RESUMEN DE LA SOLUCIÓN

Plataforma convergente de HPC

El caso para ejecutar IA y análisis en clústeres de HPC

Crear una plataforma convergente para ejecutar cargas de trabajo de simulación, modelado, inteligencia artificial (IA) y análisis en una única infraestructura de clúster permite una innovación de vanguardia, al mismo tiempo que aumenta el valor y el uso de los recursos. Los servidores con arquitectura Intel® son la base ideal para dicha convergencia. Este resumen de solución presenta los desafíos y las oportunidades en torno a la ejecución de cargas de trabajo de IA y análisis en clústeres existentes de informática de alto desempeño (HPC).

El crecimiento exponencial en el tamaño de los almacenes de datos de empresa, académicos y gubernamentales durante la pasada década ha impulsado la necesidad de contar con recursos que puedan transformar esos datos de un potencial incipiente a información procesable. Mediante el uso de motores de análisis como Apache Spark*, la base de información inteligente se convirtió en algo habitual en las industrias. Estas soluciones siguen haciéndose más sofisticadas a medida que pasa el tiempo, lo cual impulsa más valor para diversas organizaciones.

Una aplicación de IA (incluido el aprendizaje automatizado y el aprendizaje profundo) es hacer que el análisis sea aún más potente, con redes neuronales entrenadas para predecir acontecimientos futuros según entradas pasadas, por ejemplo. En un estudio realizado por Narrative Science, se informa que el 61 por ciento de los encuestados actualmente está implementando inteligencia artificial, y el análisis predictivo es el tipo de solución con IA de uso más amplio.[1] La inteligencia artificial sigue volviéndose cada vez más importante en toda una gama de organizaciones; MarketWatch* estima que el crecimiento del mercado global de IA alcanza una tasa de crecimiento anual compuesto del 36 por ciento hasta el 2024.[2]

En muchas organizaciones, existe una tendencia de adoptar las plataformas de análisis y las tecnologías de IA como entidades distintas, en lugar de enfocar su integración en las arquitecturas existentes de sistema para hacer que los procesos empresariales sean más eficaces y eficientes. En muchos casos, se pueden crear nuevos clústeres dedicados para implementar estas funciones, como se muestra en la Figura 1. Este enfoque crea silos de datos y la necesidad de costosas operaciones relacionadas con el movimiento y almacenamiento de los datos. La existencia de varios clústeres también revela un escenario de infraestructura con poco uso, particularmente en los clústeres de IA utilizados principalmente para entrenar redes de aprendizaje profundo, lo cual suele ser esporádico.

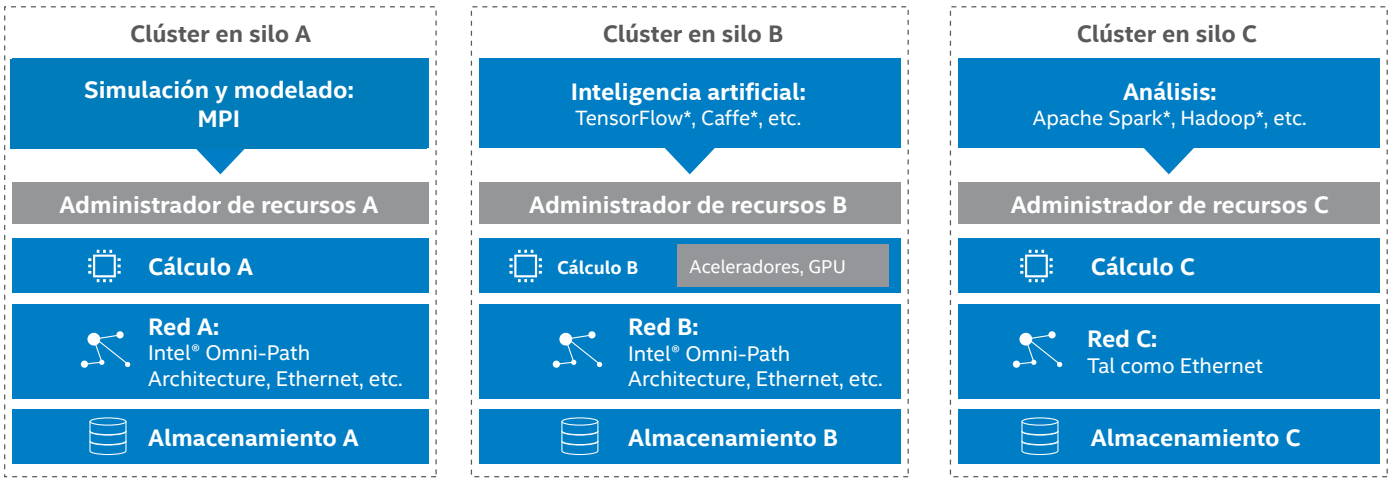

Figura 1. Cargas de trabajo repartidas en clústeres separados en silos.

Superar los retos para comprender el valor de la HPC

Puesto que las organizaciones con infraestructuras existentes de HPC adoptan inteligencia artificial y análisis, muchas de ellas mencionan que uno o más de los siguientes desafíos les impide ejecutar las tres cargas de trabajo en un sistema:

- Se deben crear **marcos y pilas de software coexistentes** para simulación, modelado, análisis e IA que trabajen juntas en armonía. Lamentablemente, las pilas de software y los marcos para HPC, análisis e IA son muy diferentes, y cada carga de trabajo debe tener su propia pila de software cargada en un clúster. En particular, los administradores de recursos encargados de distribuir recursos para estas cargas de trabajo suelen estar diseñados de manera muy diferente. Intel proporciona un conjunto creciente de arquitecturas de solución para este propósito, como se aborda en el resumen adjunto denominado *Supporting Simulation and Modeling, Analytics, and AI on a Common Platform* (Admitir la simulación, el modelado, el análisis y la IA en una plataforma común).
- El **hardware sin diseño de HPC**, como los servidores equipados con aceleradores y GPU, atraen el interés de muchas organizaciones como un medio potencial para obtener un mayor desempeño para cargas de trabajo de inteligencia artificial en particular. De hecho, las características de los sistemas que ya poseen en infraestructuras típicas de HPC (por ejemplo, núcleos informáticos sólidos conectados a estructuras de red de alto desempeño y a almacenamiento compartido de alto desempeño) son aptas para las necesidades de cargas de trabajo de análisis e inteligencia artificial. Intel ha invertido fuertemente para aumentar el desempeño de la IA en los procesadores escalables Intel® Xeon® y también para ingresar nuevas instrucciones de IA específicas para las CPU Intel® Xeon®, lo cual las convierte en una excelente elección para cargas de trabajo de inteligencia artificial.
- La **separación cultural y operativa** entre equipos de HPC se centra en clústeres locales sin sistema operativo, a diferencia de las orientaciones en la nube de muchos equipos de IA y análisis, las cuales se pueden extender a enfoques funcionales como DevOps. Además, los equipos de IA y análisis suelen usar lenguajes de alto nivel como Python*, Scala* y Java*, mientras que los equipos de HPC suelen utilizar lenguajes de nivel más bajo como C/C++ y Fortran.

Superar estos desafíos y las suposiciones que los acompañan puede ser un factor importante para guiar a una organización en la transición a una infraestructura eficiente para los tres tipos de cargas de trabajo. Los beneficios en costo de reunir las cargas de trabajo de simulación, modelado, inteligencia artificial y análisis en una sola infraestructura de clúster pueden apreciarse en dos elementos. Desde un punto de vista del gasto de capital (CAPEX), se reduce la necesidad de nuevos gastos, al mismo tiempo que se obtiene el máximo provecho de las inversiones en clúster existentes y futuras. En términos de gastos operativos (OPEX), se reduce el costo de funcionamiento y mantenimiento del entorno mediante la simplificación de la infraestructura para que ejecute un clúster en lugar de varios clústeres.

Estos beneficios se obtienen mediante pilas de soluciones ofrecidas por los mismos fabricantes de equipos originales (OEM) del servidor, con quienes ya trabajan las organizaciones de TI, pues son sus proveedores para su infraestructura existente de HPC. Como se muestra en la Figura 2, los paquetes de HPC convergente están listos para impulsar la innovación y ofrecer valor en los flujos de trabajo, incluida la HPC convencional, la inteligencia artificial basada en HPC y el análisis basado en HPC. A modo de proyección para el 2021, se espera que las implementaciones de HPC convencional sigan dominando la base instalada; sin embargo, se pronostica que el análisis y especialmente la IA muestren tasas de crecimiento aún mayores.[3]

Necesidades emergentes en simulación, modelado, inteligencia artificial y análisis

Ya que las organizaciones de TI establecen sus estrategias de arquitectura para los próximos años, muchos ven cada vez más puntos en los que la IA y el análisis se cruzan con otras cargas de trabajo y procesos de negocio. Al mismo tiempo, los conjuntos de datos que deben manejar los trabajos de simulación y modelado aumentan de forma masiva, incentivados por la actual ciencia del uso intensivo de datos, así como por los campos emergentes en computación cognitiva. Estas tendencias ponen de relieve la necesidad de una amplia interoperabilidad entre los sistemas subyacentes,

Figura 2. Niveles de inversión y crecimiento en cargas de trabajo de HPC de muestra hasta el 2021.[3]

lo cual se ve obstaculizado por la ejecución de cargas de trabajo de simulación, modelado, inteligencia artificial y análisis en clústeres separados. Por ejemplo, consideremos la ineficacia de un proceso en el que la simulación, el modelado, la limpieza de datos y la inferencia basada en inteligencia artificial ocurren en clústeres separados.

Si se ejecutan los tres al mismo tiempo en el mismo clúster, se pueden lograr varias ventajas. En primer lugar, como se mencionó anteriormente, la necesidad de copiar y almacenar los datos se redujo drásticamente, o bien se eliminó. Además, se puede evitar la demora asociada con la transferencia de grandes cantidades de datos entre clústeres, lo cual es particularmente importante para admitir flujos de trabajo en tiempo real y de forma flexible. Esta convergencia se volverá cada vez más importante a medida que las necesidades emergentes exijan nuevas capacidades dentro de las organizaciones.

Por ejemplo, la implementación de puntos de conexión de IoT en las industrias generará flujos de datos con órdenes de una magnitud que, en muchos casos, será mayor que antes. Ordenar estos flujos para identificar los puntos de datos de interés, patrones y otro tipo de información es una tarea enorme que se adapta bien al análisis basado en inteligencia artificial. Según estos resultados, se pueden identificar oportunidades para la optimización de un proceso de negocio o un sistema basado en agentes, el cual se puede investigar en todas sus variantes utilizando software de simulación y modelado. Se puede utilizar inteligencia artificial para mejorar el diseño y la calibración de los trabajos de simulación y modelado de hoy en día.

El imperativo de la interoperación entre los sistemas que cooperan mediante este tipo de flujo de trabajo general es claro. Además, a medida que estos procesos se integren en conjunto de forma más fluida, los procesos podrán volverse más eficientes y eficaces. Superar con éxito este conjunto de desafíos interrelacionados prepara el camino para un conjunto de casos de uso que abarca industrias y campos de investigación, los cuales van desde perspectivas de datos, detección de fraudes y robótica hasta medicina individualizada, meteorología y climatología.

En consecuencia, muchas empresas e instituciones tienen la necesidad de modernizar sus infraestructuras de HPC para lograr una mayor flexibilidad. Esta modernización debe esforzarse específicamente por cumplir y reunir las necesidades de las diversas partes de la organización, como se ilustra en la Figura 3. Los clústeres con propósito múltiple son ideales para abordar esta gama de intereses con la máxima flexibilidad, eficiencia y desempeño. Los servidores con arquitectura Intel son la base ideal para esta infraestructura.

Un enfoque de sistemas abiertos para una infraestructura convergente a gran escala

Los clústeres de HPC existentes, diseñados con servidores de propósito general y gran capacidad, y con la arquitectura Intel pueden ofrecer un desempeño flexible, rentable y eficaz en cargas de trabajo. Además de eliminar la necesidad de comprar, configurar, administrar y brindar soporte a recursos separados, este enfoque reduce la complejidad. A diferencia del uso de sistemas diseñados para una función específica, los trabajos se pueden programar en cualquier lugar del entorno cuando sea necesario, lo cual permite aprovechar al máximo los recursos y evitar cuellos de botella asociados con recursos específicos.

Intel trabaja en todo el ecosistema (lo que incluye contribuir en proyectos de código abierto y la ingeniería en conjunto con proveedores de hardware y software) para lograr que los clústeres de HPC convergentes funcionen de manera óptima, como se ilustra en la Figura 4. Esto agiliza la adopción para los clientes finales, ayudándolos a mitigar la complejidad y el riesgo. Esto incluye garantizar que los administradores de recursos y otros programas de software para simulación, modelado, inteligencia artificial y análisis trabajen juntos a la perfección y que los marcos estén optimizados para lograr desempeño, escalabilidad, estabilidad y seguridad en la arquitectura Intel. Intel también trabaja con proveedores de servidor para diseñar y validar sistemas que ofrezcan los mejores resultados posibles.

La base de la arquitectura es una infraestructura común de hardware y software basada en módulos Intel® y otros componentes estándar de la industria. Estos incluyen una

Figura 3. Motivación para la convergencia de HPC en la organización.

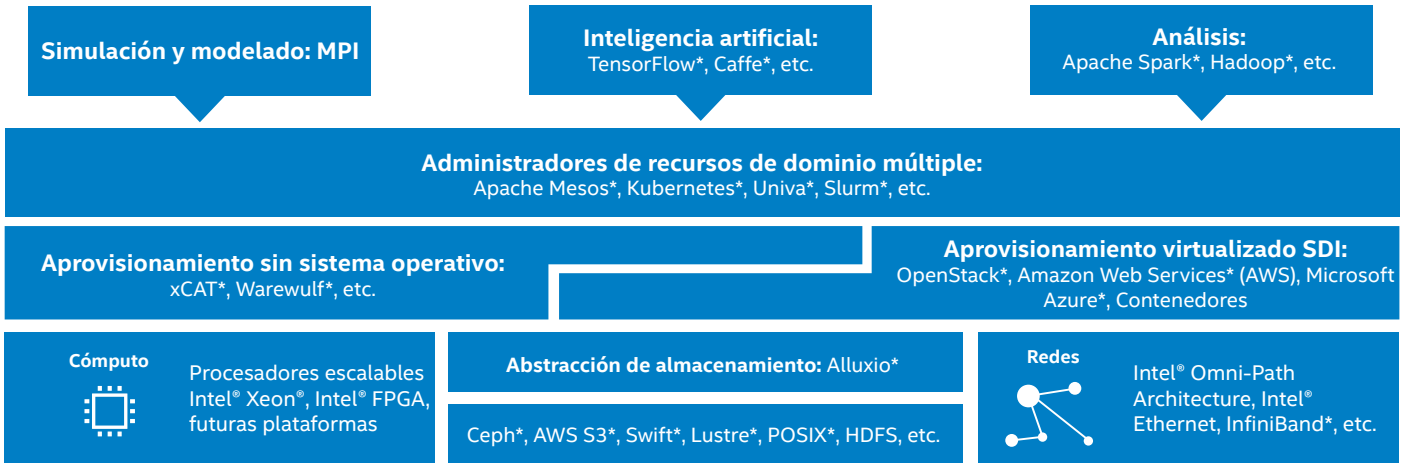

Figura 4. Arquitectura unificada en cargas de trabajo.

gama completa de interconexiones populares y almacenes de objetos, además de varios enfoques a la abstracción de almacenamiento, la cual reúne datos de los tres tipos de cargas de trabajo; lo anterior crea un único conjunto de datos que está disponible en cualquier momento en que se necesite. Por lo tanto, se eliminan el aislamiento de datos en silos y la necesidad de mover y almacenar constantemente grandes conjuntos de datos entre clústeres separados.

La arquitectura de almacenamiento unificado del modelo se basa en un almacén de objetos distribuidos que ofrece una gama de abstracciones de almacenamiento para lograr una interoperabilidad entre enfoques de acceso a datos. La arquitectura de estructura unificada armoniza el comportamiento de diversas estructuras de interconexión.

La arquitectura descrita aquí engloba los enfoques de aprovisionamiento sin sistema operativo y de aprovisionamiento virtualizado para la infraestructura definida por software (SDI). Los enfoques sin sistema operativo son el método estándar utilizado por la mayoría de las organizaciones de HPC, mientras que la SDI se centra en permitir la obtención a pedido de recursos virtuales. Ese enfoque dinámico puede ser fundamental para aumentar la agilidad y generar puestos de trabajo en función de las circunstancias en cambio constante y las necesidades de investigación o de negocios asociadas con cargas de trabajo de análisis.

En muchos sentidos, la capa de administración de recursos es el punto crucial de la plataforma convergente, ya que proporciona la abstracción sólida y eficiente de los recursos utilizados para ejecutar las cargas de trabajo de simulación, modelado, inteligencia artificial y análisis en el mismo clúster. Esta capacidad permite la integración de funciones como la asignación de recursos, la planificación de trabajos y la resolución de problemas de contención entre cargas de trabajo que, de lo contrario, serían incompatibles. Los complementos específicos del dominio en esta capa permiten que los operadores de clúster puedan adaptar el funcionamiento del entorno a sus necesidades.

Esta pila generalizada permite que las organizaciones puedan descubrir, de forma más fácil y sencilla, nuevas implementaciones de simulación, modelado, inteligencia artificial y análisis, especialmente en relación con la forma en que las capacidades de los tres se cruzan para impulsar un nuevo valor comercial y de investigación.

Desarrollo en una amplia base de tecnologías Intel®

En lugar de prescribir una solución rígida, el enfoque de arquitectura de Intel para trabajos de simulación, modelado, inteligencia artificial y análisis incluye la apertura y la flexibilidad como requisitos principales de diseño. Además de admitir módulos de hardware provenientes de los fabricantes más populares, la pila se basa en un vasto ecosistema de software. Junto con las herramientas proporcionadas directamente por Intel; muchos paquetes de software comercial y de código abierto de terceros están optimizados para obtener desempeño, escalabilidad, estabilidad y seguridad en una arquitectura Intel.

Existen combinaciones previamente validadas de módulos de hardware y software, adaptados a necesidades comerciales específicas, que están disponibles como soluciones Intel® Select, como se ilustra en la Figura 5. Esta infraestructura contribuye a acelerar el desempeño, al mismo tiempo que simplifica la implementación y reduce el riesgo para los clientes finales asociado con la modernización del centro de datos; para ello, utiliza sistemas disponibles desde una amplia variedad de fabricantes populares de servidores.

Módulos de arquitectura Intel®

Las arquitecturas de clúster se basan en una variedad de módulos de arquitectura Intel, en toda la pila de hardware, entre los que se incluyen los siguientes ejemplos:

- **Los procesadores escalables Intel® Xeon® de 2da generación** son el núcleo de los clústeres informáticos sólidos que potencian excelentes resultados en las cargas de trabajo. Por ejemplo, en pruebas recientes, se logró un promedio de desempeño hasta 3,7 veces mejor en análisis de desempeño de CPU en HPC en comparación con un sistema de tres años de antigüedad.[4] También se obtuvo un desempeño de clase mundial que es 5,8 veces mejor en el análisis de desempeño de CPU LINPACK*[5] y un desempeño de punto flotante por núcleo 1,7 veces mejor con respecto a procesadores de la competencia.[6] Estos procesadores también se diseñaron para satisfacer los requisitos de las cargas de trabajo más exigentes de inferencia de IA. Los

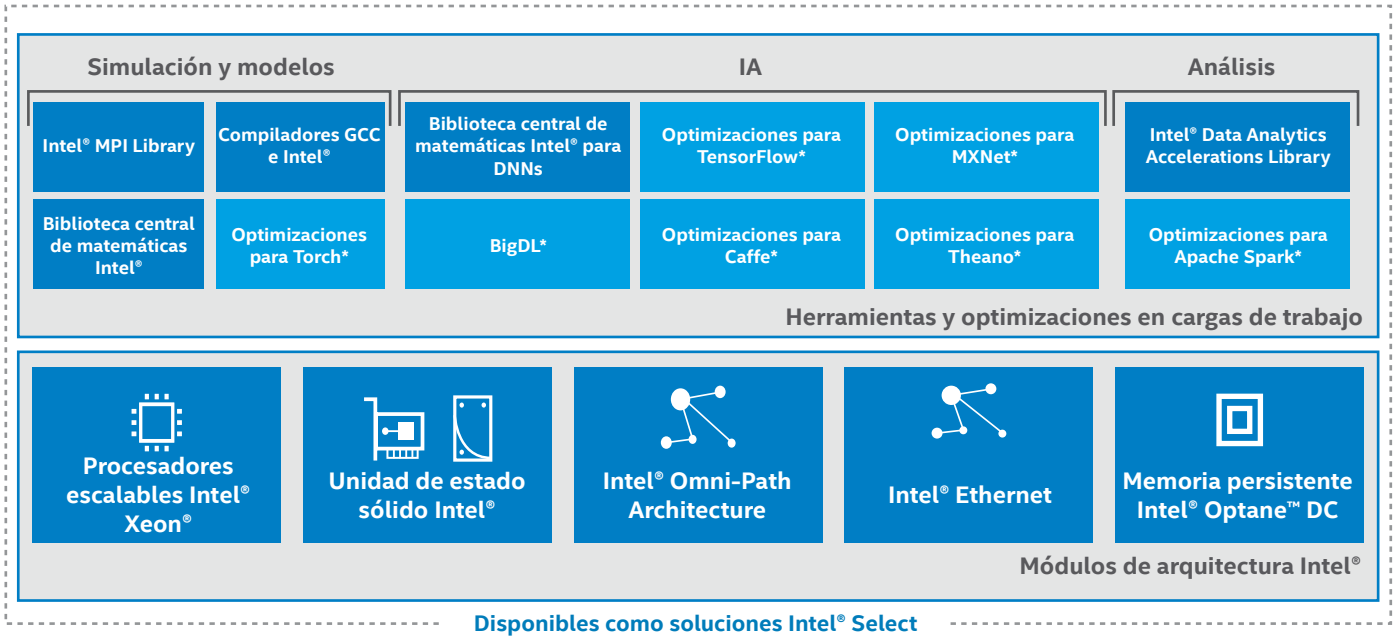

Figura 5. Pila de hardware y software de la arquitectura Intel®.

procesadores escalables Intel Xeon de 2da generación con Intel® Deep Learning Boost han demostrado mejorar el procesamiento de inferencias hasta 25 veces más, en comparación con el procesador Intel® Xeon® Platinum 8180.[7]

- **La memoria persistente Intel® Optane™ DC** ofrece una combinación sin precedentes de alta capacidad, asequibilidad y persistencia. Si el movimiento y mantenimiento de grandes cantidades de datos se realiza más cerca del procesador, la HPC con uso intensivo de datos y las cargas de trabajo de inteligencia artificial se pueden procesar rápidamente y a gran escala.
- **Las unidades de estado sólido Intel®** ofrecen una gama de opciones de almacenamiento que permiten a los clientes crear su propio equilibrio entre costo y desempeño. Las SSD Intel® Optane™ DC ofrecen una latencia mucho menor de forma más consistente, una alta resistencia y un alto desempeño equilibrado que elimina los embotellamientos de SSD NAND para liberar el desempeño del sistema.
- **Los Intel® Fabric Products** incluyen Intel® Omni-Path Architecture (Intel® OPA), una estructura con gran ancho de banda y baja latencia que optimiza el desempeño y facilita la implementación de clústeres HPC, además de Intel® Ethernet, un líder establecido de la industria con una amplia gama de opciones de velocidad, cable medio y conteo de puertos. Ambos están diseñados para funcionar sin problemas en clústeres con otras estructuras como InfiniBand*.

Herramientas y optimizaciones para la arquitectura Intel en cargas de trabajo

Herramientas y optimizaciones enfocadas en simulación y modelado

A continuación, se incluyen algunas herramientas y optimizaciones clave que abordan aplicaciones y cargas de trabajo de simulación y modelado:

- **Intel® MPI Library** es una implementación de la especificación MPICH de código abierto diseñada para crear, ejecutar, probar y mantener aplicaciones optimizadas para clústeres basados en la arquitectura Intel, admitiendo cualquiera de las varias estructuras elegidas en el tiempo de ejecución. Proporciona un entorno de tiempo de ejecución y un kit de desarrollo de software.
- **Los compiladores GCC e Intel®** permiten capacidades de optimización flexible para la arquitectura de Intel; los comportamientos similares de ambos compiladores permiten a los desarrolladores cambiar fácilmente entre los dos. Los compiladores también se integran en cadenas de herramientas populares en común, lo cual mejora más la interoperabilidad.
- **La Biblioteca central de matemáticas Intel®** es una recopilación de rutinas matemáticas previamente optimizadas que se utilizan ampliamente en la informática técnica y científica. Intel mantiene las funciones de Intel MKL para cada generación de plataforma, lo cual permite que los clientes finales puedan aprovechar las mejoras de hardware con solo volver a vincular y compilar el código.
- **Torch*** es un marco de informática científica de código abierto con soporte integrado para muchos algoritmos de aprendizaje automático. Incluye un lenguaje de secuencias de comando simple y sólido (LuaJIT), además de una implementación CUDA/C subyacente. Intel mantiene una variación del proyecto, la cual está optimizada para procesadores Intel Xeon.

Herramientas y optimizaciones enfocadas en análisis

Lo siguiente es una muestra de soluciones (entre otras muchas) que Intel proporciona o habilita para obtener mejores resultados en cargas de trabajo de análisis:

- **Intel® Data Analytics Acceleration Library (Intel® DAAL)** es una colección de rutinas optimizadas para acelerar problemas de análisis de datos a gran escala. Se diseñó para aumentar el desempeño de aplicaciones integradas en plataformas de datos populares, como Hadoop*, Apache Spark*, R* y Matlab*.
- **Apache Spark*** es un enfoque de optimización particular para Intel entre los marcos de análisis. Apache Spark* se

desarrolló originalmente en la universidad UC Berkeley; es un motor de procesamiento de datos a gran escala con módulos integrados para transmisiones, SQL, aprendizaje automático y procesamiento gráfico.

Herramientas y optimizaciones enfocadas en IA

El ecosistema para marcos de IA y otros componentes crece a pasos agigantados e Intel participa en esto ayudando a garantizar que las implementaciones de IA se ejecuten de mejor manera en la arquitectura Intel, lo que incluye los siguientes ejemplos:

- **Intel® MKL for Deep Neural Networks (Intel® MKL-DNN)** es una biblioteca de desempeño para aplicaciones de aprendizaje profundo que proporciona módulos con un alto nivel de vectores y subprocesos para implementar redes neurales profundas en sistemas con arquitectura Intel. Obtenga más información: 01.org/mkl-dnn.
- **Los marcos de aprendizaje profundo** como BigDL*, Caffe*, MXNet*, TensorFlow* y Theano* están optimizados para acelerar flujos de trabajo de IA, lo que incluye un desarrollo simplificado de aplicaciones que se benefician del rápido entrenamiento de redes neurales profundas en clústeres y servidores con arquitectura Intel. Obtenga más información: intel.ai/framework-optimizations.

Conclusión

Juntar plataformas de simulación, modelado, inteligencia artificial y análisis es una evolución necesaria para las empresas e instituciones que planean ejecutar los tres tipos de cargas de trabajo en los próximos años. Si admiten los tres tipos de tareas en un único entorno, muchas organizaciones podrán reducir el gasto de capital (CAPEX) y el gasto operativo (OPEX) a medida que abarcan el espectro completo de capacidades informáticas que se necesitarán para apoyar las nuevas necesidades empresariales y de investigación.

Acelerar la innovación de HPC:
www.intel.com/hpc

[1] *Outlook on Artificial Intelligence* in the Enterprise 2018. Narrative Science https://medium.com/@narrativesci/2018-outlook-on-artificial-intelligence-b1b63a7386f4.

[2] Comunicado de prensa, Artificial Intelligence Market to Rise at Spectacular CAGR of 36.10% During 2016-2024, Players in End-use Industries Leverage its Potential to Automate Processes: TMR. MarketWatch, 16 de agosto del 2018, www.marketwatch.com/press-release/artificial-intelligence-market-to-rise-at-spectacular-cagr-of-3610-during-2016-2024-players-in-end-use-industries-leverage-its-potential-to-automate-processes---tmr-2018-08-16.

[3] Informe de analista, The Business Value of Leading-Edge High Performance Computing. Hyperion Research, 2017, https://hpe.lookbookhq.com/c/the-business-value-p?x=1aqP6M.

[4] Media geométrica promedio de STREAM, HPCG, HPL, WRF, OpenFOAM*, LS-Dyna, VASP, NAMD, LAMMPS, Black Scholes y Monte Carlo; la carga de trabajo individual puede variar. Los resultados de desempeño se basan en pruebas realizadas en las fechas indicadas en la configuración y podrían no reflejar todas las actualizaciones de seguridad disponibles públicamente. Consulte la divulgación de configuración para obtener más información. Ningún producto o componente puede ser absolutamente seguro. Es posible que las cargas de trabajo y el software utilizados en las pruebas de desempeño se hayan optimizado para ejecutarse solo con microprocesadores Intel. Las pruebas de desempeño, como SYSmark* y MobileMark*, se miden utilizando sistemas de computación, componentes, software, operaciones y funciones específicos. Cualquier cambio de alguno de estos factores podría provocar una variación en los resultados. Debe consultar otras fuentes de información y pruebas de desempeño para que lo ayuden a evaluar de forma completa sus compras contempladas, incluido el desempeño de ese producto cuando se combina con otros. Para obtener más información, visite www.intel.la/benchmarks. Exención de responsabilidad de OpenFOAM*: Esta oferta no está aprobada ni respaldada por OpenCFD Limited, productor y distribuidor del software OpenFOAM* mediante el sitio www.openfoam.com, y propietario de la marca registrada OpenFOAM* and OpenCFD*. Configuraciones:

Ganancia promedio 3,7 veces mayor c/n procesador Intel® Xeon® Platinum 9242 en comparación con un servidor de tres años de antigüedad: Media geométrica promedio de STREAM, HPCG, HPL, WRF, OpenFOAM*, LS-Dyna, VASP, NAMD, LAMMPS, Black Scholes y Monte Carlo. La carga de trabajo individual puede variar. Procesador Intel Xeon E5-2697 v4: plataforma de referencia Intel con 2 procesadores Intel Xeon E5-2697 v4 (2,3 GHZ, 18C), 8 unidades DDR4-2400 de 16 GB, 1 SSD, Sistema de archivos de clúster: Panasas (124 TB de almacenamiento) firmware v6.3.3.a y Lustre IEEL basado en OPA, BIOS: SE5C610.86B.01.01.0027.071020182329, Microcódigo: 0xb00002e, Servidor Oracle Linux versión 7.6 (compatible con RHEL 7.6) en un kernel 7.5 usando ksplice para revisiones de seguridad, Kernel: 3.10.0-862.14.4.el7.crt1.x86_64, Pila OFED: OPA OFED 10.8 en RH7.5 con Lustre v2.10.4, HBA: 100 Gbps Intel OPA 1 puerto PCIe x16, Conmutador: Intel OPA Edge Switch serie 100 de 48 puertos. OMP STREAM 5.10, Triad, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 128.36. HPCG, Binario incluido MKL 2019U1, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 23.78. HPL 2.1, HT=ACTIVADO, Turbo=DESACTIVADO, 2 subprocesos por núcleo, puntuación: 1204.64. WRF 3.9.1.1, conus-2,5km, HT=ACTIVADO, SMT=ACTIVADO, 1 subproceso por núcleo, puntuación: 4.54. OpenFOAM* 6.0, 42M_cell_motorbike, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 3500. LS-Dyna 9.3-Explicit AVX2 binario, 3car, HT=ACTIVADO, SMT=ACTIVADO, 1 subproceso por núcleo, puntuación: 2814. VASP 5.4.4, CuC, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 384.99. NAMD 2.13, apoa1, HT=ACTIVADO, Turbo=DESACTIVADO, 2 subprocesos por núcleo, puntuación: 4.4. LAMMPS versión 12 de diciembre del 2018, Water, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 54.72. Black Scholes, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 2573.77. Monte Carlo, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 43.2. Procesador Intel Xeon 9242: Plataforma de referencia de Intel con procesadores 2S Intel Xeon 9242 (2,2 GHZ, 48C), 16 unidades DDR4-2933 de 16 GB, 1 SSD, sistema de archivos de clúster: 2.12.0-1 (servidor) 2.11.0-14.1 (cliente), BIOS: PLYXCRB1.86B.0572.D02.1901180818, Microcódigo: 0x4000017, CentOS 7.6, Kernel: 3.10.0-957.5.1.el7.x86_64, Pila OFED: OPA OFED 10.8 en RH7.5 con Lustre v2.10.4, HBA: 100 Gbps Intel OPA 1 puerto PCIe x16, Conmutador: Intel OPA Edge Switch serie 100 de 48 puertos. OMP STREAM 5.10, Triad, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 407. HPCG, Binario incluido MKL 2019U1, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 81.91. HPL 2.1, HT=ACTIVADO, Turbo=DESACTIVADO, 2 subprocesos por núcleo, puntuación: 5314. WRF 3.9.1.1, conus-2,5km, HT=ACTIVADO, SMT=ACTIVADO, 1 subproceso por núcleo, puntuación: 1.44. OpenFOAM* 6.0, 42M_cell_motorbike, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 1106. LS-Dyna 9.3-Explicit AVX2 binario, 3car, HT=ACTIVADO, SMT=ACTIVADO, 1 subproceso por núcleo, puntuación: 768. VASP 5.4.4, CuC, HT=ACTIVADO, Turbo=DESACTIVADO, 1 subproceso por núcleo, puntuación: 133.96. NAMD 2.13, apoa1, HT=ACTIVADO, Turbo=DESACTIVADO, 2 subprocesos por núcleo, puntuación: 19.9. LAMMPS versión 12 de diciembre del 2018, Water, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 276.1. Black Scholes, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 9044.32. Monte Carlo, HT=ACTIVADO, Turbo=ACTIVADO, 2 subprocesos por núcleo, puntuación: 227.62. Exención de responsabilidad de OpenFOAM*: Esta oferta no está aprobada ni respaldada por OpenCFD Limited, productor y distribuidor del software OpenFOAM* mediante el sitio www.openfoam.com, y propietario de la marca registrada OpenFOAM* and OpenCFD*. Fecha de prueba: 15 de marzo del 2019.

[5] Exención de responsabilidad de OpenFOAM*: Esta oferta no está aprobada ni respaldada por OpenCFD Limited, productor y distribuidor del software OpenFOAM* mediante el sitio www.openfoam.com, y propietario de la marca registrada OpenFOAM* and OpenCFD*. Configuraciones: LINPACK: AMD EPYC 7601: Supermicro AS-2023US-TR4 con 2 unidades AMD EPYC 7601 (2,2 Ghz, 32 núcleos), SMT DESHABILITADO, Turbo HABILITADO, BIOS versión 1.1a, 26/4/2018, microcódigo: 0x8001227, 16 unidades DDR4-2666 de 32 GB, 1 SSD, Ubuntu 18.04.1 LTS (4.17.0-041700-generic Retpoline), Linpack de alto desempeño v2.2 compilado con Intel® Parallel Studio XE 2018 para Linux, Intel MPI versión 18.0.0,128, AMD BLIS versión 0.4.0, Configuración de análisis de desempeño: Nb=232, N=168960, P=4, Q=4, Puntuación =1095GF, probado por Intel el 31 de julio del 2018, en comparación con: 1 nodo, 2 cpu Intel® Xeon® Platinum 9282 en Walker Pass con memoria total de 768 GB (24 unidades 2933 de 32 GB); ucode 0x400000A en RHEL7.6, 3.10.0-957.el7.x86_65, IC19u1, AVX512, HT deshabilitado, Turbo habilitado, puntuación=6411, probado por Intel el 16/2/2019.

[6] 1 copia de SPECrate2017_fp_base* entre Intel 8280 de 2 zócalos y AMD EPYC 7601 de 2 zócalos. Intel® Xeon®-SP 8280; plataforma de referencia basada en Intel Xeon con 2 procesadores Intel® Xeon® 8280 (2,7 GHz; 28 núcleos); BIOS versión SE5C620.86B.0D.01.0348.011820191451; 18/1/2019; microcódigo: 0x5000017; HT DESHABILITADO; Turbo HABILITADO; 12 DDR4-2933 de 32 GB; 1 SSD; Red Hat EL 7.6 (3.10.0-957.1.3.el7.x86_64); 1 copia del análisis de referencia SPECrate2017_fp_rate compilada con Compilador Intel 19.0.1.144; -xCORE-AVX512 -ipo -O; ejecutada en 1 núcleo utilizando taskset y numactl en el núcleo 0. Puntuación estimada = 9,6, al 6/2/2019 probado por Intel con mitigaciones de seguridad para las variantes 1,2,3,3a y L1TF.

AMD* EPYC* 7601: Supermicro AS-2023US-TR4 con 2S AMD EPYC* 7601 con 2 procesadores AMD* EPYC* 7601 (2,2 GHz; 32 núcleos); BIOS versión 1.1c; 4/10/2018; SMT desactivada; turbo activado; 16 DDR4-2666 de 32 GB; 1 SSD; Red Hat* EL 7.6 (3.10.0-957.5.1.el7.x86_64); 1 copia del análisis de referencia base SPECrate2017_fp_rate compilado con AOCC versión 1.0 -Ofast; -march = znver1; ejecutado en 1 núcleo utilizando taskset y numactl en el núcleo 0. Puntuación estimada = 5,56 al 8/2/2019 según las pruebas de Intel.

[7] Mejora 5 veces mayor del procesamiento de inferencias en procesador Intel® Xeon® Platinum 9242 con Intel® DL Boost: probado por Intel el 26/2/2019. Plataforma: Dragon rock, 2 zócalos, Intel® Xeon® Platinum 9242 (48 núcleos por zócalo), HT ACTIVADO, Turbo ACTIVADO, memoria total de 768 GB (24 ranuras/32 GB/2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, Kernel Centos 7 3.10.0-957.5.1.el7.x86_64, Marco de aprendizaje profundo: Optimización de Intel® para Caffe versión: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN Versión: V0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), modelo: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, Sin syntheticData de capa de datos: 3x224x224, 48 instancias/2 zócalos, Tipo de datos: INT8 vs prueba realizada por Intel el 11 de julio del 2017: CPU 2S Intel® Xeon® Platinum 8180 @ 2,50 GHz (28 núcleos), HT desactivado, turbo desactivado, regulador de escala definido como "desempeño" a través del controlador intel_pstate, RAM DDR4-2666 ECC de 384 GB. CentOS Linux versión 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC serie S3700 (800 GB, 2,5 pulgadas, SATA de 6 Gb/s, 25 nm, MLC) Desempeño medido con: variables del entorno: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, frecuencia de la CPU establecida con configuración de frecuencia de energía de cpu de desempeño -d 2,5 G -u 3,8 G -g. Caffe: (https://github.com/intel/caffe), revisión f96b759f71b2281835f690af267158b82b150b5c. Inferencia medida con el comando "caffe time --forward_only", formación medida con el comando "caffe time". En el caso de las topologías "ConvNet", se utilizó el conjunto de datos sintético. Para otras topologías, se almacenaron los datos en el almacenamiento local y en la memoria caché antes de la formación. Especificaciones de topología de https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Compilador Intel C++ versión 17.0.2 20170213, bibliotecas pequeñas de Intel MKL versión 2018.0.20170425. Caffe ejecutado con "numactl -l".

El software y las cargas de trabajo utilizados en las pruebas de desempeño podrían haber sido optimizados en términos de desempeño únicamente para microprocesadores Intel®. Las pruebas de desempeño, como SYSmark* y MobileMark*, se miden utilizando sistemas de computación, componentes, software, operaciones y funciones específicos. Cualquier cambio de alguno de estos factores podría provocar una variación en los resultados. Debe consultar otras fuentes de información y pruebas de desempeño para que lo ayuden a evaluar de forma completa sus compras contempladas, incluido el desempeño de ese producto cuando se combina con otros.

Los resultados de desempeño se basan en pruebas realizadas en las fechas indicadas en la configuración y podrían no reflejar todas las actualizaciones de seguridad disponibles públicamente. Consulte la divulgación de configuración para obtener más información. Ningún producto o componente puede ser absolutamente seguro. Es posible que las cargas de trabajo y el software utilizados en las pruebas de desempeño se hayan optimizado para ejecutarse solo con microprocesadores Intel. Las pruebas de desempeño, como SYSmark* y MobileMark*, se miden utilizando sistemas de computación, componentes, software, operaciones y funciones específicos. Cualquier cambio de alguno de estos factores podría provocar una variación en los resultados. Debe consultar otras fuentes de información y pruebas de desempeño para que lo ayuden a evaluar de forma completa sus compras contempladas, incluido el desempeño de ese producto cuando se combina con otros. Para obtener más información, visite www.intel.la/benchmarks.

Intel no ejerce control ni inspección algunos sobre los datos de análisis de desempeño o los sitios web de terceros a los que se hace referencia en este documento. Visite el sitio web citado y confirme si los datos mencionados son exactos. Las características y ventajas de las tecnologías Intel dependen de la configuración del sistema y es posible que se necesite la habilitación de hardware o software, o bien la activación del servicio. El desempeño varía según la configuración del sistema. Ningún sistema de computación puede ser absolutamente seguro. Consulte al fabricante del sistema o al distribuidor minorista. O bien, puede encontrar más información en intel.la.

Las características y ventajas de las tecnologías Intel dependen de la configuración del sistema y es posible que se necesite la habilitación de hardware o software, o bien la activación del servicio. El desempeño varía según la configuración del sistema. Consulte al fabricante del sistema o al distribuidor minorista. O bien, puede encontrar más información en intel.la.