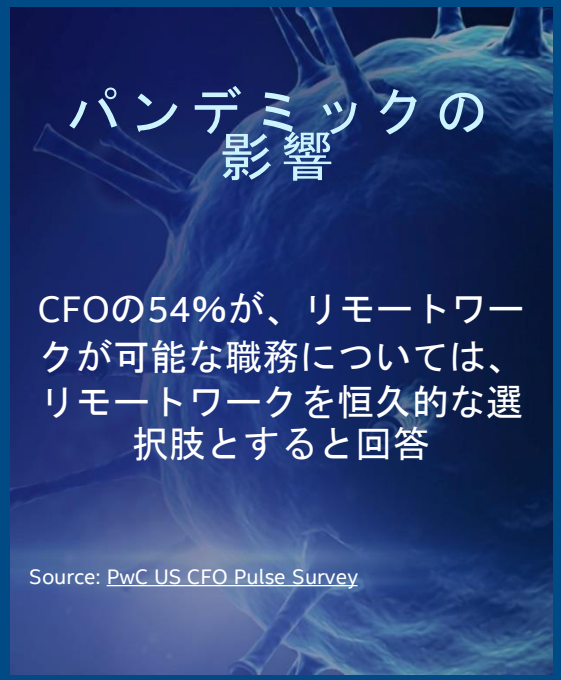
正確性を確認するには、他の情報源を参照してください。

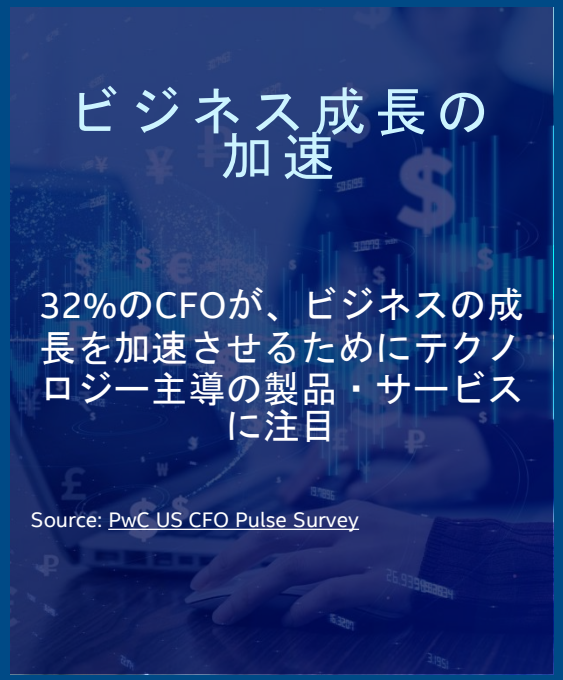
すべての製品計画とロードマップは、予告なく変更される場合があります。本資料に記載されている将来の計画や予想に関する記述は、将来の見通しに関する記述です。 これらの記述は、現時点での予測にもとづくものであり、多くのリスクや不確実性を内包するため、実際の結果はこれらの記述で明示または黙示されたものと異なる可能性があります。

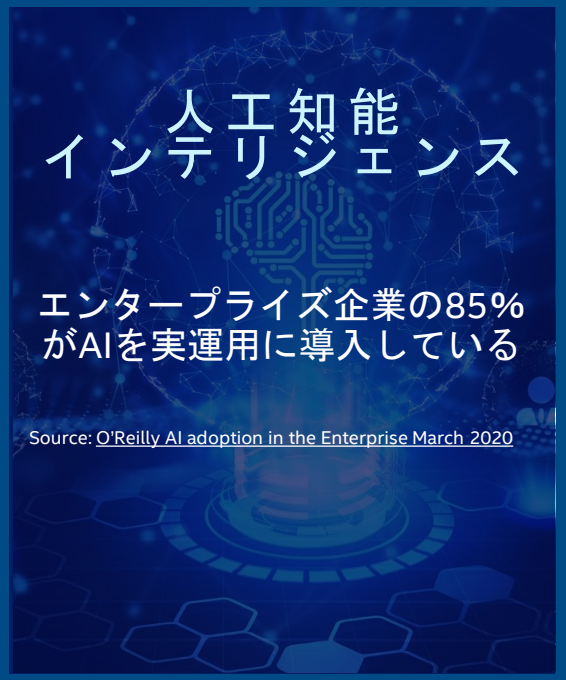
実際の結果を大きく異なるものにする要因の詳細については、最新の決算発表およびSEC提出書類（www.intc.com）を参照してください。

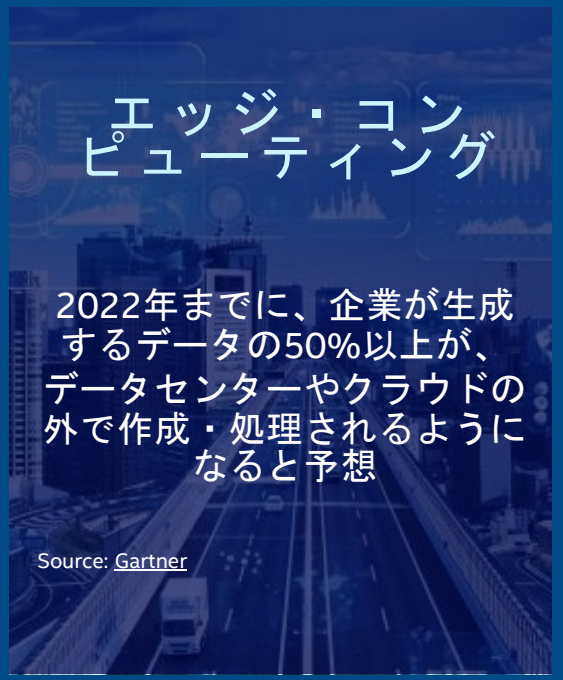
企業トレンドと要因

ビジネスの回復と成長のために求められることは、今日の変化と明日の不確実性によって再定義される









「クラウドは、デジタルエコノミーにおける企業経営の基本であり、
デジタル・エコノミーにおける企業の経営と推進の基礎となる

企業課題

ビジネス
レジリエンス

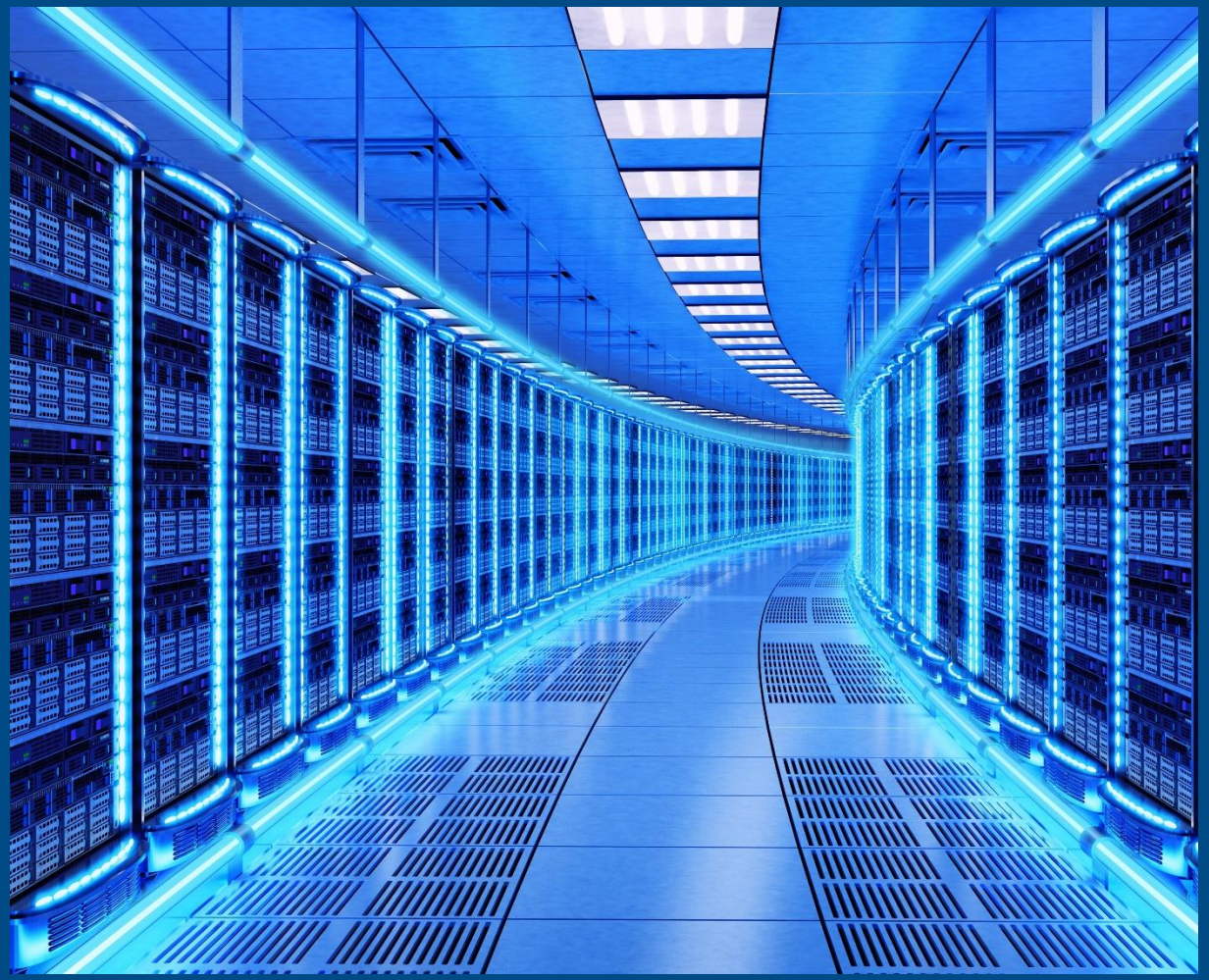
ワークロード
戦略

クラウド
マイグレーション

セキュリティー

コスト最適化

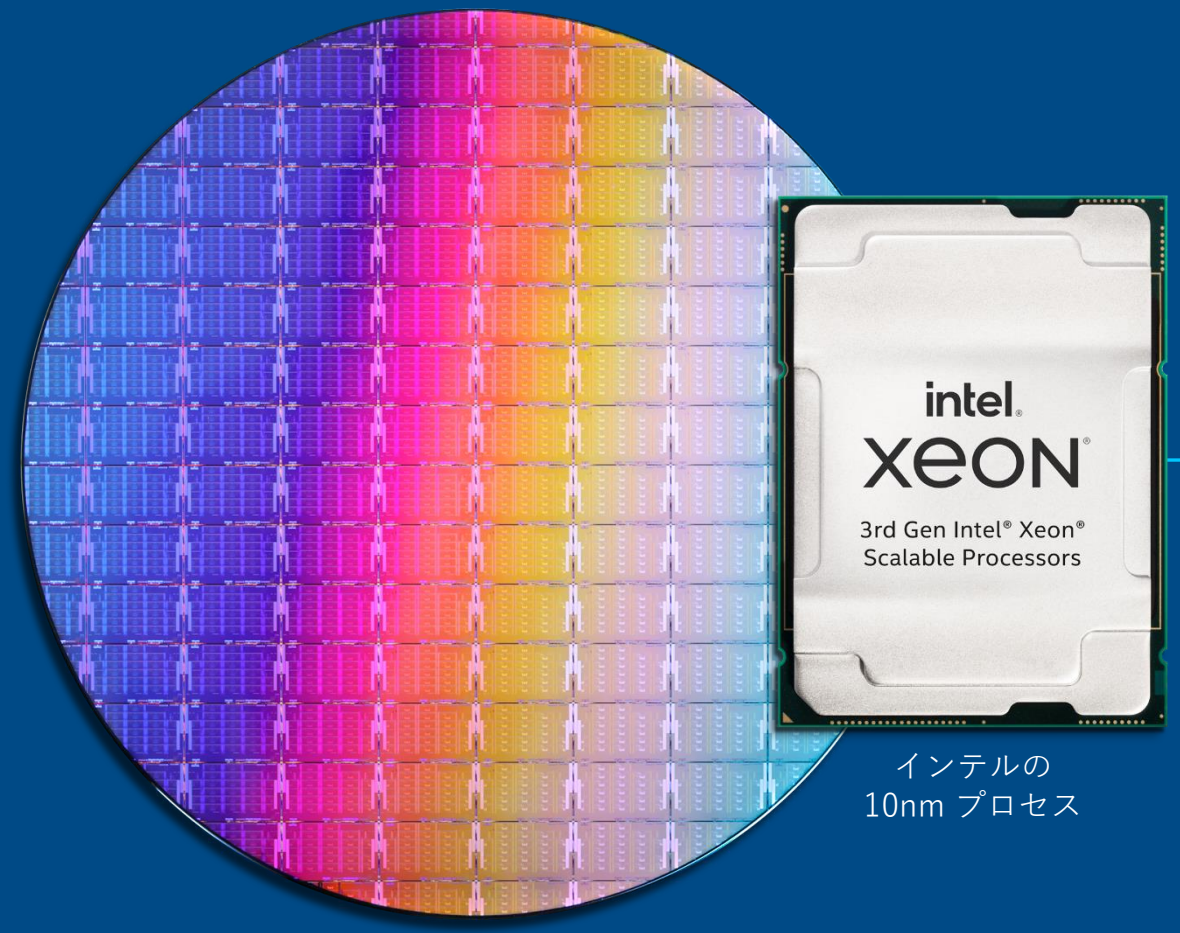
企業が直面するITの課題とは？



- 増加するリモートおよびモバイル・ワーカーをサポート
- VDIユーザーの重いワークロードの処理要求への対応
- eコマースサイトにアクセスする大量の顧客の処理
- ストレージを仮想化し、重いユーザーのワークロードをサポートできるソリューションの開発
- ラックあたりのユーザー密度の向上
- 増加するI/O需要への対応
- データをリアルタイムで活用するために、データの分類と分析の高速化
- 仮想化を拡張しながら、一般的なアプリケーションのパフォーマンスを向上させる方法

第3世代インテル® Xeon® スケーラブル・プロセッサー

パフォーマンスの柔軟性を向上



インテルの
10nm プロセス

最大 40 コア
(プロセッサー当たり)

IPC (クロックあたりの処理命令数) が
前世代と比較して20%向上
(28コア、ISO Freq. ISO Compiler)

最大 1.46 倍のパフォーマンス向上
Geomean of Integer, Floating Point, Stream Triad, LINPACK
8380 vs. 8280

1.74 倍の AI 推論向上 (前世代との比較)
8380 vs. 8280　BERT (自然言語処理)

最大 2.65 倍のパフォーマンス向上
5 年前のシステムとの比較
8380 と E5-2699v4 の比較

性能は、使用状況、構成、その他の要因によって異なります。構成については Appendix を参照 [1,3,5]

第3世代インテル® Xeon® スケーラブル・プロセッサー

Ice Lake-SP † (1 ~ 2ソケット) Whitley platform †

Performance made flexible

最大1.46X
General compute performance gain
vs Cascade Lake †

40 cores
最大 1.42X more cores
vs Cascade Lake

3.7 GHz
Top single-core turbo frequency
Platinum & Gold CPUs

組込みAI＆セキュリティー
DL Boost, SGX, TME, Crypto
New & Enhanced

競争力のある価格設定
Similar or lower price points
vs Cascade Lake†

プラットフォーム

new 最大 6TB
System Memory Capacity (Per Socket)
DRAM + PMem

new 最大 8CH
DDR4-3200 2 DPC (Per Socket)

new 最大 2.6X
Memory Capacity Increase vs. Cascade Lake

new 最大 64
Lanes PCI Express 4 (Per Socket)

先進のセキュリティ技術

new Intel Software Guard Extensions

new Intel Platform Firmware Resilience

new Intel Total Memory Encryption

new Intel Crypto Acceleration

拡張性、柔軟性、カスタマイズ性

new Intel Speed Select Technology

new Intel Optane persistent memory 200 series

Optimized Software

用途に最適化
お客様のニーズに合わせた設計

Network optimized | Cloud optimized | IoT optimized
Security optimized | Liquid cooled | Single socket value

組込みアクセラレーター
最も要求の厳しいワークロードに対応

AI w/ DL Boost | AVX-512 | Speed Select
VBMI | DDIO | Crypto

†開発コード名

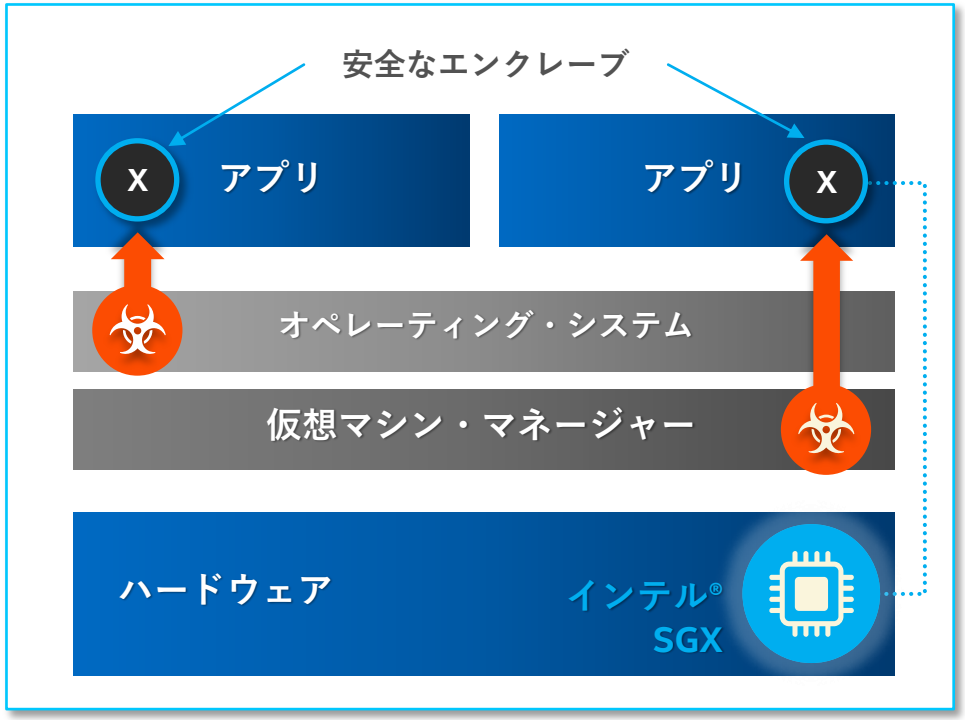
No product or component can be absolutely secure . Performance varies by use, configuration and other factors. For workloads and configurations see [125] at https://www.intel.com/3gen-xeon-config.

Information and performance details referenced above are preliminary and subject to change without notice.

DELL Technologies intel

信頼できる実行環境:
インテル® ソフトウェア・ガード・エクステンションズ (インテル® SGX)

オペレーティング・システムやハードウェア構成に関係なく、アプリケーション・データのセキュリティー保護を強化



- **ソフトウェア攻撃に対する防御が可能:** OS / ドライバー / BIOS / VMM / SMM への攻撃に対抗
- **機密情報 (データ、鍵など) の保護を強化:** 攻撃者にプラットフォームが完全制御されてしまった場合でも防御が可能
- **攻撃の防御が可能:** メモリーバスのスヌーピング、メモリー改ざん、RAM のメモリー内容に対する「コールドブート」攻撃などから防御

最小サイズのトラステッド・コンピューティング・ベース (TCB)

ほかのテクノロジーでは信頼境界内で権限付きソフトウェアが実行されてしまう

防御が難しい領域の保護を強化

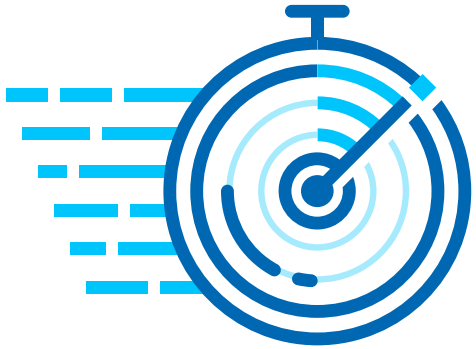
第 3 世代インテル® Xeon® スケーラブル・プロセッサー
ワークロード・アクセラレーションを実行する命令セット

	アプリケーション	使用例 / ワークロード	第 3 世代インテル® Xeon® スケーラブル・プロセッサー (Ice Lake†)	第 2 世代インテル® Xeon® スケーラブル・プロセッサー (Cascade Lake†)	第 3 世代 AMD EPYC (Milan†)
VNNI	8 ビット整数処理のアクセラレーション	AI ワークロードのパフォーマンス・アクセラレーション	対応	対応	未対応
AVX-512	演算要件の厳しいタスクに対するベクトル・アクセラレーション	HPC など、演算負荷の高いアプリケーション	対応	対応	未対応 AVX2
ベクトル AES (Advanced Encryption Standard) **ベクトル CLMUL**	暗号化 (CTR、CBC、XTS) と認証付き暗号化 (AES-GCM)	ディスクとの間で移動するデータ、またはすべての AES 処理。例: データベース暗号化、クラウド暗号化 (Hadoop など)	対応 AVX-512 (light)	未対応	対応 AVX2
VPMADD52	公開鍵の暗号生成 (RSA & DH)	SSL フロントエンド・ウェブサーバー接続 (NGINX、HA-Proxy、WordPress)	対応	未対応	未対応
SHA (Secure Hash Algorithm) 拡張機能	SHA-1 と SHA-256 のハッシュ化アルゴリズム	ハッシュ化、SSL、TLS、IPSec、データの重複排除、ブロックチェーン、ビットコイン	対応	未対応	対応
VBMI **AVX 512 (ベクトルバイト操作命令)**	小さな単位でデータをアルゴリズムに入力しながら即時演算を行うインライン圧縮	インメモリー・データベース (IMDB) のワークロード、圧縮 / 解凍のパフォーマンスを向上	対応	未対応	未対応

第 3 世代インテル® Xeon® スケーラブル・プロセッサーは差別化された命令セットを実装し、
最先端のワークロードで抜群のパフォーマンスを発揮

出典: インテルの命令については https://software.intel.com/content/www/us/en/develop/articles/intel-sdm.html#combined (英語) を参照。AMD の命令については、https://www.amd.com/system/files/TechDocs/24594.pdf (英語) を参照。

† 開発コード名



暗号処理を拡張する内蔵機能
インテルの暗号化アクセラレーション

インテルは暗号化アルゴリズムのパフォーマンスにかかる演算コストの削減において業界をリードし続けており、より多くのデータをより強固に保護することが可能

公開鍵の暗号化

5.6 倍

OpenSSL の RSA 署名 2048 ビットでシングルスレッドの公開鍵暗号生成を前世代と比較

対称暗号化

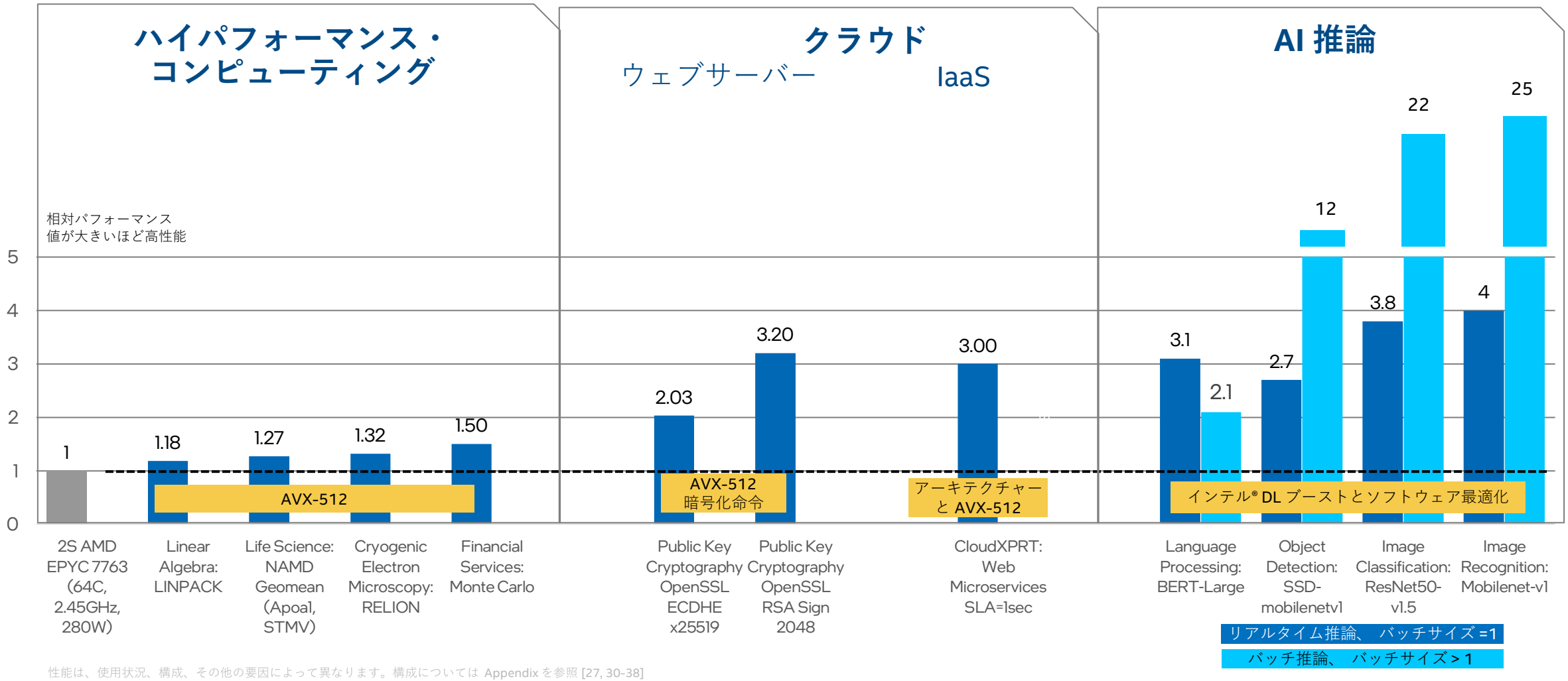
3.3 倍

AES-GCM モードでシングルスレッドの対称暗号化を前世代と比較

ワークロードと構成については、https://www.intel.com/3gen-xeon-config/ (英語) の [70,71] を参照。結果は変わる可能性があります。

HPC、クラウド、AI のパフォーマンス比較

2 ソケットのインテル® Xeon® Platinum 8380 プロセッサー (40C) と 2 ソケットの AMD EPYC 7763 (64C) の比較



性能は、使用状況、構成、その他の要因によって異なります。構成については Appendix を参照 [27, 30-38]

インテル® AVX-512

3D モデリング、オーディオ / ビデオ処理、財務分析など演算負荷の高いワークロードに対応

金融サービス

金融サービス業界におけるモンテカルロ・シミュレーションのパフォーマンス向上

50%

2 ソケットのインテル® Xeon® Platinum 8380 プロセッサー (40C) と 2 ソケットの AMD EPYC 7763 (64C) を比較

分子動力学

NAMD のパフォーマンス向上

27%

2 ソケットのインテル® Xeon® Platinum 8380 プロセッサー (40C) と 2 ソケットの AMD EPYC 7763 (64C) を比較

ワークロードと構成については、https://www.intel.com/3gen-xeon-config/ (英語) の [36,37] を参照。結果は変わる可能性があります。

第 3 世代インテル® Xeon® スケーラブル・プロセッサー
内蔵アクセラレーションによる性能向上

インテル® DDIO 有効 / 無効の比較

最大

1.3 倍

高速

DPDK L3 パケット転送

6338N

インテル® AVX-512 インテル® AVX2 との比較

最大

1.62 倍

パフォーマンス向上

LINPACK

8380

インテルの暗号化アクセラレーション 有効 / 無効の比較

最大

4.2 倍

パフォーマンス向上

NGINX ECDHE-X25519-RSA2K

6338N と 6252N の比較

インテル® DL ブースト (VNNI) int8 と fp32 の比較

最大

4.3 倍

パフォーマンス向上

ResNet50-v1.5

8380

性能は、使用状況、構成、その他の要因によって異なります。構成については Appendix を参照 [14,15,16.51]

柔軟性の高いパフォーマンスで極めて要求の厳しいワークロードにも対応

インテリジェント・エッジからクラウド最大、前世代からの大幅なパフォーマンス向上

クラウド	5G	IoT	HPC	人工知能 (AI)
最大 1.5 倍[1]	最大 1.62 倍	最大 1.56 倍	最大 1.57 倍	最大 1.74 倍
レイテンシー要件の厳しいワークロードで向上	ネットワーク / 通信ワークロードで向上	画像分類の推論で向上	ワクチン研究のモデリングで高速化	自然言語処理推論の向上

性能は、使用状況、構成、その他の要因によって異なります。構成については Appendix を参照 [5, 7, 17, 19]

第3世代 インテル® Xeon® スケーラブル・プロセッサーを搭載した15GのDell EMC PowerEdgeサーバー

インテルとデル様は、最も要求の厳しいワークロードを処理するための次世代サーバーを推進する信頼できるソリューションの包括的なポートフォリオと、連携するリソースを構築しています。

Dell EMC PowerEdgeサーバは、第 3 世代インテル® Xeon® スケーラブル・プロセッサーを搭載し、企業を支援します。

- 最新のワークロードを高速化
- レスポンシブ・データベースの高速化
- アクセラレータ・ソリューション
- 最も安全なインフラを実現

SQLサーバーの高速化

インテル® データセントリック・ポートフォリオ

最適化されたパフォーマンス

vSAN上のSQLオーダーの高速化

最大 **2x**

6248搭載のr640と比較して、
より多くのオーダー/分を実現[1]

ユーザー密度の向上

SQLユーザー密度の向上

最大 **60%**

6248搭載のr740XDと比較して
より多くのVMを搭載[2]

SQLトランザクションの高速化

最大 **28%**

r740XD w/6248 と比較して、クエリートランザクションが高速[2]

TCOの削減

より多くのメモリー

最大 **30%**

インテル® PMEMとDRAMの比較で、
1ドルあたりのメモリー使用量が増加

PowerEdge

VxRail

intel. XEON

intel. OPTANE PERSISTENT MEMORY

intel. OPTANE SSD

intel. ETHERNET

1 https://www.principledtechnologies.com/Dell/PowerEdge-R650-vs-previous-generation-OPM-0621.pdf
2 https://www.principledtechnologies.com/Dell/PowerEdge-R750-analytics-comparison-0621.pdf

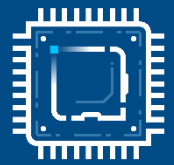
Performance made flexible.

オラクルをアクセラレート

インテル® データセントリック・ポートフォリオ

最適化されたパフォーマンス

オラクル・データベース



最大 50%

2ソケットシステムにおけるオラクルライセンスあたりの性能向上[12]

ライセンスコストの削減

オラクル・インメモリー コラムストア

最大 40.2%

従来のDRAMのみのテストケースと比較して、タスクの実行時間が短縮されました。[13]

オラクルTimesTen

最大 6X

インテル® Optane™ パーシステント・メモリーへの同期書き込みを使用した場合の耐久性のあるトランザクション・パフォーマンス[14]

TCOの削減

オラクルRAC

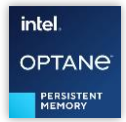
最大 1.11M

PowerFlexサーバーの場合、1msあたりのIOPS*

PowerEdge

VxRail

パワーフレックス



12 http://www.oracle.com/us/corporate/contracts/processor-core-factor-table-070634.pdf
13 https://www.delltechnologies.com/resources/en-us/asset/technical-guides-support-information/solutions/h18295-inteloptanepmem-oracleinmem-dg.pdf - ページ 29
14 https://blogs.oracle.com/timesten/oracle-timesten-intel-optane-persistent-memory
*デルのホワイトペーパーの完成待ち

Performance made flexible.

Ai推論速度の倍増

インテル® データセントリック・ポートフォリオ

物体検出の高速化

オベクト検出
MLPerf INT8 推論

最大 2X

28CのCPUを使用し、1秒間に処理する画像数

オンライン分類の高速化

画像分類
MLPerf INT8 推論

最大 30%

28CのCPUを使用し、1秒あたりの処理を高速化

オフライン分類の高速化

オフライン画像分類
MLPerf INT8推論

最大 7%

28CのCPUを使用し、1秒あたりの処理を高速化

PowerEdge

VxRail

intel. XEON

パフォーマンスデータとコンフィギュレーション： https://www.delltechnologies.com/resources/en-us/asset/tech-notes/products/servers/mlperf-inference.pdf

Performance made flexible.

インテルデータセンタ戦略

インテル・テクノロジーに基づくエッジ - クラウド ソリューション

マルチクラウドを実現するビジョン

5Gネットワーク・モメンタムを加速

インテリジェント・エッジの育成

開発者コミュニティの活性化

Xeonを基盤とした、包括的なテクノロジー・ポートフォリオ

Move Faster

Intel® Ethernet
Intel® Silicon Photonics
Intel® Tofino

Process Everything

欧州のデジタルの未来を強化する。
世界トップレベルのエコ・システムの構築

今後10年間で最大**800**億ユーロの投資
研究開発から設計、先端チップパッケージング、製造ファウンドリーサービス

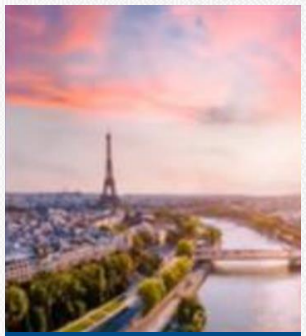
第1フェーズで330億ユーロを投資。

ドイツ

最先端-Fab

アイルランド

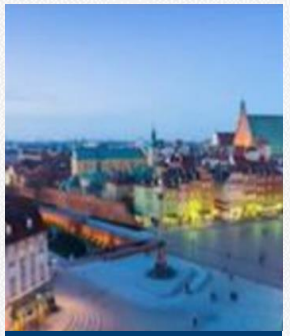
ファブ・エクステンション



フランス

新しい研究開発・デザイン拠点
デザインハブ

イタリア



ポーランド

スペイン

ベルギー

オランダ

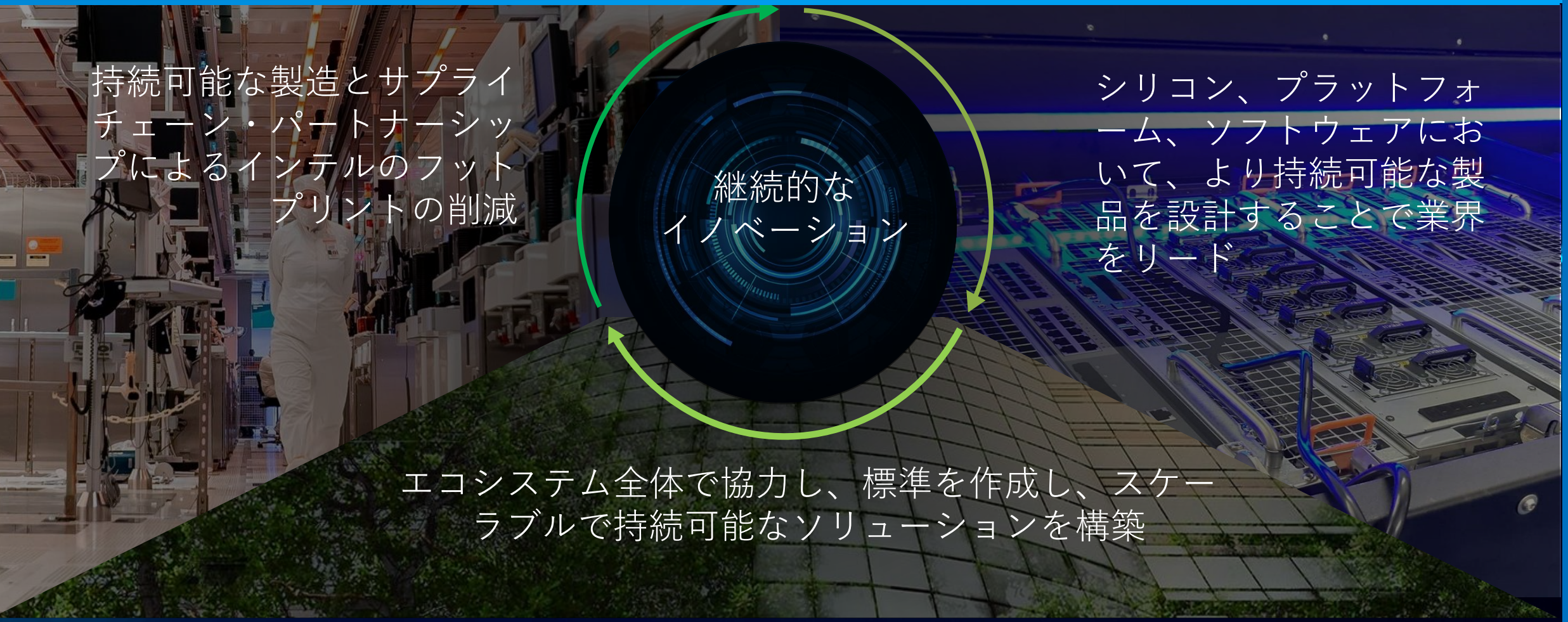
研究開発、製造、最先端パッケージング技術、ファウンドリーサービスにおける能力の拡大

私たちは、まさに**共に歴史を築いているのです**

サステナビリティー インテルが描くデータセンターのビジョン

#1
2022年 米国で最も持続可能な企業[1]
バロンズ

製造から製品、ソリューションまで、より持続可能なコンピューティング業界を構築する



1https://www.barrons.com/articles/most-sustainable-companies-51644564600

インテル® Xeon® プラットフォーム

パフォーマンスとサステナビリティーの向上

#1
2022年 米国で最も持続可能な企業[1]
バロンズ

電力の80%は
再生可能エネルギー
カーボンの低減

AIアクセラレーター
4倍の perf / watt 向上を実現[2]
AI推論ワークロードのために - TensorFlow

アクセラレーターを内蔵し、 perf / watt を向上
セキュリティー、ネットワーク・ワークロード、ハイパフォーマンス・コンピューティング・ワークロードなどに利用可能

テレメトリー機能搭載
エネルギー効率測定が可能

2030年最大にエネルギー効率を10倍に
クライアントおよびサーバーCPUの場合

廃棄物に関する循環型経済戦略の構築
埋立廃棄物量5%*。

1 https://www.barrons.com/articles/most-sustainable-companies-51644564600
2 結果は異なる場合があります。構成の詳細については、https://www.servethehome.com/stop-leaving-performance-aws-ec2-m6i-intel-instances/ を参照してください*。インテル 2021-22 年度 CSR 報告書

まとめ ...

今後も増加し続けるITの課題に対して、インテルはこれらの課題を克服するためのハードウェア、ソフトウェア、サービスなどのテクノロジーを開発しています

第3世代 インテル® Xeon® スケーラブル・プロセッサーを搭載した 15世代 Dell PowerEdge サーバー ビジネスの差別化を支援します

インテルは、より持続可能なコンピュート・インダストリーを実現するために努力を継続します

We are the makers of wonderful.

ご清聴ありがとうございました。

Appendix

27. **3.00x higher CloudXPRT Web Microservices with SLA < 1 sec.** Ice Lake: 2-socket Intel® Xeon® 8380 (40C/2.3GHz, 270W TDP) on Intel Software Development, HT on, Turbo on, SNC off, 512GB (16x32GB DDR4-3200), ucode x270, Ubuntu 20.04 LTS, 5.8.0-40-generic, CloudXPRT version 1.1, Tested by Intel and results as of February 2021. Milan: 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92, SMT on, Boost on, Power deterministic mode, NPS=1, 512 GB (16 x32GB DDR4-3200), ucode 0xa001114, Ubuntu 20.04 LTS, 5.8.0-40-generic, CloudXPRT version 1.1. Tested by Intel and results as of March 2021. Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

28. 1.5x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC 7763 (64C Milan): (geomean of 20 workloads including logistic regression inference, logistic regression fit, ridge regression inference, ridge regression fit, linear regression inference, linear regression fit, elastic net inference, XGBoost Fit, XGBoost predict, SSD-ResNet34 inference, Resnet50-v1.5 inference, Resnet50-v1.5 training, BERT Large SQuaD inference, kmeans inference, kmeans fit, brute_knn inference, SVC inference, SVC fit, dbscan fit, traintestsplit)
8380: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic/5.4.0-64-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, tested by Intel, and results as of March 2021. 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8/INTEL SSDSC2KG019T8, tested by Intel, and results as of March 2021.
ResNet50-v1.5 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=128, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart , ResNet50-v1.5 AMD : gcc-9.3.0, oneDNN 1.6.4, BS=128, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5
ResNet50-v1.5 Training Intel : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart , ResNet50-v1.5 Training AMD : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5 SSD-ResNet34 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, AMD : SSD-ResNet34, gcc-9.3.0, oneDNN 1.6.4, BS=1, FP32, TensorFlow- 2.4, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/object_detection/tensorflow/ssd-resnet34 BERT-Large SQuAD Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, AMD : BERT-Large SQuAD, gcc-9.3.0, oneDNN 1.6.4, BS=1, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/language_modeling/tensorflow/bert_large Python : Python 3.7.9, SciKit-Learn : Sklearn 0.24.1, oneDAL : Daal4py 2021.2, XGBoost : XGBoost 1.3.3 : Benchmarks: https://github.com/IntelPython/scikit-learn_bench

29. 1.3x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. NVIDIA A100 GPU: (geomean of 20 workloads including logistic regression inference, logistic regression fit, ridge regression inference, ridge regression fit, linear regression inference, linear regression fit, elastic net inference, XGBoost Fit, XGBoost predict, SSD-ResNet34 inference, Resnet50-v1.5 inference, Resnet50-v1.5 training, BERT Large SQuaD inference, kmeans inference, kmeans fit, brute_knn inference, SVC inference, SVC fit, dbscan fit, traintestsplit) 8380: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, tested by Intel, and results as of March 2021.
DL Measurements on A100: 1-node, 2-socket AMD EPYC 7742 (64C) with 256GB (8 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x8301038, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-42-generic, INTEL SSDSC2KB01, NVIDIA A100-PCIe-40GB, HBM2-40GB, Accelerator per node =1 , tested by Intel, and results as of March 2021. ML Measurements on A100 : 1-node, 2-socket AMD EPYC 7742 (64C) with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0x8301034, HT on, Turbo on, Ubuntu 18.04.5 LTS, 5.4.0-42-generic,NVIDIA A100 (DGX-1) , 1.92TB M.2 NVMe, 1.92TB M.2 NVMe RAID tested by Intel, and results as of March 2021.
ResNet50-v1.5 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart ResNet50-v1.5 NVIDIA :A100 (7 instance/GPU), BS=1,TensorFlow -1.5.5 (NGC: tensorflow:21.02-tf1-py3), https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Classification/ConvNets/resnet50v1.5, TF AMP (FP16+TF32); ResNet50-v1.5 Training Intel : gcc-9.3.0, oneDNN 1.6.4, BS=256, FP32, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart , ResNet50-v1.5 Training NVIDIA :A100, BS=256,TensorFlow -1.5.5 (NGC: tensorflow:21.02-tf1-py3),https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Classification/ConvNets/resnet50v1.5, TF32;
BERT-Large SQuAD Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/ A100 : BERT-Large SQuAD, BS=1, A100 (7 instance/GPU), TensorFlow -1.5.5 (NGC: tensorflow:20.11-tf1-py3), https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT,TF AMP (FP16+TF32) ; SSD-ResNet34 Intel : gcc-9.3.0, oneDNN 1.6.4, BS=1, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, SSD-ResNet34 NVIDIA :A100 (7 instance/GPU), BS=1,Pytorch – 1.8.0a0 (NGC Container, latest supported): A100 : SSD-ResNet34 (NGC: pytorch:20.11-py3), https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Detection/SSD, AMP (FP16 +TF32) ;
Python : Intel: Python 3.7.9, Scikit-Learn : Sklearn 0.24.1, OneDAL : Daal4py 2021.2, XGBoost: XGBoost 1.3.3 Python : NVIDIA A100 : Python 3.7.9, Scikit-Learn : Sklearn 0.24.1, CuML 0.17, XGBoost 1.3.0dev.rapidsai0.17, Nvidia RAPIDS : RAPIDS 0.17, CUDA Toolkit : CUDA 11.0.221 Benchmarks: https://github.com/IntelPython/scikit-learn_bench

Appendix

30. **LINPACK**. Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG96, App Version: The Intel Distribution for LINPACK Benchmark; Build notes: Tools: Intel MPI 2019u7; threads/core: 1; Turbo: used; Build: build script from Intel Distribution for LINPACK package; 1 rank per NUMA node: 1 rank per socket, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: AMD official HPL 2.3 MT version with BLIS 2.1; Build notes: Tools: hpc-x 2.7.0; threads/core: 1; Turbo: used; Build: pre-built binary (gcc built) from https://developer.amd.com/amd-aocl/blas-library/; 1 rank per L3 cache, 4 threads per rank, tested by Intel and results as of March 2021

31. **Monte Carlo FSI Kernel**. Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG96, App Version: v1.1; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 1; Turbo: used; Build knobs: -O3 -xCORE-AVX-512 -qopt-zmm-usage=high -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: v1.1; Build notes: Tools: Intel MKL 2020u4, Intel C Compiler 2020u4, Intel Threading Building Blocks 2020u4; threads/core: 2; Turbo: used; Build knobs: -O3 -march=core-avx2 -fimf-precision=low -fimf-domain-exclusion=31 -no-prec-div -no-prec-sqrt tested by Intel and results as of March 2021

32. **NAMD Geomean of Apoa1, STMV**. Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG96, App Version: 2.15-Alpha1 (includes AVX tiles algorithm); Build notes: Tools: Intel MKL , Intel C Compiler 2020u4, Intel MPI 2019u8, Intel Threading Building Blocks 2020u4; threads/core: 2; Turbo: used; Build knobs: -ip -fp-model fast=2 -no-prec-div -qoverride-limits -qopenmp-simd -O3 -xCORE-AVX-512 -qopt-zmm-usage=high, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: 2.15-Alpha1 (includes AVX tiles algorithm); Build notes: Tools: Intel MKL , AOCC 2.2.0, gcc 9.3.0, Intel MPI 2019u8; threads/core: 2; Turbo: used; Build knobs: -O3 -fomit-frame-pointer -march=znver1 -ffast-math, tested by Intel and results as of March 2021

33. **RELION Plasmodium Ribosome**. Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 256 GB (16 slots/ 16GB/ 3200) total DDR4 memory, ucode 0x261, HT on, Turbo on, CentOS Linux 8.3.2011, 4.18.0-240.1.1.el8_3.crt1.x86_64, 1x Intel_SSDSC2KG96, App Version: 3_1_1; Build notes: Tools: Intel C Compiler 2020u4, Intel MPI 2019u9; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -g -debug inline-debug-info -xCOMMON-AVX-512 -qopt-report=5 –restrict, tested by Intel and results as of March 2021 EPYC 7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=4, Red Hat Enterprise Linux 8.3, 4.18, 1x Samsung_MZ7LH3T8, App Version: 3_1_1; Build notes: Tools: Intel C Compiler 2020u4, Intel MPI 2019u9; threads/core: 2; Turbo: used; Build knobs: -O3 -ip -g -debug inline-debug-info -march=core-avx2 -qopt-report=5 -restrict tested by Intel and results as of March 2021

34. **3.88x higher INT8 real-time inference throughput & 22.09x higher INT8 batch inference throughput on ResNet-50 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan 8380**: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, ResNet50-v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow 2.5 (container-intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, ResNet50-v1.5, gcc-9.3.0, oneDNN 1.6.4, BS=1,128, FP32, TensorFlow- 2.4.1, Model : https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/resnet50v1_5, tested by Intel, and results as of March 2021.

35. **2.79x higher INT8 real-time inference throughput & 12x higher INT8 batch inference throughput on SSD-MobileNet-v1 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan 8380**: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, SSD-MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,448, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow 2.5 (container-intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, SSD-MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,448, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/object_detection/tensorflow/ssd-mobilenet, tested by Intel, and results as of March 2021.

Appendix

36. Upto 25x higher AI performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC 7763 (64C Milan) ,4.01x higher INT8 real-time inference throughput & 25.05x higher INT8 batch inference throughput on MobileNet-v1 with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan : 1-node, 2x Intel Xeon Platinum 8380 processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260 , HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,56, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, test by Intel on March 2021. 1-node, 2x AMD Epyc 7763 on GigaByte with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, MobileNet-v1, gcc-9.3.0, oneDNN 1.6.4, BS=1,56, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/image_recognition/tensorflow/mobilenet_v1, tested by Intel and results as of March 2021.

37. **3.18x higher INT8 real-time inference throughput & 2.17x higher INT8 batch inference throughput on BERT Large SQuAD with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost vs. FP32 AMD EPYC Milan 8380:** 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3GHz, 270W TDP) processor on Intel Software Development Platform with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Intel_SSDSC2KG96, Intel SSDPE2KX010T8, BERT Large SQuAD , gcc-9.3.0, oneDNN 1.6.4, BS=1,128, INT8, TensorFlow 2.4.1 with Intel optimizations for 3rd Gen Intel Xeon Scalable processor, upstreamed to TensorFlow- 2.5 (container- intel/intel-optimized-tensorflow:tf-r2.5-icx-b631821f), Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/quickstart/, tested by Intel, and results as of March 2021.
7763: 1-node, 2-socket AMD EPYC 7763 (64C/2.45GHz, 280W cTDP) on GIGABYTE R282-Z92 server with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, SMT on, Boost on, Power deterministic mode, NPS=1, Ubuntu 20.04 LTS, 5.4.0-65-generic, 1x Samsung_MZ7LH3T8, BERT Large SQuAD , gcc-9.3.0, oneDNN 1.6.4, BS=1,128, FP32, TensorFlow- 2.4.1, Model zoo: https://github.com/IntelAI/models/tree/icx-launch-public/benchmarks/language_modeling/tensorflow/bert_large, tested by Intel, and results as of March 2021.

38. **3.20x higher OpenSSL RSA Sign 2048 performance, 2.03x higher OpenSSL ECDHE x25519 performance** 8380 : 1-node, 2x Intel(R) Xeon(R) Platinum 8380 CPU on M50CYP2SB2U with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xd000270, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x INTEL_SSDSC2KG01 , OpenSSL 1.1.1j, GCC 9.3.0, QAT Engine v0.6.4, Tested by Intel and results as of March 2021. 7763 : 1-node, 2x AMD EPYC 7763 64-Core Processor on R282-Z92-00 with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xa001114, HT On, Turbo Off, Ubuntu 20.04.1 LTS, 5.4.0-65-generic, 1x SAMSUNG_MZ7LH3T8 , OpenSSL 1.1.1j, GCC 9.3.0, Tested by Intel and results as of March 2021.

39. 1) 5G vRAN: Results have been estimated or simulated: Based on 2x throughput from 32Tx32R (5Gbps) on 2nd Gen Intel® Xeon® Gold 6212U processor to 64Tx64R (10Gbps) on 3rd Gen Intel® Xeon® Gold 6338N processor at ~185W

    2) VDI: https://www.principledtechnologies.com/VMware/VMware-HCI-Intel-Optane-VDI-0420.pdf

    3) Azure stack HCI – Ice Lake Configuration:

    4 Node, 2x Intel® Xeon® Gold 6330 CPU, 1x Intel® Server Board M50CYP, Total Memory: 256GB (16 x 16 GB 3200MHz DDR4 RDIMM,HyperThreading: Enable, Turbo: Enable ,Storage (boot): 1 x Intel® SSD D3-S4510 Series (480GB, 2.5in SATA 6Gb/s, 3D2, TLC), Storage: 4x Intel® SSD DC P4610 Series (3.2TB) (NVMe), Network devices: 1 x 100 GbE Intel(R) Ethernet Network Adapter E810-C-Q2, Network speed: 25 GbE, 1 x 10 GbE Intel(R) Ethernet Converged Network Adapter X550-T2, Network Speed: 10 GbE, OS/Software: Microsoft Azure Stack HCI build 17763, Benchmarks: DiskSpd (QD=8,30w:70r): 1.396M IOPS @4.03ms(r), @6.95ms(w) for 90% requests, Tested by Intel as of 12-Mar-2021.

    Azure stack HCI – Cascade Lake Configuration:

    4 Node, 2x Intel® Xeon® Gold 6230, 1x Intel® Server Board S2600WFT, Total Memory: 512 GB Intel® Optane™ ™ DC persistent Memory, 4 slots/128 GB/2666 MT/s and 192 GB, 12 slots/16 GB/2666 MT/s,HyperThreading: Enable, Turbo: Enable, Storage (boot): 1x 480 GB Intel® SSD 3520 Series M.2 SATA, Storage (cache): 2x 375 GB Intel® Optane™ DC SSD P4800X, Storage (capacity):4x 4 TB Intel® SSD DC P4510 PCIe NVMe, Network devices: 1x 25 Gbps Chelsio* Network Adapter, Network speed: 25GbE, OS/Software: Windows Server* 2019 Datacenter Edition build 17763, Benchmarks: DiskSpd (QD=8,30w:70r): 588K IOPS @4.99ms(r), @19.54ms(w) for 90% requests, Tested by Intel as of 22-Feb-2019.