Der Sweetspot des modernen Enterprise Computing

Research von:

Peter Rutten
Research Director, Infrastructure Systems, Platforms and Technologies Group, Performance Intensive Computing Solutions
Global Research Lead, IDC

September 2021 | IDC Dok. Nr. US48197921

Navigation im Whitepaper

Klicken Sie auf eine Überschrift oder eine Seitenzahl, um zu den einzelnen Abschnitten zu gelangen.

Die Einschätzung von IDC

Die IT-Landschaft gleicht heutzutage bisweilen der Quadratur des Kreises. Beim Bestreben, ein digitales Unternehmen zu werden, und im ständigen Bemühen, die Bedürfnisse immer anspruchsvollerer Kunden erfüllen, stehen Firmen unter dem Druck, schier Unmögliches zu leisten.

- Märkte können sich von einem Moment auf den anderen ändern und dadurch Spitzenwerte in beiden Richtungen verursachen. Diese Volatilität ist keine Ausnahmeerscheinung mehr. Volatilität ist heute der Standard.
- Systeme müssen das Auf und Ab bei der Workloadnachfrage mühelos dynamisch auffangen können, ohne dass dazu ein überdimensioniertes, kostspieliges und energieintensives Rechenzentrum benötigt wird, das nur zu Spitzenzeiten ausgelastet ist. Nachhaltigkeit ist kein leeres Schlagwort mehr, das nur der Außendarstellung dient.
- Bei Analyse und Nutzung von Märkten können wir uns darüber hinaus nicht mehr auf herkömmliche Intelligenz und Erfahrung allein verlassen. Mittlerweile muss in großem Umfang künstliche Intelligenz (KI) hinzugezogen werden, die in Echtzeit zahllose Variablen und immense Datenmengen verarbeiten kann. Künstliche Intelligenz wird sich in sämtlichen Bereichen immer mehr durchsetzen und erfordert speziell dafür ausgelegte Hardwarefunktionalität.
- Angesichts der Forderung nach ununterbrochener Verfügbarkeit können die Workloads, die die Grundlage des digitalen Unternehmens bilden, nicht verlangsamt oder ausgebremst, geschweige denn vollständig unterbrochen werden. Im heutigen rund um die Uhr laufenden Geschäft kann jede Ausfallzeit einer Katastrophe gleichkommen.
- Da im digitalen Unternehmen alles digital und vernetzt sein muss, ist alles ständig neuartigen Angriffen und Bedrohungen ausgesetzt. Cyberkriminelle führen mit einem gewaltigen Arsenal an Cybertools und Angriffsstrategien einen organisierten und dauerhaften Krieg gegen Unternehmen auf der ganzen Welt. Robuste, umfassende Sicherheit muss deshalb jetzt die Grundlage für alle Bereiche bilden.

Ausgehend von dieser Prämisse muss eine IT-Plattform der Enterprise-Klasse, die als Motor des digitalen Unternehmens dienen soll, unter allen Umständen sicher, zuverlässig, skalierbar, nachhaltig, im Rahmen eines Hybridansatzes mit der Cloud integrierbar und für KI ausgelegt sein. In diesem Whitepaper werden diese Aspekte aus der Infrastruktur- und Implementierungsperspektive näher betrachtet und es wird untersucht, inwieweit der neue IBM Power10-Prozessor und die neue IBM Power-Plattform der Enterprise-Klasse E1080 den Anforderungen gerecht werden.

Die Situation im Überblick

Nach Ansicht von IDC sind die folgenden Punkte für den Erfolg eines digitalen Unternehmens in den komplexen und anspruchsvollen Märkten unserer Zeit von zentraler Bedeutung:

- **Security Sicherheit als unabdingbare Voraussetzung**
- **Das Zuverlässigkeitsmandat**
- **Skalierbarkeit und Nachhaltigkeit**
- **Die richtige hybride IT-Infrastruktur (Hybrid Cloud und cloudnative Anwendungen)**
- **Künstliche Intelligenz - Stellenwert und Implementierungsvarianten**

In den nachfolgenden Abschnitten werden diese Punkte näher ausgeführt.

Sicherheit als unabdingbare Voraussetzung

Sicherheit hat sich für das digitale Unternehmen zur wichtigsten Anforderung überhaupt entwickelt. Wenn Unternehmen von IDC nach ihren Prioritäten befragt werden, wird Sicherheit ausnahmslos an erster Stelle oder zumindest auf den ersten Plätzen genannt. So wird von Unternehmen bei der Frage, welche KI-Infrastrukturelemente im Angebot ihrer Server- und Speicherlieferanten zu wünschen übrig lassen, wenig überraschend Sicherheit am häufigsten genannt: 30 % gaben an, dass die Sicherheitsfeatures sie nicht zufriedenstellen.[1]

Diese Unzufriedenheit in Bezug auf die Sicherheit wird auch dadurch deutlich, dass viele Unternehmen die Verwendung durch andere Workloads nicht bei Speichereinheiten zulassen, die Daten für ihre KI-Workloads enthalten. Begründet wird diese Maßnahme am häufigsten (45 %) mit dem Zugriffs- und Datenschutz. Darüber hinaus hat eine IDC-Studie ergeben, dass Sicherheit ein Hauptanliegen bei Public Cloud IaaS (Infrastructure as a Service) ist: 37 % der Unternehmen gaben an, dass ihr Hauptaugenmerk bei derartigen Implementierungen auf der Sicherheit liegt.[2] Mehr als bei allen anderen Workloads nutzen Unternehmen KI in ihren sicherheitsbezogenen Workloads, um Verstöße besser vorhersagen und vermeiden zu können.

Zurzeit gilt die Aufmerksamkeit vor allem der Sicherheit von Anwendungen und Netzwerkstacks, und in diesem Bereich wird auch am meisten investiert. Eine Vielzahl von Angriffen richtet sich jedoch auf Low-Level-Funktionen und auf die Hardware. Angreifer machen sich oft Sicherheitslücken in den Prozessoren und/oder im Mikrocode als Einfallstor zunutze. Diese Angriffe sind raffiniert und schwer zu erkennen.

IDC beobachtet daher bei Unternehmen ein zunehmendes Interesse am Einsatz von Confidential Computing für die geschäftskritischen Plattformen. Confidential Computing ermöglicht die Isolierung sensibler Daten für die Verarbeitung über ein designiertes und geschütztes Prozessorsubsystem (auch als „sichere Enklave“ bekannt). Heutzutage werden Daten zwar oft im „ruhenden“ Zustand im Speicher (Data at Rest) und bei der Übertragung (Data in Transit) verschlüsselt, jedoch nicht bei der Verwendung im Hauptspeicher (Data in Use). Die Möglichkeiten zum Schutz von

[1] Quelle: IDC AI Infrastructure View 2021

[2] Quelle: IDC IaaSView 2020

Daten und Code im Hauptspeicher sind bei vielen Plattformen eingeschränkt. Dennoch können es sich Unternehmen, die mit sensiblen Daten wie personenbezogenen Daten (PII), Kontodaten oder Gesundheitsdaten arbeiten, nicht leisten, auf Vorkehrungen gegen Angriffe zu verzichten, die auf die Anwendung bzw. die Daten im Hauptspeicher abzielen.

Beim Confidential Computing sind die auf Hardwareebene verschlüsselten Inhalte des Subsystems nur für autorisierten Code innerhalb eines Programms zugänglich. Ein Zugriff auf die Inhalte ist von außen nicht möglich, weder über anderen Code noch über andere Systeme oder Operatoren. Unbefugte Instanzen können die Daten und den Ausführungsprozess des autorisierten Codes weder einsehen noch manipulieren. Eine umfassende Confidential Computing-Lösung schützt Data in Use ebenso wie Data at Rest. Dies kann durch eine Verschlüsselung von Inhalten in flüchtigem oder nicht flüchtigem Systemspeicher und persistenten Datenspeichern, entweder auf Flashmedien oder rotierenden Medien, erreicht werden.

In modernen Confidential Computing-Infrastrukturen – insbesondere Infrastrukturen in gemeinsam genutzten Multi-Tenant-Umgebungen – werden separate Koprozessoren zur Auslagerung von priviligierten Prozessoroperationen verwendet, die durch Sicherheitslücken bei der Ausführung von Low-Level-Code kompromittiert werden können. Dieser Ansatz ist noch relativ neu, für zentrale Workloads großer Unternehmen jedoch sehr vielversprechend. Bis er sich durchsetzt, werden Unternehmen verschiedene hard- und softwarebasierte Sicherheitsstrategien kombinieren.

Das Zuverlässigkeitsmandat

Während Sicherheitsstrategien für den Schutz von Daten, Anwendungen und Hardware vor Angriffen von entscheidender Bedeutung sind, ist ein weiterer maßgeblicher Aspekt des digitalen Unternehmens die uneingeschränkte Zuverlässigkeit der IT-Umgebung, einschließlich der Infrastruktur. Hohe Verfügbarkeit ist kein neues Konzept mehr, und Unternehmen können mittlerweile auf Plattformen mit bis zu 99,999 % und Speicherplattformen mit 99,99999 % Verfügbarkeit zurückgreifen. Diese Werte werden jedoch nur mit der richtigen Hard- und Software und den entsprechenden Richtlinien erreicht. Auf dem Servermarkt hat IDC lediglich neun Serverplattformen von sechs Anbietern der Kategorie AL4[3] zugeordnet, die für höchste Verfügbarkeit und uneingeschränkte Fehlertoleranz steht.

- IDC-Untersuchungen[4] zufolge handelt es sich bei den drei Hauptursachen für Anwendungsausfallzeit um Netzfehler (16,2 %), Serverfehler (15,5 %) und Malware (10,3 %). Zu den gängigsten Ursachen von Serverausfällen gehören die Überlastung von Speicher (DRAM) oder CPUs sowie Speicherfehler und -defekte.
- Der Zuwachs an Transaktionsvolumen hat dramatische Ausmaße angenommen, und Unternehmen sind auf immer höhere Transaktionsgeschwindigkeiten angewiesen, um ihre Kunden zufriedenzustellen.
- Betriebs- und geschäftskritische Workloads nehmen zu, und Funktionen zur Unterstützung des Geschäftsbetriebs, die bisher auf einer niedrigen Verfügbarkeitsstufe laufen konnten – z. B. über Virtualisierung oder Clustering –, gelten zunehmend als geschäftskritisch.
- Die Kosten für Ausfallzeiten steigen an, da Unternehmen bei ihrem Tagesgeschäft immer mehr auf ihre Infrastruktur angewiesen sind. Untersuchungen von IDC zufolge betragen die Kosten für Ausfallzeiten bei 20,7 % der Unternehmen 5.000 bis 10.000 USD pro Stunde, bei 18,4 % 10.000 bis 25.000 USD, bei 17 % 25.000 bis 100.000 USD und bei einigen Unternehmen sind es (1,4 %) 500.000 USD pro Stunde.
- Seit geregelte Geschäftszeiten keine Rolle mehr spielen und von Unternehmensanwendungen die Verfügbarkeit für Kunden rund um die Uhr erwartet wird, ist die für den Betrieb dieser Anwendungen erforderliche Infrastruktur einem enormen Druck ausgesetzt. Geplante und außerplanmäßige Ausfallzeiten sind, wenn überhaupt, dann nur in minimalem Umfang möglich.
- Die Toleranz für Ausfälle, Verzögerungen, Datenverluste und Datenbeschädigungen ist bei den Unternehmen wie bei den Kunden gleich Null. Datenschutzverstöße und Fehler können für den Ruf eines Unternehmens fatal sein.

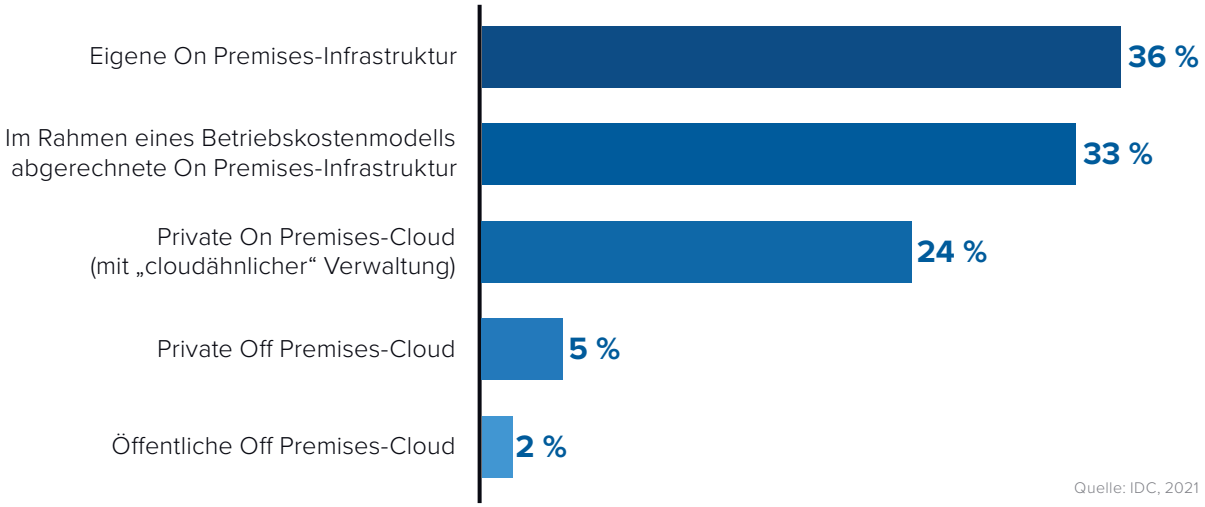
[3] Quelle: IDC Worldwide AL4 Server Market Shares, 2019: *Fault-Tolerant Systems Become Digital Transformation Platforms*

[4] Quelle: IDC *Server Storage Infrastructure Availability Survey,* 2018

- Da Unternehmen immer häufiger und über zunehmend unterschiedliche Wege digital mit Kunden, Bürgern und anderen Unternehmen interagieren, ist die Einhaltung von nationalen und internationalen Vorschriften zu Datenverfügbarkeit, Sicherheit und Datenschutz von höchster Bedeutung.
- Auch wenn sich Verfügbarkeit und Sicherheit in der öffentlichen Cloud deutlich verbessert haben, wird eine echte Fehlertoleranz nach wie vor On Premises- oder Hybrid-Cloud-Lösungen zugeschrieben, nicht der öffentlichen Cloud (siehe **Abb. 1**).

ABB. 1

Infrastruktur für die höchste Verfügbarkeitsstufe

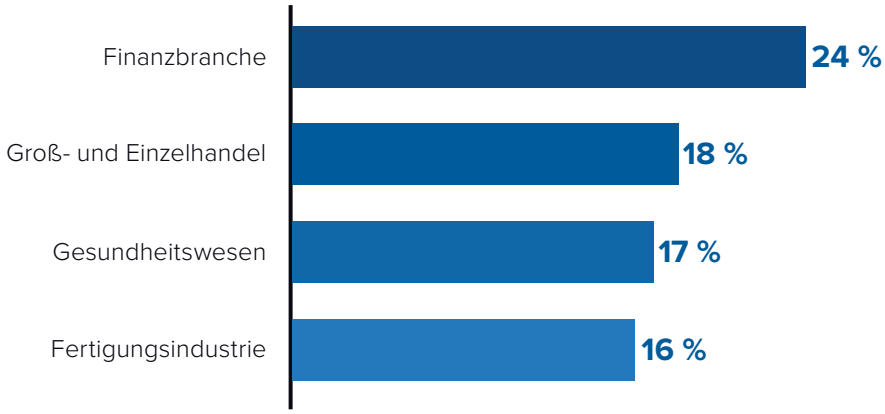


Quelle: IDC, 2021

Der Prozentsatz der Systeme, die eine hohe Verfügbarkeit aufweisen müssen, nimmt infolgedessen zu. In allen Branchen betreiben über 60 % der Unternehmen 21–30 % ihrer gesamten Server in der höchsten Verfügbarkeitsstufe. Aus **Abb. 2** geht der Prozentsatz der Systeme in verschiedenen Branchen hervor, die hoch verfügbar sein müssen.

ABB. 2

Prozentsatz der Systeme nach Branche, bei denen Hochverfügbarkeit verlangt wird



Quelle: IDC, 2021

Die derzeit am meisten verbreiteten AL4-Plattformen haben bedeutende Fortschritte gemacht und sich zu vollständig integrierten Plattformen im Rechenzentrum entwickelt, die bei der digitalen Transformation eines Unternehmens nicht nur einbezogen werden, sondern diese de facto antreiben. Diese Systeme verarbeiten die geschäftskritischsten und wertvollsten Daten vieler Unternehmen, oft in größerem Umfang als alle sonstigen Datentypen. Unternehmen müssen diese Daten deshalb erschließen und nutzen, wenn sie es mit dem Wandel zum digitalen Unternehmen ernst meinen.

Die Notwendigkeit von Skalierbarkeit und Nachhaltigkeit

Unternehmen müssen Workloads skalieren, die immer größere Datenmengen in immer weitläufigeren IT-Umgebungen verarbeiten.

Gleichzeitig müssen sie schnell auf teils unkalkulierbare Nachfrageschwankungen reagieren können, die mitunter die Form von heftigen Ausschlägen annehmen. All dies verlangt größere Rechenzentren, mehr Ausrüstung, mehr Erneuerung bestehender Ausrüstung und mehr Energie für den Betrieb und gleichzeitig auch für die Kühlung der Anlagen.

KI-Workloads stellen den am schnellsten wachsenden Anteil der Workloads dar, die Daten verarbeiten und die IT-Investitionen der Unternehmen antreiben. Derzeit investieren 21 % der Unternehmen in IT-Technologien für die Parallelverarbeitung, die für Training und Inferenzierung bei Deep Learning-Netzwerken erforderlich ist, und weitere 9 % planen derartige Investitionen für 2021. Weiters investieren 46 % der Unternehmen in Beschleunigungstechnologien für Workloads wie GPUs, FPGAs und ASICs, und weitere 7 % planen diese Investitionen für 2021.[5] Vor allem ASICs führen in Rechenzentren bereits heute zu Problemen hinsichtlich Energieverbrauch und Kühlung. Der häufigste Anwendungsfall für Beschleunigung ist Deep Learning-Inferenzierung (Verwendung eines mit einem DNN [Deep Neural Network] entwickelten KI-Modells). Zurzeit nutzen 38 % der Unternehmen Beschleunigung für KI-Inferenzierung, während nur 27 % Beschleunigung für das Training eines DNN verwenden.[6] Die Entwicklung, dass die IT-Investitionen für KI-Inferenzierung die Investitionen in KI-Training übersteigen, wurde erwartet. KI ist dabei nicht die einzige Workload, die zu höheren Investitionen in eine Beschleunigung mit derartigen Koprozessoren führt. Datenanalyse, High-Performance Computing, Finanzplanung, Cybersicherheit und Betrugserkennung sowie Finanzhandel sind weitere Beispiele für Workloads, die zunehmend auf GPUs, FPGAs oder ASICs zurückgreifen, wobei diese Workloads bei der Mehrzahl der Unternehmen on Premises betrieben werden.

21 % der Unternehmen investieren derzeit laut eigener Auskunft in IT-Technologien, die die für Training und Inferenzierung in Deep Learning-Netzwerken benötigte Parallelverarbeitung ermöglichen.

Ein großes Problem ist jedoch, dass die meisten Rechenzentren nicht die Wattstunden und Kühlkapazität bereitstellen können, die durch den erhöhten Energiebedarf und die höhere Wärmeentwicklung beschleunigter Server beim Betrieb mehrerer Racks mit beschleunigten Rechenknoten benötigt werden. Nach Angaben des US-Energieministeriums (2020) gehören Rechenzentren zu den energiezehrendsten Gebäudetypen und verbrauchen 10- bis 50-mal mehr Energie pro Fläche als ein normales Geschäfts- oder Bürogebäude. Untersuchungen von IDC haben ergeben, dass 17,6 % des Betriebsbudgets von Rechenzentren für Elektrizität ausgegeben werden, mehr als für jeden anderen Haushaltsposten. In den Vereinigten Staaten sind Rechenzentren für 2 % des gesamten Stromverbrauchs im gewerblichen Bereich verantwortlich.

Gleichzeitig sind jedoch viele Unternehmen, insbesondere in der Tech-Branche, bestrebt, ihre CO2-Bilanz zu verbessern. Tech-Firmen führen die Liste grüner Unternehmen der amerikanischen Umweltbehörde EPA an, und auch IDC hat im Tech-Bereich enorme Investitionen in erneuerbare Energien sowie Investitionen in energiesparende Hard- und Software beobachtet, die zu einer Senkung des Energieverbrauchs beitragen. IDC stellte fest, dass der Energieverbrauch durch diese Bestrebungen im Schnitt um 26 % gesenkt werden konnte.

[5] Quelle: IDC IT Infrastructure Plans for 2021 Survey, 2020

[6] Quelle: IDC IT Infrastructure for Compute Survey, 2021

Viele Unternehmen haben sich von Cloud-Service-Providern zu mehr Nachhaltigkeit im IT-Bereich animieren lassen, vor allem durch Wiederverwendung und Recycling der Geräte. 33 % der Befragten gaben bei einer IDC-Umfrage[7] an, dass dies ihrer Meinung nach eine Rolle beim Erreichen einer höheren Nachhaltigkeit spielt. Wiederverwendung und Recycling der Geräte kann die CO2-Bilanz eines Rechenzentrums tatsächlich erheblich verbessern. Es mag zwar Gründe geben, bestimmte Komponenten eines Servers zu aktualisieren, jedoch überschreitet beim Übergang zu einer neuen Servergeneration die Menge der unverzichtbaren neuen Serverkomponenten nicht die Anzahl der Komponenten, die beibehalten und wiederverwendet werden könnten.

Das Bewusstsein für diese Wiederverwendungsmöglichkeiten zum Schutz der Umwelt wächst und IDC prognostiziert, dass bis 2025 90% der G2000-Unternehmen wiederverwendbare Materialien in der Lieferkette für IT-Hardware, Klimaneutralitätsbestrebungen bei den Anlagen ihrer Provider und einen niedrigeren Energieverbrauch zur Vorbedingung für weitere Geschäfte machen werden.[8] Derartige Maßnahmen helfen den Unternehmen gleichzeitig Kosten zu sparen, sei es durch niedrigeren Energieverbrauch oder geringere Hardwareinvestitionen.

Die richtige hybride IT-Infrastruktur

Der Trend zur Hybrid Cloud

Derzeit sind 54 % der Anwendungen von Unternehmen weiterhin lokal implementiert.[9] IDC sieht bei diesem Anteil keine Anzeichen für einen spürbaren Rückgang. Die Unternehmen erwarten nach eigenen Angaben, dass in zwei Jahren weiterhin 52 % ihrer Anwendungen lokal betrieben werden. Von diesen On Premises-Anwendungen werden 56 % in einer privaten Cloud betrieben, ein Wert, der in zwei Jahren voraussichtlich auf 60 % ansteigen wird. Auf die Frage, ob die private Cloud ihren Ansprüchen gerecht wird, sagen 61 % der Unternehmen, dass sie ihre Erwartungen nicht nur erfüllt, sondern sogar übertrifft.

Viele dieser Anwendungen, insbesondere die geschäftskritischen Anwendungen, sind auf komplexe Weise miteinander verbunden. Laut den Unternehmen weisen im Schnitt 49 % ihrer Geschäftsanwendungen gewisse Abhängigkeiten und 27 % komplexe Interdependenzen auf. Heute gelten nur 18 % der gesamten Anwendungen als cloudnativ, d. h. als modulare, unterteilte Mikroservices, die Suiten von unabhängig bereitstellbaren Services darstellen. Dagegen sind 32 % der Anwendungen weiterhin monolithisch. Dies wird sich jedoch sehr schnell ändern. Unternehmen erwarten, dass in zwei Jahren nur noch 21 % der geschäftskritischen Anwendungen monolithisch sein werden, während der Anteil der cloudnativen Geschäftsanwendungen auf 44 % ansteigen wird.

Unternehmen gehen auch davon aus, dass sie verschiedene On Premises- und Off Premises-Clouds einsetzen werden. Dieser oft als Hybrid-Cloud bezeichnete Ansatz wird von IDC als schnell wachsendes Bereitstellungssegment eingeschätzt. Abbildung 3 verdeutlicht, dass das gängigste Cloudszenario heute in der Kombination mehrerer Clouds besteht, um Workloads und Daten zwischen den Clouds verschieben zu können. Bei der Kombination von privater Cloud mit öffentlicher Cloud geben 40 % der Unternehmen an, dass die beiden Implementierungen in ihren Unternehmen miteinander verknüpft werden, also als mehr oder weniger integrierte Hybrid Cloud dienen.

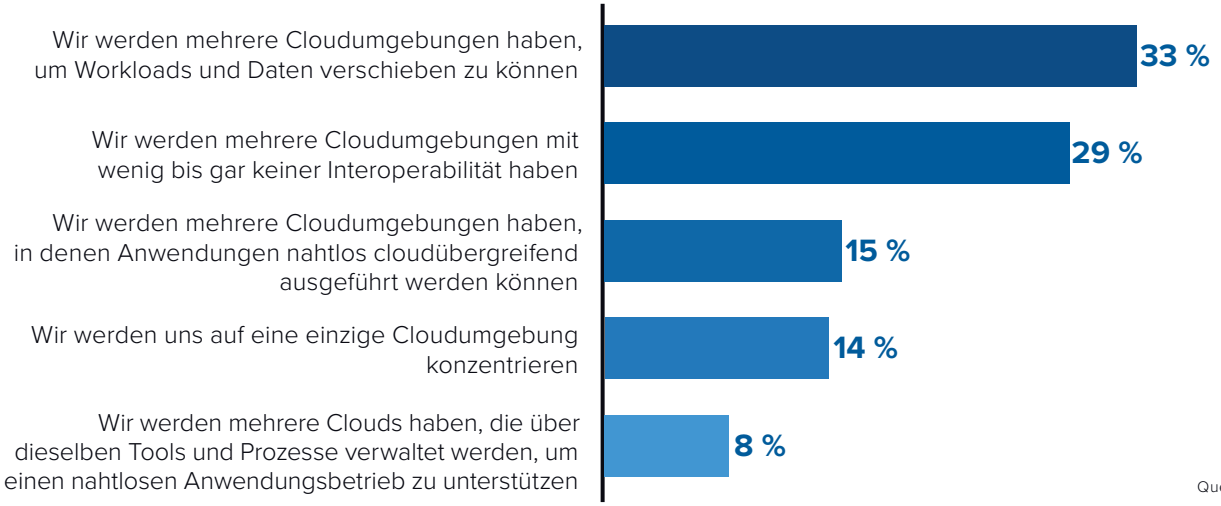
Bemerkenswert ist hierbei, dass Unternehmen den lokalen Teil einer Hybrid Cloud mit großer Mehrheit (84 %) von einem Capex- auf ein Opex-Modell umstellen wollen. Derzeit werden 42 % der IT-Budgets von Unternehmen über ein Opex-Konzept finanziert. Vor drei Jahren lag diese Zahl noch bei 36 %.

Bemerkenswert ist hierbei, dass Unternehmen den lokalen Teil einer Hybrid Cloud mit großer Mehrheit (84 %) von einem Capex- auf ein Opex-Modell umstellen wollen.

[7] Quelle: IDC 2021 Datacenter Operational Survey
[8] Quelle: IDC Worldwide Future of Digital Infrastructure 2021 Predictions
[9] Quelle: IDC 1Q21 Cloud Pulse Survey, Mai 2021

ABB. 3

Nutzung von On Premises- und Off Premises-Cloudumgebungen

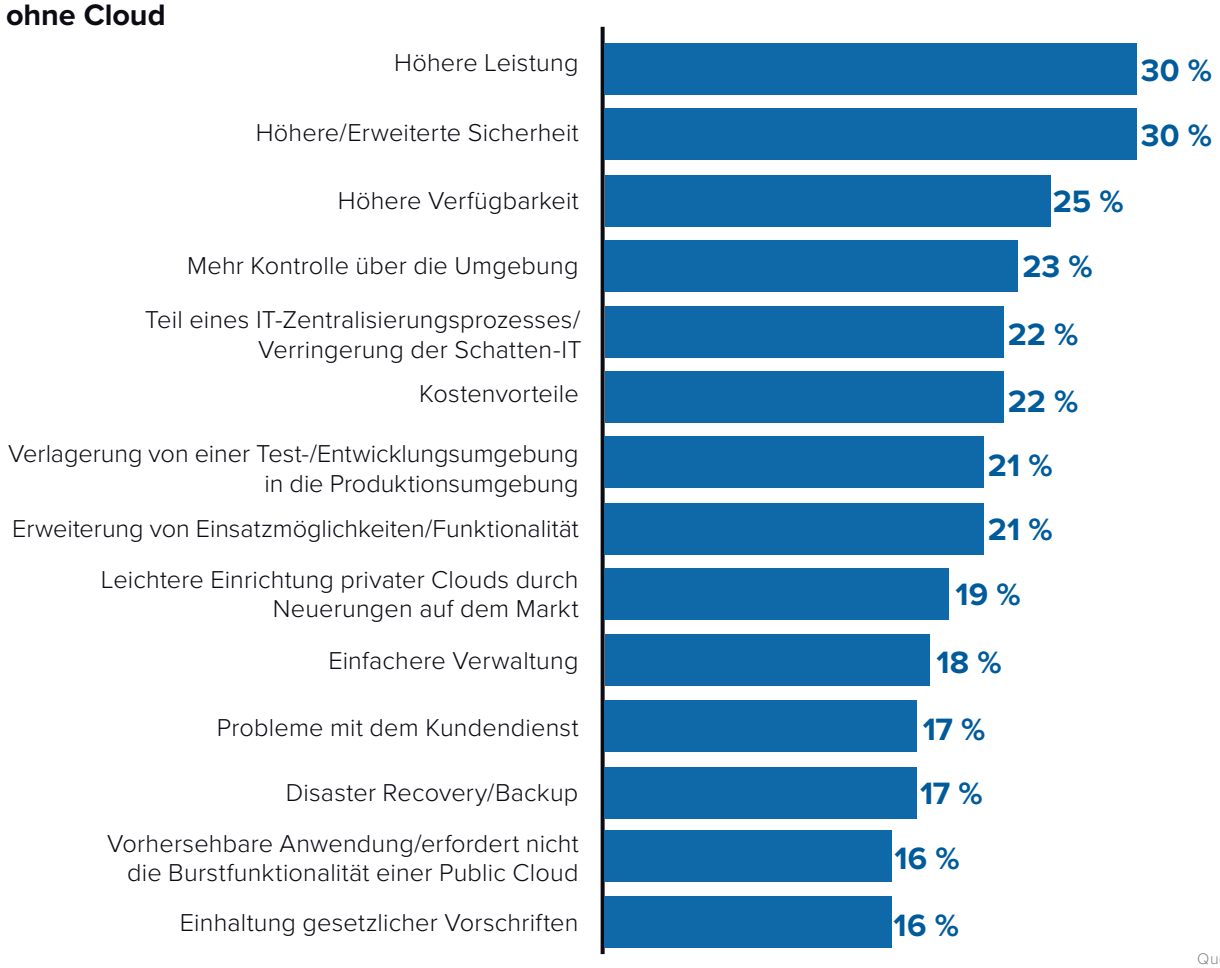


Quelle: IDC, 2021

Da sich die Hybrid Cloud immer mehr durchsetzt, ist auch die Rückkehr von einer öffentlichen Cloud zu einer privaten Cloud verbreitet: 66 % der Unternehmen geben an, dass sie Anwendungen in ihre private Cloud verlagern oder ganz aus der Cloud nehmen. Dies hat unterschiedliche Gründe. Leistung, Zugriffsschutz und Verfügbarkeit stehen jedoch an oberster Stelle (siehe **Abb. 4**).

ABB. 4

Gründe für die Verlagerung von Anwendungen von IaaS in eine private Cloud oder Umgebung ohne Cloud



Quelle: IDC, 2021

Hybrid Cloud und cloudnative Anwendungen

Eine richtig angelegte Hybrid Cloud stellt eine ideale Plattform für Entwicklung und Betrieb cloudnativer Anwendungen dar. Immer mehr Unternehmen betrachten dies als einen entscheidenden Vorteil für ihren digitalen Wandel. Bei ihren Investitionen in eine für Entwicklung und Betrieb von cloudnativen Anwendungen geeignete Cloudstrategie erachtet die Mehrzahl der Unternehmen den Untersuchungen von IDC zufolge die Implementierung bestimmter Funktionalitäten als „wichtig" bis „extrem wichtig" für die Erfüllung ihrer Geschäftsanforderungen. Dabei geht es z. B. um Folgendes:

- Höhere Leistung, Verfügbarkeit, Portierbarkeit und Verwaltung im Anwendungsbereich
- Verbesserte Datenintegration, Orchestrierung, Observability, API-Management und AIOps über mehrere Cloudumgebungen hinweg
- Schnellere Entwicklungszyklen und Markteinführung mit CI/CD (kontinuierliche Entwicklung und Bereitstellung) sowie Automatisierung
- Umfassende Sicherheitsrichtlinien, Risikomanagement, Disaster Recovery-Strategien und Einhaltung gesetzlicher Vorschriften
- Opex-Modell mit Kostenzuordnungsfunktionen anstelle eines Capex-Modells
- Optimierte Mitarbeiterproduktivität, Effizienz und Qualifikationen

Unternehmen, die ihre Investitionen in eine Hybrid Cloud erhöhen wollen, müssen diese Punkte berücksichtigen, wenn sich der erwartete ROI einstellen soll.

Künstliche Intelligenz – Stellenwert und Implementierungsvarianten

IDC erwartet, dass der weltweite Markt für KI-Serverplattformen bis 2025 auf 27 Milliarden US-Dollar anwächst.[10]

Dieses Wachstum wird durch zunehmende Akzeptanz von Dialogtechnologien, Verarbeitung natürlicher Sprache (NLP), Bild- und Videoanalyse, Deep Learning, maschinelles Lernen (ML), Hypothesenbildung und Predictive Analytics befeuert. KI-Serverplattformen werden infolgedessen bis 2025 21 % des gesamten weltweiten Servermarkts ausmachen.

Zuvor wurde bereits der wachsende Bedarf an Koprozessoren thematisiert, der sich sowohl durch KI-Trainingsworkloads als auch durch KI-Inferenzworkloads ergibt. Angesichts der Tatsache, dass die lokale private Cloud das bevorzugte Implementierungsszenario für KI ist und On Premises-Umgebungen ohne Cloud an zweiter Stelle stehen, führt dieser Bedarf in den Unternehmen unmittelbar zu erheblichen Investitionen für zusätzliche GPUs, FPGAs und ASICs. In Bezug auf KI-Training sind diese Investitionen mehr oder weniger unvermeidlich – das Training eines DNN-Algorithmus lässt sich schlichtweg nicht auf einem Hostprozessor durchführen. Bei der KI-Inferenzierung gibt es jedoch viele KI-Modelle, die sehr gut auf einem modernen Hostprozessor oder einem Hostprozessor mit integriertem KI-Spezialprozessor betrieben werden können. Diese Szenarien bieten Unternehmen spürbare Kostenvorteile, denn bereits einige wenige zusätzliche GPUs in einem Server können den Preis des gesamten Pakets ohne Weiteres verdoppeln.

Angesichts der Tatsache, dass die lokale private Cloud das bevorzugte Implementierungsszenario für KI ist und die On Premises-Umgebung ohne Cloud das zweithäufigste, impliziert dies unmittelbar erhebliche Investitionen.

Dies wirft die Frage auf, warum Unternehmen bei ihren KI-Anwendungen weiterhin in erster Linie auf On Premises-Umgebungen setzen. Was spricht dagegen, KI-Anwendungen beispielsweise in der Cloud zu betreiben und ganz auf das Capex-Modell zu verzichten? Natürlich erfolgt das KI-Training in gewissem Umfang in öffentlichen Clouds auf den KI-Plattformen der Provider, und diese Modelle verbleiben nach der Entwicklung gelegentlich als Produktionsworkloads in der Cloud.

[10] IDC Worldwide AI Server Forecast, 2021–2025, Juli 2021

Der wichtigste Aspekt bei der Wahl zwischen Cloud und On Premises sind die Daten. Die folgenden Fragen erläutern die Gründe für die Entscheidung für oder gegen die Cloud.

Welche Daten werden für die Entwicklung des Modells benötigt?

Handelt es sich um Daten aus zentralen Unternehmensanwendungen, z. B. Transaktionsdaten, ist einem Verbleib auf der transaktionsorientierten Plattform der Vorzug zu geben, nicht zuletzt aus Latenzgründen.

Wie sensibel sind diese Daten?

Handelt es sich um sensible und streng zu schützende Daten, wird eine Verlagerung in die Cloud – sei es für Trainings- oder Inferenzzwecke – nicht erwünscht sein.

Welche rechtlichen Rahmenvorschriften für die Verarbeitung von Daten sind zu beachten?

Einige Daten können aus rechtlichen Gründen nicht in eine öffentliche Cloud verlagert werden. Dies dürfte bei zentralen Unternehmensdaten überwiegend der Fall sein. Unternehmen sind an eine Vielzahl gesetzlicher Vorschriften gebunden, von nationalen Datenschutzverordnungen über die DSGVO, branchenspezifische Regelungen wie HIPAA und ISO-Regelungen bis hin zum kalifornischen Verbraucherschutzgesetz.

Was darf mit den Daten gemacht werden, ohne gegen die Vorschriften zu verstoßen, was nicht?

Sobald Daten an verschiedene Stellen übertragen werden, lässt sich die Compliance nur schwer gewährleisten.

Wie umfangreich sind die Daten?

Je mehr Daten für das Training des KI-Modells benötigt werden oder je mehr Daten bei der Inferenzierung einbezogen werden, insbesondere bei echtzeitnaher Inferenzierung, desto schwieriger wird eine Umsetzung in der Cloud.

Wie eng sind die Anwendungen integriert, die die Daten nutzen?

Die für die Transaktionen verwendete Plattform wird aller Wahrscheinlichkeit nach über mehrere Anwendungen verfügen, die für die Durchführung von Analysen und anderen Funktionen eng in die Datenbank integriert sind. Für eine Verlagerung in die Cloud ist dies ungünstig.

Wie teuer ist der Speicher für die Daten?

Die Kosten für umfangreichen Cloudspeicher können die Kapitalaufwendungen, die bei On Premises-Speicher anfallen, schnell übersteigen.

All diese Überlegungen bestärken viele Unternehmen darin, ihre Workloads für KI-Training und -Inferenzierung in On Premises-Umgebungen zu belassen. Das Training wird von den Unternehmen möglicherweise weiterhin hinter einer Firewall in einer separaten IT-Umgebung des Rechenzentrums durchgeführt, das trainierte Modell wird jedoch anschließend zurück auf die Plattform verlagert, auf der die zentralen Unternehmensanwendungen für die Inferenzierung laufen. Wenn die Plattform eine robuste Inferenzierung ermöglicht, können Unternehmen die KI für zentrale Daten verwenden, die früher tabu waren.

IBM Power10 und IBM Power E1080

Um den digitalen Wandel erfolgreich vollziehen zu können benötigen Unternehmen Plattformen, die jede Art von Volatilität auffangen können, kompromisslose Sicherheit bieten, problemlos skalieren und dabei den Flächenbedarf ebenso wie den CO2-Fußabdruck der Unternehmen verringern, ein Höchstmaß an Ausfallsicherheit bieten und Echtzeit-KI für unzählige Transaktionen verarbeiten – alles als Teil einer nahtlosen Hybrid Cloud. Der neue Power10-Prozessor von IBM und die auf Power10 basierende IBM Power-Plattform E1080 bieten eine Reihe von Innovationen, die diese Anforderungen auf eine neue und interessante Weise erfüllen.

Der neue Power10-Prozessor

Die neue IBM Power10-Architektur und der Power10-Prozessor von IBM zeichnen sich durch bedeutsame neue Technologien aus, die für Unternehmen bei rechen- und speicherintensiven Workloads sowie bei Workloads interessant sind, die viel Bandbreite beanspruchen, darunter neue Technologien für eine schnellere KI-Inferenzierung auf dem Chip ohne zusätzliche Hardware basierend auf einem integrierten, speziell dafür ausgelegten MMA (Matrix Math Accelerator).

Im Hinblick auf Sicherheit implementiert Power10 Arbeitsspeicherverschlüsselung ohne Leistungseinbußen (im Gegensatz zu softwarebasierter Arbeitsspeicherverschlüsselung), bietet über Hard- und Software ko-optimierte Containersicherheit für Container-Isolation und beinhaltet Sicherheitseinrichtungen, die der drohenden Gefahr einer Entschlüsselung herkömmlicher Verschlüsselungsschlüssel durch Quantencomputing zuvorkommen.

Die Skalierbarkeit mit Power10 erreicht durch verschiedene Bandbreiteninnovationen ein neues Niveau. IBM hat die PowerAXON-Verbindungstechnologie erweitert und durch OMI (Open Memory Interface) ergänzt, die Übertragungsrate liegt jeweils bei 32 GT/s. Die AXON-Schnittstelle von Power10 verbindet bis zu 16 Sockets zu einem großen, skalierbaren System. Das OMI kommuniziert mit DDR4-DRAM-Speicher über 16 DDR-Ports pro Socket und bietet eine Bandbreite von bis zu 409 GB/s pro Socket. Mit diesen beiden Schnittstellen lassen sich sehr flexible und sogar dem Composable-Prinzip entsprechende Compute-Lösungen bereitstellen.

Power10 ist der erste 7-Nanometer-Prozessor von IBM, der laut IBM im Vergleich zu IBM Power9 eine 3-fach höhere Effizienz in Bezug auf Rechenleistung (Anzahl der Benutzer, Anzahl der Transaktionen) und Energie erreicht.[11] In Verbindung mit dem anhaltenden Fokus von IBM auf die Hybrid Cloud führt dies unmittelbar zu einem geringeren Platzbedarf im Rechenzentrum und deutlich niedrigeren Energieverbrauch. Auf dem Chip befinden sich 15 Prozessorkerne, und Power10 wird über PCI Generation 5 verfügen, einen Standard, der sich in der Branche durchzusetzen beginnt.

IBM Power E1080

IBM Power E1080 ist die erste IBM Plattform der Enterprise-Klasse mit dem Power10-Prozessor. Das System kann mit bis zu 16 Prozessoren ausgestattet werden und ist eindeutig auf die drängendsten IT-Themen von Unternehmen ausgerichtet, die die Anforderungen eines digitalen Unternehmens meistern müssen.

Sicherheit

Um Sicherheitsprobleme dauerhaft und ohne Kompromisse zu lösen, hat IBM eine Verschlüsselung in den Power10-Prozessor integriert. Auf diese Weise können Daten ohne Abstriche bei der Systemleistung verschlüsselt werden. Das System wurde darüber hinaus mit zusätzlichen Sicherheitsfeatures ausgerüstet, um vor Angriffen mit Return-Oriented

[11] Die dreifache Leistung basiert auf technischer Pre-Silicon-Analyse von Integer-, Enterprise- und Gleitkomma-Umgebungen auf einem POWER10 2-Socket-Serverangebot mit 2x30-Kernmodulen im Vergleich zu einem POWER9 2-Socket-Serverangebot mit 2x12-Kernmodulen; beide Module haben das gleiche Energieniveau.

Programming zu schützen, bei denen Angreifer Schadsoftware ungeachtet von Schutzvorrichtungen einschleusen können. Der Power E1080 bietet einen erweiterten Datenschutz mit Transparent Memory Encryption (TME), dem Sicherheitsfeature auf Hardwareebene für Data in Use, das dem Confidential Computing zugrunde liegt, und weist gegenüber dem Vorgänger das Vierfache an Krypto-Beschleunigern auf. Partitionen auf der Plattform sind besser isoliert, und das System ist durch Post-Quanten-Kryptographie (PQC) und homomorphe Verschlüsselung (FHE) vor zukünftigen auf Quantencomputing basierenden Bedrohungen geschützt. Bei FHE ist die Entschlüsselung von im System eingehenden Daten nicht erforderlich, sodass diese Eingaben auch bei einer Nutzung durch nicht vertrauenswürdige Personen nicht offengelegt werden.

Ausfallsicherheit

IDC stuft die Power-Serverfamilie der Enterprise-Klasse als AL4-Systeme ein. Diese höchste Verfügbarkeitsstufe beinhaltet eine uneingeschränkte Fehlertoleranz und damit eine Verfügbarkeit von mindestens 99,999 %. Mit Power10 geht der IBM Power E1080 noch einen Schritt weiter als sein Vorgänger. Mit dem neuen Open Memory Interface wird eine sehr hohe Bandbreite und Zuverlässigkeit, Verfügbarkeit und Wartungsfreundlichkeit (RAS) beim Speicher bereitgestellt. Der Prozessor kann sporadisch auftretende Fehler automatisch erkennen, isolieren und eine Wiederherstellung einleiten, ohne dass der Betrieb dabei unterbrochen oder auf das Fehlermanagement oder die automatische Fehlerbehebung des Betriebssystems zurückgegriffen wird. Darüber hinaus verfügt das System über erweiterte Reparaturfunktionen bei eingeschalteter Einheit wie z. B. SMP-Kabel (Sub Miniature Push-on) für Verbindungen zwischen Knoten, um die Ausfallzeit von Anwendungen zu verringern.

Skalierbarkeit und Nachhaltigkeit

Hinsichtlich Skalierbarkeit und Nachhaltigkeit profitiert der IBM Power E1080 enorm von dem Umstand, dass die Familie der Power-Server außergewöhnlich gut integriert ist, da es sich vom Prozessor über die Firmware bis hin zu Betriebssystem und Hardware bei allen Komponenten um IBM-Komponenten handelt. Die Effizienz der Plattform in Bezug auf Software und OpenShift-Container setzt laut IBM neue Maßstäbe. Infolgedessen erreicht die Plattform mit dem neuen Power10-Prozessor im Vergleich zum Power E980 bei demselben Platz- und Energiebedarf 50 % mehr Leistung.[12] Nach Aussage von IBM bedeutet dies bei derselben Workload außerdem einen um 33 % niedrigeren Energieverbrauch.[13] Die höhere Effizienz ermöglicht es Unternehmen, ihren CO2-Fußabdruck deutlich zu reduzieren und gegebenenfalls Kosten für Hard- und Software durch eine Konsolidierung von Workloads einzusparen.

Hybrid Cloud

Der Power E1080 unterstützt drei Betriebsumgebungen – AIX, IBM i und Linux – auf derselben Plattform und ist dafür konzipiert, Hybrid Cloud-Bereitstellungen für alle drei Betriebsumgebungen zu unterstützen. AIX ist als umfassend modernisiertes UNIX-Betriebssystem von IBM nach wie vor ein Favorit für die hochskalierte Power-Plattform der Enterprise-Klasse. IBM i ist die IBM-Betriebsumgebung, die die Datenbank und andere Unternehmenssoftware in das Betriebssystem integriert und das Plattformmanagement erheblich vereinfacht. Für viele mittlere Unternehmen steht IBM i im Mittelpunkt des operativen Geschäfts. AIX und IBM i sind extrem Open Source-freundlich, unterstützen moderne und beliebte Entwicklersprachen und werden vollständig als Hybrid Cloud betrieben. Wie seine Vorgänger kann auch der Power E1080 mit denselben Sicherheits-, Verfügbarkeits- und Skalierbarkeitsfeatures vollständig oder teilweise unter Linux betrieben werden und bietet Unternehmen dadurch eine Gelegenheit, die transaktionsorientierten und analytischen Workloads auf eine reine Open Source-Plattform zu verlagern.

Dass Unternehmen ihre Power-Plattform der Enterprise-Klasse mit AIX, IBM i und Linux für eine sichere, hochverfügbare, cloudbasierte Workloadmodernisierung nutzen können, ist in erster Linie den folgenden IBM Power-Softwarekomponenten zu verdanken:

IBM PowerVM

- IBM Power Server-Workloads sind virtuell, mobil und vollständig cloudfähig mit PowerVM, das kürzlich mit mehreren neuen Features erweitert wurde, darunter die Datenkomprimierung und -verschlüsselung von LPM (Live Partition Mobility), die eine automatische Verschlüsselung und Komprimierung von Daten bei der unterbrechungsfreien Migration einer aktiven Partition von einem Power-Server auf einen anderen beinhaltet – ein entscheidendes Sicherheits- und Leistungsmerkmal.

IBM PowerVC

- Das Virtualisierungsmanagementtool PowerVC basiert auf OpenStack und vereinfacht das Management virtueller Ressourcen in Power-Umgebungen. Die Software wurde kürzlich funktional erheblich erweitert und beinhaltet z. B. eine neue Export-/Import-Funktionalität für eine gemeinsame Nutzung von VM-Images über mehrere Rechenzentren hinweg.

[12] Informationen von IBM. Basierend auf veröffentlichten rPerf-Ergebnissen für Power E980/12-Kern im Vergleich zu internen rPerf-Messungen von IBM (unter Verwendung derselben Methodik) für Power E1080/15-Kern

[13] Power9 (12 Kern): 5081 rPerf @ 16.520 Watt (0,31 rPerf/Watt), Power10 (15 Kern): 7998 rPerf @ 17.320 Watt (0,46 rPerf/Watt) 0,46 / 0,31 = 1,48 höhere rPerf/Watt

IBM PowerSC

PowerSC ist das Sicherheitsportfolio der Plattform. Die Software vereinfacht das Sicherheits- und Compliance-Management durch verschiedene Funktionen, z. B. Funktionen für die Automatisierung der Compliance, die Erkennung von Malware-Angriffen und das Patchmanagement. Sie wurde mit verschiedenen Funktionen und sogar neuen Produktangeboten erweitert, darunter ein Angebot für Multifaktor-Authentifizierung (MFA), eine weitere bedeutsame Sicherheitsfunktion. Insgesamt wird die Sicherheit auf IBM Power mit AIX durch eine umfassende Lösung erreicht, die den Prozessor, die Firmware, den Hypervisor und die zahllosen Zugriffsschutzfunktionen des Betriebssystems selbst umfasst, um Daten auf allen Ebenen zu schützen.

IBM PowerHA und VM Recovery Manager HA sowie Disaster Recovery

Bei PowerHA handelt es sich um eine Hochverfügbarkeitstechnologie, die eine nahezu kontinuierliche Anwendungsverfügbarkeit ermöglicht und die Servicezuverlässigkeit verbessert. PowerHA gibt mit den Ausschlag dafür, dass IBM Enterprise Power von IDC als fehlertolerant (AL4) eingestuft wird, und wurde mit diversen Features wie erweiterten Failovermetriken und clusterübergreifender Verifizierung (z. B. für den Vergleich zwischen einem Entwicklungs- und einem Testcluster) verbessert. VM Recovery Manager (VMRM) stellt eine vereinfachte HA/DR-Lösung basierend auf VM-Replikation und -Neustart dar, die unabhängig vom verwendeten Betriebssystem einsetzbar ist und Anwendungsüberwachungsagenten für beispielsweise Db2, Oracle und SAP HANA enthält.

Cloud Management Console

Cloud Management Console (CMC) bietet einen umfassenden Einblick in die Leistungs-, Bestands- und Prokollierungsdaten der Power-Infrastruktur, ob On Premises oder Off Premises. CMC wird in IBM Cloud gehostet, sodass Unternehmen auf Pflege und Wartung von Infrastrukturüberwachungssoftware verzichten können, und vereinfacht die Verwaltung von Hybrid Cloud-Implementierungen ebenso wie die Überwachung und Verwaltung der Infrastruktur.

Enterprise Cloud Edition 2.0

Als Ergänzung zu PowerVM vereint Enterprise Cloud Edition alle Schlüsselkomponenten einer vereinfachten Cloud-Managementinfrastruktur, einschließlich PowerSC, MFA, PowerVC, CMC, VMRM und Aspera. Diese Cloudedition ermöglicht eine schnelle Bereitstellung und ein effizientes Management einer privaten Cloud, ein vereinfachtes Sicherheits- und Compliance-Management, eine einfachere Bereitstellung hoher Verfügbarkeit sowie eine beschleunigte Übertragung großer Dateien zwischen Clouds. Enterprise Cloud 2.0 kann mit AIX 7.2 erworben werden.

Red Hat Ansible Automation Platform

Red Hat Ansible Automation Platform ermöglicht eine skalierbare und sichere Automatisierung verschiedener Aspekte der IT-Unternehmensoperationen, einschließlich Ressourcenbereitstellung, Management des Anwendungslebenszyklus und Netzwerkoperationen. Die Plattform besteht aus Ansible Engine, Ansible Tower und Ansible Hosted Services. Alle weiteren Produkte des Red Hat-Portfolios können über Red Hat Ansible Automation Platform integriert werden. Red Hat Ansible Automation Platform ermöglicht mithilfe programmgesteuerter Methoden für Bereitstellung, Verwaltung und Sicherheit von Infrastrukturressourcen Konsistenz im Rechenzentrum.

Red Hat OpenShift

Als zertifizierte Kubernetes-Plattform der Enterprise-Klasse ermöglicht Red Hat OpenShift das Erstellen, Bereitstellen und Verwalten containerisierter Anwendungen (Kubernetes ist eine Container-Orchestrierung). Red Hat OpenShift kann als vollständig verwalteter Service unter verschiedenen Cloud-Providern genutzt oder vom Kunden über Red Hat OpenShift Container Platform oder Red Hat OpenShift Kubernetes Engine verwaltet werden. Die Plattform kann lokal auf Bare-Metal-Servern, Virtualisierungsplattformen (Red Hat Virtualization, VMware oder Red Hat OpenStack) oder über große Cloud-Anbieter wie IBM Cloud, AWS, Google oder Azure bereitgestellt werden. Außerdem kann Red Hat Advanced Cluster Management for Kubernetes dazu eingesetzt werden, mehrere Red Hat OpenShift-Cluster und -Anwendungen von einer einzelnen Konsole aus gemäß integrierter Sicherheitsrichtlinien zu verwalten, und ermöglicht Kunden so die Nutzung einer offenen Hybrid Cloud. Red Hat OpenShift wird von IBM Power-, IBM Z- und x86-basierten Plattformen unterstützt und kann in Verbindung mit AIX, IBM i und Linux eingesetzt werden.

IBM Cloud Paks

Die in Containern angebotenen IBM Cloud Paks werden als beliebte Softwareprodukte zunehmend eingesetzt. Sie sind eng in verschiedene OpenShift Services integriert und können schnell und ohne großen Aufwand auf OpenShift bereitgestellt werden. Anwendungsentwicklern bieten IBM Cloud Paks Tools, Daten und KI-Services sowie quelloffene

Middleware-Software. Sie werden auf der Red Hat OpenShift-Cloudplattform betrieben. Verschiedene Cloud Paks sind in Verbindung mit IBM Power besonders interessant, z. B.:

- ***Cloud Pak for Data:*** Stellt den Kunden KI-Funktionalität bereit und ermöglicht es, mehr Erkenntnisse aus Daten zu gewinnen.
- ***Cloud Pak for Integration:*** Umfasst Integrationstools für Daten, Anwendungsservices und Cloud-Services, um die Integration von Anwendungen, Daten, Cloud-Services und APIs zu unterstützen.
- ***Cloud Pak for Watson AIOps:*** Bietet in Hinblick auf die Verbreitung von Multi-Cloud-Bereitstellungen cloudübergreifende Datensichtbarkeit, Governance und Automatisierung.

Künstliche Intelligenz

Der Power E1080 ermöglicht laut IBM eine um das Zehnfache höhere KI-Inferenzierungsleistung als seine Vorgänger. Er kommt dabei ohne Spezialhardware wie z. B. einen Koprozessor (GPU, FPGA oder ASIC) aus. Die Inferenzierung erfolgt stattdessen auf einem MMA (Matrix Math Accelerator). Um Matrizenoperationen effizient durchführen zu können, ist in jeden Kern des Power10-Chips ein eigener MMA integriert. Diese Operationen sind über eine breite Palette von Datentypen hinweg für verschiedene Genauigkeiten optimiert, die für Deep Learning relevant sind – von doppelter Genauigkeit und einfacher Genauigkeit bis hin zu zwei Typen halber Genauigkeit, einschließlich Bfloat-16 sowie Int-16, Int-8 und Int-4. KI-Inferenzierungsleistung wurde in jede Schicht des Prozessors eingebracht. Der L2-Cache wurde vervierfacht: Die Lade-/Speichereinheiten und SIMD wurden verdoppelt. Infolgedessen kann eine transaktionsorientierte Workload mit eingebetteten KI-Komponenten auf demselben Power10-Prozessor sowohl die Transaktionen als auch die KI-Inferenzierung durchführen, ohne dass es dazu eines Koprozessors bedarf.

Inferenzierung auf dem Chip hat auch den Vorteil, dass alle Sicherheitsfeatures des Prozessors und des Systems zum Schutz der für die Inferenzierung verwendeten Daten zur Verfügung stehen. Außerdem ist die Plattform ONNX-freundlich (Open Neural Network Exchange). ONNX ist ein quelloffenes KI-Ökosystem von Technologie-Unternehmen und Forschungseinrichtungen, die daran arbeiten, offene Standards für die Repräsentation von KI-Algorithmen und Tools zu etablieren, um Innovationen und Zusammenarbeit im KI-Bereich zu fördern. Unternehmen mit IBM Power E1080 können ONNX-Modelle unverändert auf der Plattform einbringen und betreiben und so bei der Inferenzierung von den RAS-Funktionen der Plattform profitieren.

Herausforderungen und Chancen

Für Unternehmen

Unternehmensplattformen für die zentralen transaktionsorientieren und analytischen Workloads eines Unternehmens werden tendenziell in den Rechenzentren als Silos implementiert, auch wenn sie mit umfassenden Features und Technologien konzipiert und erstellt wurden, die genau dies verhindern sollen. Diese Plattformen werden oftmals von erfahrenen IT-Mitarbeitern „geschützt“, die sich mit dem System auskennen, jedoch davor scheuen, die Daten verfügbar zu machen, die Plattform mit der Cloud zu integrieren, Open Source auf der Plattform zu betreiben und KI-Modelle für Echtzeitdaten auszuführen. Für Unternehmen besteht die Herausforderung jetzt darin, mit dieser Kultur der Zurückhaltung so bald wie möglich aufzuräumen. Plattformen der Enterprise-Klasse müssen als die offenen Systeme wahrgenommen werden, die sie sind. Nur so können sie den Unternehmen als Plattformen für die digitale Transformation dienen, die neue Gewinnchancen eröffnen. Gleichzeitig bieten diese Plattformen Unternehmen eine Gelegenheit, sich ernsthaft Nachhaltigkeitsthemen zuzuwenden und den CO2-Fußabdruck zu reduzieren. Das Betreiben von KI auf einer Unternehmensplattform ohne kostenintensive, stromfressende Koprozessoren ist ein zentrales Anliegen, da immer mehr Kernanwendungen mit KI-Funktionalität ausgestattet werden.

Für IBM

Mit der neuen Power-Plattform E1080 fördert IBM weiterhin die Entwicklung von Unternehmen in Richtung Offenheit, Hybrid Cloud, KI und Nachhaltigkeit auf einer ausgesprochen sicheren, leistungsfähigen und zuverlässigen Plattform. IBM ist bekannt dafür, Herausforderungen im Innovationsmanagement mit interessanten neuen Technologien zu begegnen, die in einigen Fällen bahnbrechend und der Konkurrenz voraus sind. Ein Beispiel ist der MMA im Power10-Prozessor. Die größte Herausforderung für IBM sind nicht die Innovationen. Die wirkliche Herausforderung für IBM liegt darin, ein Umdenken bei den Kunden zu erreichen. Die Unternehmensplattformen dürfen nicht mehr als Silo-System oder bestenfalls als vorsichtig zu öffnendes System betrachtet werden, sondern müssen offensiv mit den übrigen Bereichen des Rechenzentrums und mit der Cloud integriert werden, damit das gesamte Leistungsspektrum vollständig ausgeschöpft werden kann. Auf diese Weise können neue Wege eingeschlagen und höhere Einnahmen aus den Kerndaten, die auf der Plattform vorliegen, generiert werden. IBM muss seine Kunden durch Aufklärung, Anreize und ROI-Studien weiter ermuntern, im Umgang mit der Unternehmensplattform Mut und Kreativität an den Tag zu legen.

Fazit

Moderne Unternehmen brauchen Compute-Plattformen, die extreme Marktschwankungen auffangen, robuste Sicherheit bieten, mühelos und umweltschonend skalieren, ein Maximum an Ausfallsicherheit liefern, KI in Echtzeit ermöglichen und als Hybrid Cloud betrieben werden können. Der neue Power10-Prozessor von IBM und die auf Power10 basierende IBM Power-Plattform E1080 der Enterprise-Klasse sind genau auf diese Anforderungen zugeschnitten. Wer bei den neuen IBM Power-Prozessoren an einen kleinen Entwicklungsschritt denkt, unterschätzt ihr Potenzial. Mit ihnen ist IBM ein großer und zukunftsweisender Sprung gelungen.

Der Prozessor ermöglicht den Einsatz von Confidential Computing-Technologie für eine hardwarebasierte Verschlüsselung, mit der Data in Transit geschützt werden können. Die Bandbreite des Power10-Prozessors wurde erheblich erhöht, um eine leistungsstarke Skalierbarkeit auf bis zu 16 Sockets zu ermöglichen. Die Ausfallsicherheit wurde durch die Fähigkeit erweitert, sporadische Fehler automatisch zu erkennen, zu isolieren und für eine Recovery zu sorgen, ohne dass der Betrieb unterbrochen oder auf das Betriebssystem zurückgegriffen werden muss. Der MMA auf dem Chip ermöglicht KI-Inferenzierung in Echtzeit, ohne dass es dafür eines Koprozessors bedarf. Darüber hinaus lässt die Kombination von Red Hat-Lösungen und IBM Cloud-Software einen uneingeschränkten Betrieb als Hybrid-Cloud zu. Mit dem Power10-Chip als Motor der neuen Power-Plattform E1080 unterstützt IBM die Entwicklung des Enterprise Computing in Richtung eines Sweetspots, der das Beste aller Welten vereint: Offenheit, Rechenleistung, Hybrid-Cloud, künstliche Intelligenz, Zugriffsschutz, Skalierbarkeit, Nachhaltigkeit und Zuverlässigkeit auf einer einzigen Plattform.

Der Analyst

Peter Rutten

Research Director, Infrastructure Systems, Platforms and Technologies Group,
Performance Intensive Computing Solutions Global Research Lead, IDC

Peter Rutten ist Forschungsleiter im Bereich der weltweiten Infrastrukturpraxis von IDC, der auch die Marktforschung zu IT-Plattformen beinhaltet. IDC-Forschungsarbeiten zu durchsatzintensiven IT-Lösungen und Anwendungsfällen auf der ganzen Welt unterliegen seiner Leitung, die Marktforschung zu künstlicher Intelligenz (KI), Modellierung und Simulation (M&S) sowie BDA-Infrastruktur (Big Data and Analytics) und zugehörige Lösungsstacks eingeschlossen. Seine Arbeit auf dem Gebiet der Datenverarbeitung im Hochleistungssegment erstreckt sich über Supercomputing, High-End-Computing, In-Memory Computing, IT-Infrastruktursysteme mit Beschleunigern sowie heterogene IT-Infrastruktursysteme, Plattformen und Technologien. Compute-Plattformen mit GPUs, FPGAs, ASICs und anderen Beschleunigern, die in der Cloud oder lokal bereitgestellt werden, gehören ebenso zu seinem Spezialgebiet wie die Marktforschung zu missionskritischen x86-Plattformen, Mainframes und RISC-basierten Systemen sowie deren Betriebsumgebungen (Linux, z/OS, UNIX). Peter Rutten befasst sich darüber hinaus mit neuen Technologien und Plattformen wie Quantencomputing, neuromorphen Rechenverfahren und Technologien, die das Potenzial zur Umwälzung reifer Infrastrukturmärkte bergen. Im Rahmen seines Aufgabenbereichs führt Peter Rutten quantitative (Markteinschätzung und Prognose) und qualitative (auf primärer Marktforschung basierende) Analysen sowie kundenspezifische Markteinschätzungen für IDC-Kunden durch.

Weitere Informationen zu Peter Rutten

IDC Custom Solutions

Diese Veröffentlichung wurde von IDC Custom Solutions erstellt. Als weltweit führender Anbieter von Marktinformationen, Beratungsdienstleistungen und Veranstaltungen auf dem Gebiet der Informationstechnologie, der Telekommunikation und der Verbrauchertechnologie hilft die Custom Solutions-Gruppe von IDC ihren Kunden bei einer fundierten Planung und bei der Entwicklung erfolgreicher Marketing- und Vertriebsstrategien, mit denen sich die Unternehmen auf den globalen Märkten behaupten können. Wir generieren zielführende Marktforschung und überzeugende Content-Marketing-Programme, die messbare Ergebnisse liefern.

IDC @idc @idc idc.com