WHITE PAPER

Intel®, Jülich Supercomputing Centre (JSC), Megware

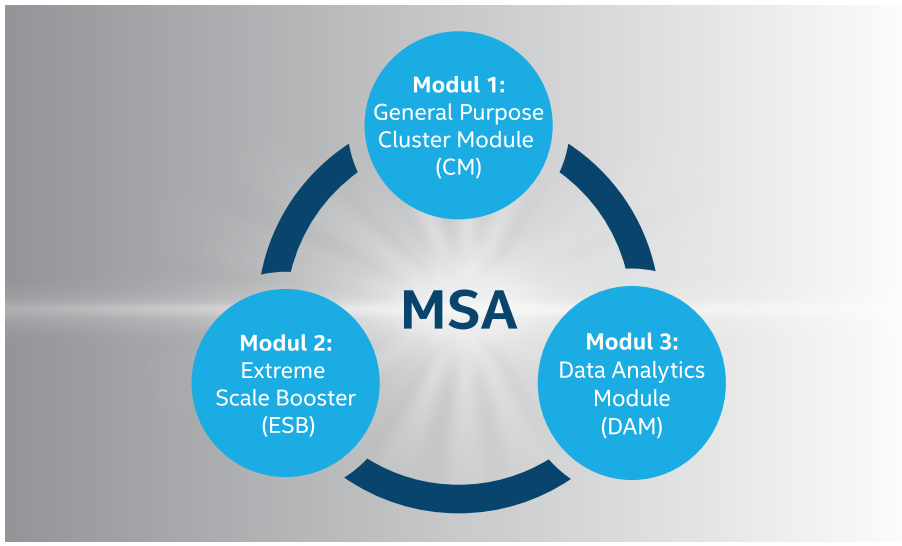
Die Blaupause für den Erfolg

Co-Design-Ansatz zur Entwicklung der Modular Supercomputing Architecture (MSA)

Forschungseinrichtungen wie das Jülich Supercomputing Centre (JSC), eine der führenden High-Performance-Computing (HPC) Forschungsstätten in Europa, definieren mit einem EU-geförderten Projekt die Zukunft des Supercomputing neu. Die HPC-Lösung ist Teil des kollaborativen R&D Projekts DEEP-EST (Dynamical Exascale Entry Platform – Extreme Scale Technologies). DEEP-EST hat ein Budget von 15 Millionen Euro und das Ziel, ein voll funktionsfähiges energieeffizientes Prototyp-System der Modular Supercomputing Architecture (MSA) zu entwickeln und zu bauen.

Die Modular Supercomputing Architecture (MSA)

Das DEEP-EST MSA Prototyp-System besteht aus den drei folgenden Verarbeitungsmodulen: dem General Purpose Cluster Module (CM), dem Extreme Scale Booster (ESB) und dem Data Analytics Module (DAM). Alle Module benutzen Intel® Xeon® Scalable Prozessoren und speziell das DAM verwendet die neue Intel® Optane™ Technologie als DRAM-Alternative und als schnelle SSD sowie spezialisierte Grafikkarten (GPU) und Intel® FPGA Programmable Accelerator Cards (PACs), die sich je nach Anwendung individuell programmieren lassen. In diesem Projekt wurden die Elemente durch den auf HPC spezialisierten Systemintegrator und Intel Partner Megware effizient entworfen und gebaut.



Kurz gesagt liegt der maßgebliche Vorteil des DAM darin, sehr große Mengen an Speicher und Rechenleistung zielgenau für die richtigen Teile der Anwendung zur Verfügung zu stellen. Insbesondere Anwendungen aus Hochleistungs-Datenanalyse (High Performance Data Analytics: HPDA) und Machine Learning (ML)

nutzen große Datenmengen und Speicherarrays, welche sie im Gegensatz zu vielen herkömmlichen HPC-Anwendungen nicht linear, sondern wahlfrei adressieren. Mit dem modularen Ansatz, bei dem die Komponenten durch eine „Network Federation" verbunden sind, ist es nicht notwendig, alle Knoten mit sehr viel Arbeitsspeicher oder mit Beschleunigern wie GPUs und FPGAs aufzurüsten. Das kommt sowohl dem Stromverbrauch als auch den Kosten zugute. Während eine herkömmliche monolithische Architektur alle Anwendungsfälle auf der gleichen Zielhardware berechnet, nutzt ein modularer Supercomputer speziell zugeordnete Knoten und Module mit genau passenden Fähigkeiten, wie eben das DAM für HPDA- und ML-Workloads.

Anforderungsanalyse mit Hilfe eines umfassenden Co-Design-Ansatzes

Das Projekt DEEP-EST startete bereits Mitte 2017 mit umfangreichem Anforderungsmanagement und Technologie-Reviews. Alle drei Module sind fertiggestellt und in Betrieb.

Forschung und Entwicklung im Bereich des High Performance Computing ist Teamarbeit. Bei DEEP-EST sind unter der Projektleitung des Jülich Supercomputing Centre 16 Partner aus der EU involviert. Darunter sind sechs wissenschaftliche Institute, die den Anwendungsteil abdecken, sowie neben Intel auch der Supercomputer-Spezialist Megware, der die Module baut. Weitere Partner sind u.a. Astron, das CERN in Genf, die Universität Edinburgh, das Barcelona Supercomputing Center (BSC) sowie das Leibniz-Rechenzentrum (LRZ) in München.

Die Basis für den Rechenerfolg: General Cluster Module

Das Cluster Module liefert die Basis für die gesamte modulare Supercomputer-Architektur. Ziel war es, möglichst universell Workloads mit zuverlässiger Rechenleistung abzuarbeiten: Jede Anwendung muss auf dem Cluster Module eine ausreichend gute Leistung erzielen können, ohne speziell optimiert oder angepasst zu werden. Insbesondere waren bestmögliche per-thread Leistungen bei Fließkomma- und bei Ganzzahlenberechnungen gefragt, weil sie intensiv von allen Anwendungen, unabhängig von ihrer Parallelisierbarkeit genutzt werden.

Dazu muss das Cluster Module alle üblicherweise in den Anwendungsgebieten verwendeten Programmumgebungen und Software-Stacks effizient bedienen. Während HPC normalerweise eine begrenzte Anzahl von Programmiermodellen und APIs wie MPI, OpenMP, OpenCL und CUDA nutzt, kommen bei Datenanalyse und ML mittlerweile zahlreiche Programmiersprachen und Frameworks zum Einsatz. Die Auswahl reicht von Python über R, Caffe, Tensorflow, Keras und Theano bis hin zu Apache Spark. Das Data Analytics Module wird zwar den größten Teil der Analyseaufgaben übernehmen, trotzdem erwarten die Entwickler, dass das Cluster Module als Entwicklungsplattform für HPDA-Code dienen kann.

In der aktuellen Ausprägung nutzt jeder der 50 CM-Knoten zwei Intel Xeon Gold 6146 Scalable Prozessoren mit jeweils 12 Kernen, 3,2 GHz Taktfrequenz und 24,75 MB L3-Cache. 192 GB DDR4 RAM Speicher und Intel® NVMe PCIe3 SSDs stellen den notwendigen Speicher für die Verarbeitung bereit.

Jeder Clusterknoten kann über vier PCIe Generation 3 Links ca. 1,0 GByte/s auf die angeschlossene SSD schreiben. Anwendungen, die auf mehreren Nodes laufen, profitieren so von einem fast linearen Geschwindigkeitsgewinn für I/O-Operationen. Im Prototyp-System hat Megware jeweils zehn Knoten in einem sechs Höheneinheiten großen Chassis realisiert. Berücksichtigt man den Platz für Kühlung, Strominfrastruktur und die InfiniBand-Switches, passen bis zu 50 Knoten in ein voll ausgebautes, für Energieeffizienz, Warmwasser gekühltes Rack.

Spezialist für Machine Learning: Das Data Analytics Module (DAM)

Das DAM ist speziell für Machine Learning, Künstliche Intelligenz sowie Deep-Learning-Workloads optimiert. Trotz dieser Spezialisierung kommt es mit Standardkomponenten wie GPU- und FPGA-Karten, ohne kostspielige Eigenentwicklungen, aus. Die Intel Optane Technologie ist eine essentielle Komponente des DAM: Zusätzlich zu 384 GB DDR4 RAM verfügt jeder DAM-Knoten über bis zu 3 TB Intel Optane Persistent Memory (PMem). Intel Optane PMem lässt sich in zwei Modi verwenden: Als persistenter, schneller, Byte-adressierbarer Speicher (AppDirect) oder als flüchtige und transparente Erweiterung des installierten DRAM, wodurch Anwendungen mit bis zu 3 TB Arbeitsspeicher mit einer ähnlichen Leistung wie DRAM zur Verfügung stehen. Intel Optane PMem bietet extrem geringe Latenz und ein sehr gutes Preis-Kapazitätsverhältnis – ideale Eigenschaften für den HPC- und HPDA-Bereich.

Darüber hinaus verfügt jeder Knoten über zwei 1,5 TB Intel Optane SSD Data Center Platten, die je nach konkreter

Anwendung als temporäre Benutzerpartitionen, Checkpoint/Neustart-Speicher oder für BeeOND (BeeGFS on-demand paralleles Dateisystem) genutzt werden können.

Das DAM besteht aus 16 Nodes mit je zwei Intel Xeon Platinum 8260M Scalable Prozessoren der 2. Generation (jeweils mit 24 Kernen, 2,4 GHz Taktfrequenz und 35,75 MB L3-Cache) sowie einer 40 GbE Cluster-Fabric für die Anbindung. Ein 100 Gbps EXTOLL Interconnect mit Tourmalet PCIe3 Add-In-Karten rundet die Hardwareausstattung ab. Das DAM wurde dafür ausgelegt, HPDA Codes wie unbeaufsichtigtes Clustering (DBSCAN oder KMeans), überwachtes Clustering mit Support Vector Machines (SVMs), Random Forests und natürlich Deep Learning (DL) auszuführen. Solche Anwendungen werden beispielsweise bei den Co-Design-Workloads zur Bildanalyse der KU Leuven und der University of Iceland eingesetzt.

Schneller und großer Speicher für optimale Machine-Learning-Leistung

Während HPC-Anwendungen Daten vor allem große Matrizen und Arrays sequentiell lesen und verarbeiten, greifen Machine Learning und Data-Analytics-Anwendungen zufällig auf den Speicherinhalt großer Speicher zu. Es werden mehr Operationen auf kleinen Dateneinheiten ausgeführt und kleinere Datentypen genutzt. Das DAM, das FPGAs umfasst, ist sehr wichtig, da es optimal zu diesen spezifischen Anforderungen passt – im Gegensatz zu traditionellen HPC-Architekturen. Ein skalierbares System wie das DAM ist eine energieeffiziente Lösung für parallele Anwendungen, da es die benötigte Speicherkapazität für Workloads bereitstellt, indem es eine passende Anzahl von Knoten zuweist.

I/O-Durchsatz spielt im Bereich von HPDA und ML eine deutlich größere Rolle als bei traditionellen Simulationen. Trotzdem kommt es auf hohe Fließkommaleistung an, um die eingebaute Optimierungen in vielen Machine-Learning-Algorithmen zu nutzen, beispielsweise bei der Co-Design Workload des Partners KU Leuven. Durch das Design des DAM können speicherhungrige Anwendungen aus dem Bereich der Bioinformatik und des maschinellen Lernens effizient ablaufen. Darüber hinaus recycelt der zeitaufwändige Modellierungsprozess von HPDA-Anwendungen häufig große Datensätze. Persistenter Speicher, wie ihn Intel Optane bietet, kann solche, immer wieder angeforderten Datensätze, schneller verfügbar machen, als herkömmliche oder flüchtige Speichersysteme.

Die eingebauten Intel FPGA PAC D5005 (PCIe3 x16) Karten sind frei programmierbar und stellen die einfachste Möglichkeit dar, vorhandene ML-Frameworks zu nutzen. Davon gibt es für Intel FPGA PACs bereits eine große Zahl, sowie optimierte Bibliotheken und Frameworks. Diese vorhandene und etablierte Software beschleunigt die Anwendung von Workloads auf dem DAM. Eigene Anpassungen und Code-Entwicklung beschränken sich auf ein Minimum. Natürlich können die Anwender trotzdem maßgeschneiderte Codes erstellen, wenn die Anwendung nicht durch die vorhandene Software abgedeckt wird. So ist von High-Level-Sprachen wie OpenCL und Data Parallel C++ (als Teil von Intel® oneAPI) bis hin zum sehr hardwarenahen VHDL alles möglich. Die Intel FPGA PACs nutzen den Intel® Acceleration Stack für Intel Xeon CPUs mit FPGAs, das optimierte und vereinfachte Hardwareschnittstellen und Software Application Programming Interfaces (APIs) bietet. Damit sparen Entwickler wertvolle Zeit, so dass sie sich auf die Mehrwerte ihrer Lösung konzentrieren können. In den Workloads, die zum Test der Modular Supercomputing Architecture genutzt werden, sind beide Ansätze – bestehende Frameworks und Programmierung – vertreten.

Zwischenbilanz und weitere Schritte

Auch wenn das Projekt noch in vollem Gange ist, zeigen die bisherigen Erfahrungen bereits, dass der eingeschlagene Weg großes Potential für zukünftige Anwendungen und für andere Forschungseinrichtungen im institutionellen oder im industriellen Bereich hat. So erwarten die Projektpartner für geeignete Workloads eine erhebliche Leistungssteigerung im Vergleich zu monolithischen Systemen. Das trifft insbesondere für Workloads zu, die von der Berechnung auf spezialisierten Modulen profitieren. Alle am Projekt beteiligten Partner erhalten die Mess- und Testergebnisse, so dass an den Institutionen eigene Varianten der Architektur erprobt und weiterentwickelt werden können. Daraus sollen vorgefertigte Softwarepakete entstehen, die bereits einen Großteil der Anpassung für bestimmte Workloads enthalten und noch schneller eingesetzt werden können.

Das JSC wird seine zukünftige Beschaffung am Konzept von DEEP-EST ausrichten. Ab Mitte 2020 wird ein Booster-Modul für das Produktionssystem JUWELS installiert. Die aktuelle DEEP-EST Testplattform hat die Anwender bereits so weit überzeugt, dass klar ist, hier einen sehr guten Blueprint für eine universelle Plattform gefunden zu haben. Die beteiligten Wissenschaftler der Partneruniversitäten und Institute befinden sich zwar noch mitten in der Evaluierung, schon heute zeigt aber die Vielfalt und Aktualität der Anwendungen, dass mit der modularen Supercomputer Architektur die richtigen Weichen für die Zukunft des HPC gestellt wurden.

Highlights der Lösung

- Kombination von Technologien zur Beschleunigung komplexer wissenschaftlicher Simulationen, Datenanalysen und Maschine-Learning-Anwendungen
- Heterogene Anwendungen und Workloads laufen auf exakt zugeschnittenen Compute- und Speicherressourcen
- Intel Xeon Platinum 8260M Prozessoren mit VNNI Instruktionen und Intel FPGA PAC D5005 für Datenanalyse & ML Inferenz
- Intel Optane Persistent Memory als Erweiterung des Hauptspeichers oder als ultraschneller persistenter Datenpuffer sowie Intel Optane SSDs als schneller Datenspeicher
- Möglichkeit, schnell weitere Systemmodule anzukoppeln, zum Beispiel spezialisierte Plattformen für Künstliche Intelligenz
- Volle Nutzung des Intel-Spektrums von Werkzeugen (Compiler, Analyse, NN-Frameworks) und Bibliotheken (Intel® Math Kernel Library, Intel® Data Analytics Acceleration Library)
- Deutliche Verringerung von Zeit- und Energieaufwand bis zur Lösung

Eine der wichtigsten Anforderungen der Modularen Supercomputer Architektur (MSA) ist es, möglichst viele wissenschaftliche Einsatzgebiete effizient zu unterstützen. Mehrere namhafte Partner nutzen den Prototypen am Jülich Supercomputing Centre bereits für Referenzprojekte:

Radioastronomie, Astron

Weltweit beobachten große Antennenfelder das Universum im Bereich der Radiowellen. Die empfangenen Daten der einzelnen Antennen müssen zusammengeführt und analysiert werden. Dabei werden die Daten in einem sequentiellen Prozess bearbeitet. Zwei Schritte sind besonders rechenintensiv: die Kombination (Correlator) der Daten und die Errechnung eines Bildes (Imager). Correlator und Imager wurden für mehrere Plattformen portiert (GPUs, Intel Xeon CPUs, DSP und FPGA). Der MSA-Prototyp wird verwendet, um jeweils die beste Plattform und den energieeffizientesten Lösungsansatz für diese beiden Schritte zu finden.

Weltraumwetter, Katholieke Universiteit Leuven

Die Sonne ist eine große Kugel aus Plasma – ein Pool von Molekülen, der so heiß ist, dass Elektronen aus den Atomkernen herausgerissen werden. Diese elektrisch geladenen Teilchen sind im starken Magnetfeld unserer Sonne gefangen. Die Magnetfelder dehnen sich allmählich in das Sonnensystem aus und tragen das Plasma nach außen. Wenn sie allerdings auf der Sonnenoberfläche zu stark werden, können sie riesige magnetische Eruptionen, sogenannte Flares, erzeugen. Magnetfelder und Plasmaeruptionen können Störungen in elektronischen Geräten verursachen und Satelliten beeinträchtigen. Mit ihrem neuen Simulationsmodell versucht die Universität Leuven, die Auswirkungen der Sonnenaktivität auf unseren Planeten vorherzusagen: Ein Machine-Learning-Code, der auf dem DAM läuft, sagt Wirkungen auf die Magnetosphäre der Erde grob voraus, und ein traditioneller HPC-Code, der auf den anderen Modulen läuft, leistet im Bedarfsfall die detaillierte Simulation und Vorhersage.

Neurowissenschaft, Norwegian University of Life Sciences, Ås

Die NEST-Software wird verwendet, um neuronale Netzwerkmodelle zu simulieren. Sie stellen Netzwerke auf der Ebene vereinfachter Neuronen und Synapsen dar, die durch Spikes – stereotype elektrische Impulse – interagieren. Die Simulationen sind für vielfältige Forschungsthemen relevant: von neuen Lernparadigmen für die Neurorobotik bis hin zu Modellen des tierischen und möglicherweise sogar des menschlichen Gehirns. Bereits NEST allein schöpft die Möglichkeiten herkömmlicher Supercomputer-Architekturen sehr gut aus, hinterlässt den Wissenschaftlern allerdings sehr große Mengen an Ausgabedaten, die weiterverarbeitet und analysiert werden müssen. Mit der MSA können die von NEST auf dem Compute Module erzeugten Daten direkt in das statistische Analysewerkzeug Elephant eingespeist werden, das auf dem DAM läuft. Ziel ist es, parallel zu großen Gehirnsimulationen relevante Signale aus den Rohdaten zu extrahieren.

High-Energy Physik, CERN Genf

Der Large Hadron Collider (LHC) des CERNs ist der weltweit größte und leistungsstärkste Teilchenbeschleuniger. Mit ihm werden Teilchen auf nahezu Lichtgeschwindigkeit beschleunigt und zur Kollision gebracht, wodurch sich riesige Datenmengen erzeugen lassen. Nach dem Zusammenstoß werden die Ergebnisse anhand der Daten von großen Detektoren rekonstruiert. Das DAM hilft dabei, den entstandenen Teilchen ihre respektiven Energien zuzuordnen und durch die Kombination aller Datenpunkte die Kollision zu rekonstruieren. Weil das DAM große Datenmengen ad hoc verarbeiten kann, ist es besonders gut für diese Anwendung geeignet.

Datenanalyse in Geowissenschaften, University of Iceland

Fernerkundung wird in zahlreichen Feldern eingesetzt, unter anderem im Bereich Geowissenschaften, wenn die Landbedeckung der Erde beobachtet werden muss, um beispielsweise Umweltveränderungen zu erfassen. Dieses Vorgehen lässt sich zum Beispiel durch Satellitensensoren oder laserbasiertes LiDAR (Light Detection and Ranging) umsetzen. Die University of Iceland setzt maschinelles Lernen ein, um die Klassifizierung der Landbedeckung anhand von Satellitenbildern zu automatisieren und um LiDAR-Punktwolken zu einzelnen Objekten zu clustern. Das Training von neuronalen Netzen und Support Vector Machines profitiert von den schnellen und energieeffizienten Beschleunigern und dem großen Speicher des DAM.

Weitere Informationen über Intel-Lösungen für HPC und KI finden Sie unter Intel.de/hpc.

Intel Technologien erfordern möglicherweise eine Hardware-, Software- oder Dienstaktivierung.

Kein Produkt oder Bauteil kann absolut sicher sein.

Ihre Kosten und Ergebnisse können abweichen.

© Intel Corporation. Intel, Intel Core das Intel Logo und andere Intel Produkte sind Marken der Intel Corporation oder ihrer Tochtergesellschaften.
Andere Marken oder Produktnamen sind Eigentum der jeweiligen Inhaber.

OpenCL und das OpenCL-Logo sind Marken von Apple Inc., die mit Genehmigung von Khronos verwendet werden.