



FlexPod für die Genomik

FlexPod

NetApp
October 30, 2025

Inhalt

FlexPod für die Genomik	1
TR-4911: FlexPod Genomics	1
Umfang	2
Zielgruppe	2
Auf FlexPod implementierte Krankenhausfunktionen	2
Vorteile der Implementierung genomischer Workloads auf FlexPod	3
Hardware- und Softwarekomponenten der Lösungsinfrastruktur	9
Genomik - GATK Einrichtung und Ausführung	13
Von der Probe bis zur Variantenidentifikation, Anmerkung und Vorhersage	13
Genom Analysis Toolkit (GATK)	15
Einrichtung von FlexPod	15
GATK-Einrichtung und -Ausführung	19
Ausgabe zur Ausführung des GATK unter Verwendung der JAR-Datei	23
Ausgabe zur Ausführung des GATK mit dem Skript ./gatk	26
Ausgabe zur Ausführung von GATK mit Cromwell Engine	28
GPU-Einrichtung	32
Schlussfolgerung	41
Wo Sie weitere Informationen finden	42
Versionsverlauf	43

FlexPod für die Genomik

TR-4911: FlexPod Genomics

JayaKishore Esanakula, NetApp

In der Medizin gibt es nur wenige Gebiete, die für das Gesundheitswesen und die Biowissenschaften wichtiger sind als Genomik. Genomik wird für Ärzte und Krankenschwestern schnell zu einem wichtigen klinischen Werkzeug. Genomik hilft uns in Kombination mit medizinischer Bildgebung und digitaler Pathologie zu verstehen, wie die Gene eines Patienten durch Behandlungsprotokolle beeinflusst werden können. Der Erfolg von Genomik im Gesundheitswesen hängt zunehmend von Dateninteroperabilität nach Maß ab. Ziel ist es, die enormen Mengen an genetischen Daten zu erkennen und klinisch relevante Zusammenhänge und Varianten zu identifizieren, die die Diagnose verbessern und die Präzisionsmedizin zur Realität machen. Genomik hilft uns dabei zu verstehen, woher Krankheiten kommen, wie sich Krankheiten entwickeln und welche Behandlungen und Strategien effektiv sein können. Die Genomik bietet offensichtlich viele Vorteile, die nicht nur Prävention, sondern auch Diagnostik und Behandlung umfassen. Gesundheitseinrichtungen haben verschiedene Herausforderungen mit sich:

- Bessere Versorgungsqualität
- Wertbasierte Versorgung
- Datenexplosion
- Präzisionsmedizin
- Pandemien
- Wearables, Fernüberwachung und Pflege
- Cyber-Sicherheit

Standardisierte klinische Behandlungspfade und klinische Protokolle sind eine der kritischen Komponenten der modernen Medizin. Einer der wichtigsten Aspekte der Standardisierung ist die Interoperabilität zwischen Gesundheitsanbietern: Nicht nur bei medizinischen Unterlagen, sondern auch bei Genomdaten. Die große Frage lautet: Werden Gesundheitseinrichtungen sich anstelle der Patienteneigentum an ihren persönlichen Genomdaten und den damit verbundenen medizinischen Unterlagen auf Genomdaten verzichten?

Interoperable Patientendaten sind der Schlüssel zu Präzisionsmedizin, eine der ausschlaggebenden Faktoren bei der kürzlich explosionsartigen Zunahme des Datenwachstums. Das Ziel der Präzisionsmedizin ist es, die Gesundheitsvorsorge, die Prävention, die Diagnose und die Behandlungslösungen effektiver und genauer zu gestalten.

Das Datenwachstum war exponentiell. Anfang Februar 2021 wurden in den USA ca. 8,000 COVID-19 Stämme pro Woche sequenziert. Die Anzahl der sequenzierten Genome war bis April 2021 auf 29,000 pro Woche erhöht. Jedes vollständig sequenzierte menschliche Genom ist etwa 125GB groß. Daher würde die gesamte Genomspeicherung im Ruhezustand mit einer Rate von 29,000 Genomen pro Woche mehr als 180 Petabyte pro Jahr betragen. Verschiedene Länder haben sich für die Genomepidemiologie engagiert, um die genomische Überwachung zu verbessern und sich auf die nächste Welle globaler Herausforderungen im Gesundheitswesen vorzubereiten.

Die reduzierten Kosten für die Genomforschung führen zu nie da gewesenen Gentests und Forschungen. Die drei PS befinden sich an einem Wendepunkt: Computerleistung, Datenschutz und Personalisierung der Medizin. Bis 2025 schätzen Forscher, dass 100 Millionen bis 2 Milliarden menschliche Genome sequenziert werden. Damit Genomik effektiv und wertvoll sein kann, müssen Genomik-Funktionen einen nahtlosen Teil der Pflegungsprozesse sein. Er sollte leicht zugänglich sein und bei einem Patientenbesuch umsetzbar sein. Ebenso wichtig ist es, dass Patientendaten der elektronischen Krankenakten in die Genomik-Daten des Patienten integriert werden. Mit der Einführung hochmoderner konvergenter Infrastrukturen wie FlexPod können Unternehmen ihre Genomfunktionen in die alltäglichen Workflows von Ärzten, Pflegepersonal und Klinikmanagern integrieren. Aktuelle Informationen zur FlexPod Plattform finden Sie in dieser ["FlexPod Datacenter with Cisco UCS X-Series Whitepaper"](#).

Der wahre Nutzen der Genomforschung besteht für Ärzte darin, Präzisionsmedizin zu bieten und personalisierte Behandlungspläne zu entwickeln, die auf den genomischen Daten eines Patienten basieren. In der Vergangenheit gab es noch nie derartige Synergien zwischen Klinikpersonal und Datenanalysten, und die Genomik profitiert von den technologischen Innovationen der jüngsten Vergangenheit sowie von echten Partnerschaften zwischen Einrichtungen im Gesundheitswesen und Technologieführern der Branche.

Akademische medizinische Zentren und andere Organisationen im Gesundheitswesen und Life Science sind auf dem besten Weg, um in der Genomforschung das Kompetenzzentrum (COE) aufzubauen. Laut Dr. Charlie Gersbach, Dr. Greg Crawford und Dr. Tim E Reddy von der Duke University: „Wir wissen, dass Gene nicht durch einen einfachen binären Schalter ein- oder ausgeschaltet werden, sondern dass es ein Ergebnis mehrerer genregulatorischer Schalter ist, die zusammen arbeiten. Sie haben auch festgestellt, dass „keiner dieser Teile des Genoms isoliert arbeitet. Das Genom ist ein sehr kompliziertes Netz, das Evolution gewebt hat“ (["ref"](#)).

NetApp und Cisco arbeiten bereits seit über 10 Jahren intensiv an der Implementierung inkrementeller Verbesserungen in der FlexPod Plattform. Das gesamte Kundenfeedback wird gehört, bewertet und an die Value Streams und Funktionen von FlexPod gebunden. Es ist diese kontinuierliche Schleife, Zusammenarbeit, Verbesserungen und Feier, die FlexPod als vertrauenswürdige konvergente Infrastrukturplattform auf der ganzen Welt auszeichnet. Die Lösung wurde von Grund auf vereinfacht und als die zuverlässigste, robusteste, vielseitigste und agilste Plattform für Unternehmen im Gesundheitswesen konzipiert.

Umfang

Mit der konvergenten Infrastrukturplattform von FlexPod können Gesundheitseinrichtungen einen oder mehrere Genomapplikationen und andere klinische und nicht klinische Applikationen im Gesundheitswesen hosten. Dieser technische Bericht verwendet ein Open-Source-Tool für die branchenübliche Genomik mit dem Namen GATK während der Plattformvalidierung von FlexPod. Eine umfassendere Diskussion über Genomik oder GATK ist jedoch nicht im Rahmen dieses Dokuments enthalten.

Zielgruppe

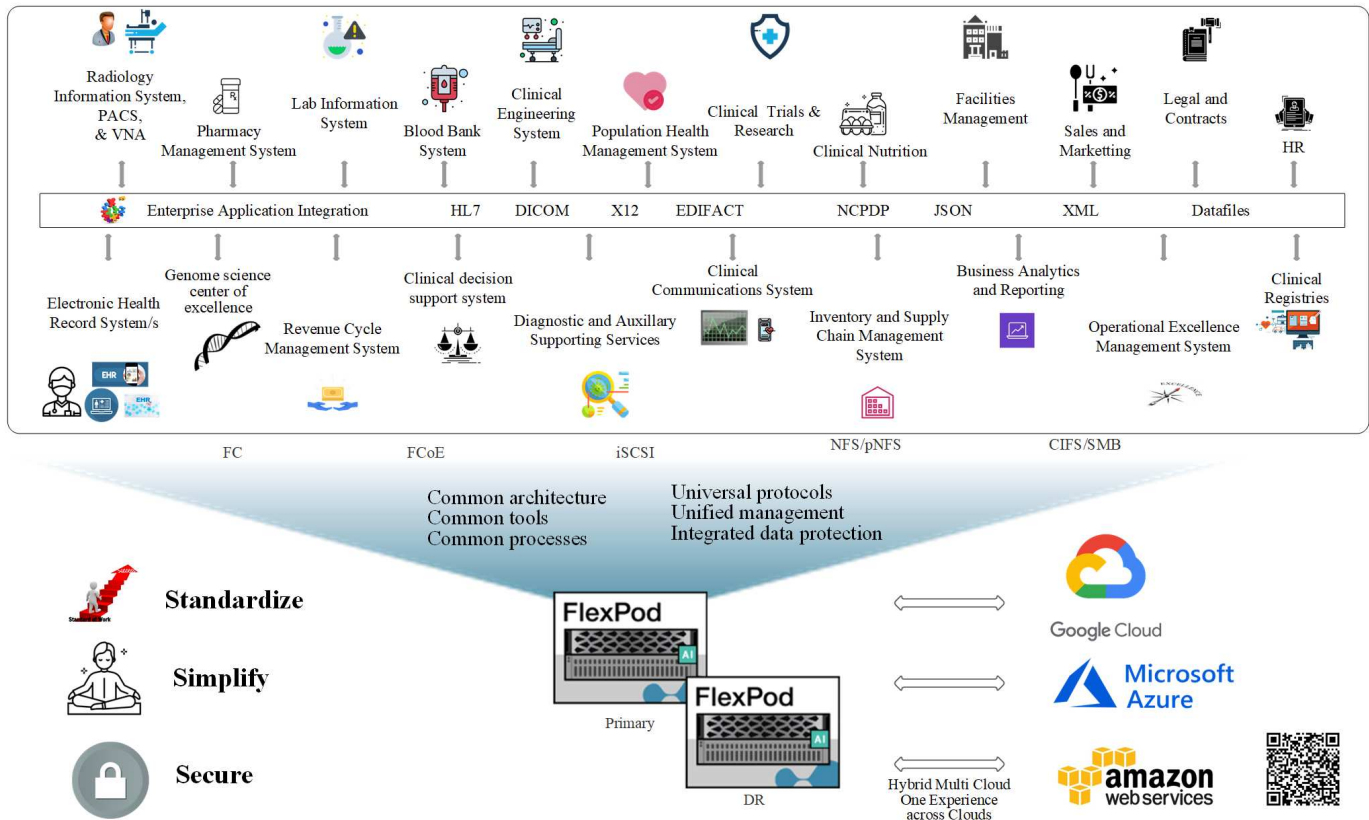
Dieses Dokument richtet sich an technische Leiter im Gesundheitswesen sowie an Lösungstechniker von Cisco und NetApp Partnern und Professional Services-Mitarbeiter. NetApp geht davon aus, dass der Leser gute Kenntnisse der Konzepte zur Berechnung der Storage- und Computing-Größenbemessung sowie der technischen Vertrautheit mit Bedrohungen für das Gesundheitswesen, mit der Sicherheit im Gesundheitswesen, MIT IT-Systemen im Gesundheitswesen, mit Cisco UCS und NetApp Storage-Systemen hat.

Auf FlexPod implementierte Krankenhausfunktionen

Ein typisches Krankenhaus verfügt über eine Reihe an IT-Systemen. Der Großteil solcher Systeme wird bei einem Anbieter gekauft, während nur sehr wenige von dem Krankenhaussystem im Haus gebaut werden. Aus diesem Grund muss das Kliniksystem eine diverse Infrastrukturumgebung in seinen Datacentern managen.

Wenn Krankenhäuser ihre Systeme in einer konvergenten Infrastrukturplattform wie FlexPod zusammenführen, können Unternehmen ihren Datacenter-Betrieb standardisieren. Mit FlexPod können Gesundheitseinrichtungen klinische und nicht klinische Systeme auf derselben Plattform implementieren und so den Datacenter-Betrieb vereinheitlichen.

Hospital capabilities deployed on a FlexPod



"Der nächste Schritt: Vorteile der Implementierung genomischer Workloads auf FlexPod."

Vorteile der Implementierung genomischer Workloads auf FlexPod

"Zurück: Einführung."

Dieser Abschnitt bietet eine kurze Übersicht über die Vorteile, die zum Ausführen eines Genomik-Workloads auf einer konvergenten FlexPod Infrastrukturplattform erzielt werden können. Lassen Sie uns kurz die Möglichkeiten eines Krankenhauses beschreiben. Die folgende Ansicht der Unternehmensarchitektur zeigt die Funktionen eines Krankenhauses, die auf einer konvergenten, Hybrid-Cloud-fähigen FlexPod Infrastrukturplattform bereitgestellt werden.

- **Vermeiden von Silos im Gesundheitswesen.** Silos im Gesundheitswesen sind ein sehr echtes Anliegen. Abteilungen sind häufig Silos in ihrer eigenen Hardware und Software, nicht nach ihrer Wahl, sondern organisch. Beispielsweise Radiologie, Kardiologie, EHR, Genomik analysen, Umsatzzyklen und andere Abteilungen enthalten jeweils einen individuellen Satz an dedizierter Software und Hardware. Einrichtungen im Gesundheitswesen haben nur wenige IT-Fachkräfte für das Management ihrer Hardware-

und Software-Ressourcen. Der Wendepunkt folgt, wenn zu erwarten ist, dass diese Gruppe von Einzelpersonen ein sehr vielseitiges Spektrum an Hardware und Software managen wird. Die Heterogenität wird durch eine unkongruente Reihe von Prozessen, die von Anbietern in die Gesundheitsorganisation eingebracht werden, noch verschlimmert.

- **Start Small and Grow.** das GATK Tool Kit ist auf die CPU-Ausführung abgestimmt, die besten Suites Plattformen wie FlexPod. FlexPod ermöglicht eine unabhängige Skalierbarkeit von Netzwerk, Computing und Storage. Klein beginnen und mit der Zeit wachsen, wenn die Genomfunktionen und die Umgebung wachsen. Unternehmen im Gesundheitswesen müssen nicht in spezialisierte Plattformen investieren, um genomische Workloads auszuführen. Stattdessen können Unternehmen vielseitige Plattformen wie FlexPod nutzen, um Genomik- und nicht-genomikfreie Workloads auf derselben Plattform auszuführen. Wenn beispielsweise die Abteilung für Pädiatrie Genomfunktionen implementieren möchte, kann die IT-Leitung Computing, Storage und Networking auf einer vorhandenen FlexPod Instanz bereitstellen. Mit dem Wachstum der Geschäftsbereiche zur Genomik können Unternehmen im Gesundheitswesen ihre FlexPod Plattform nach Bedarf skalieren.
- **Eine einzige Kontrollscheibe und unübertroffene Flexibilität.** Cisco Intersight vereinfacht den IT-Betrieb erheblich, indem Anwendungen mit Infrastruktur überbrückt werden. Dadurch werden Transparenz und Management von Bare-Metal-Servern und -Hypervisoren zu serverlosen Anwendungen ermöglicht, wodurch Kosten gesenkt und Risiken gemindert werden. Diese einheitliche SaaS-Plattform verwendet ein einheitliches Open API-Design, das sich nativ in Plattformen und Tools von Drittanbietern integrieren lässt. Außerdem kann das Betriebsteam des Datacenters direkt oder von einem beliebigen Ort aus über eine mobile App heraus managen.

Die Benutzer schöpfen in ihrer Umgebung schnell greifbaren Wert aus, indem sie Intersight als Managementplattform nutzen. Durch die Möglichkeit der Automatisierung vieler täglicher manueller Aufgaben beseitigt Intersight Fehler und vereinfacht Ihre täglichen Abläufe. Dank der erweiterten Support-Funktionen von Intersight können Anwender zudem Probleme voraus bleiben und die Problembeseitigung beschleunigen. Gemeinsam geben Unternehmen deutlich weniger Zeit und Geld für ihre Applikationsinfrastruktur aus und gewinnen mehr Zeit für die eigentliche Geschäftsentwicklung.

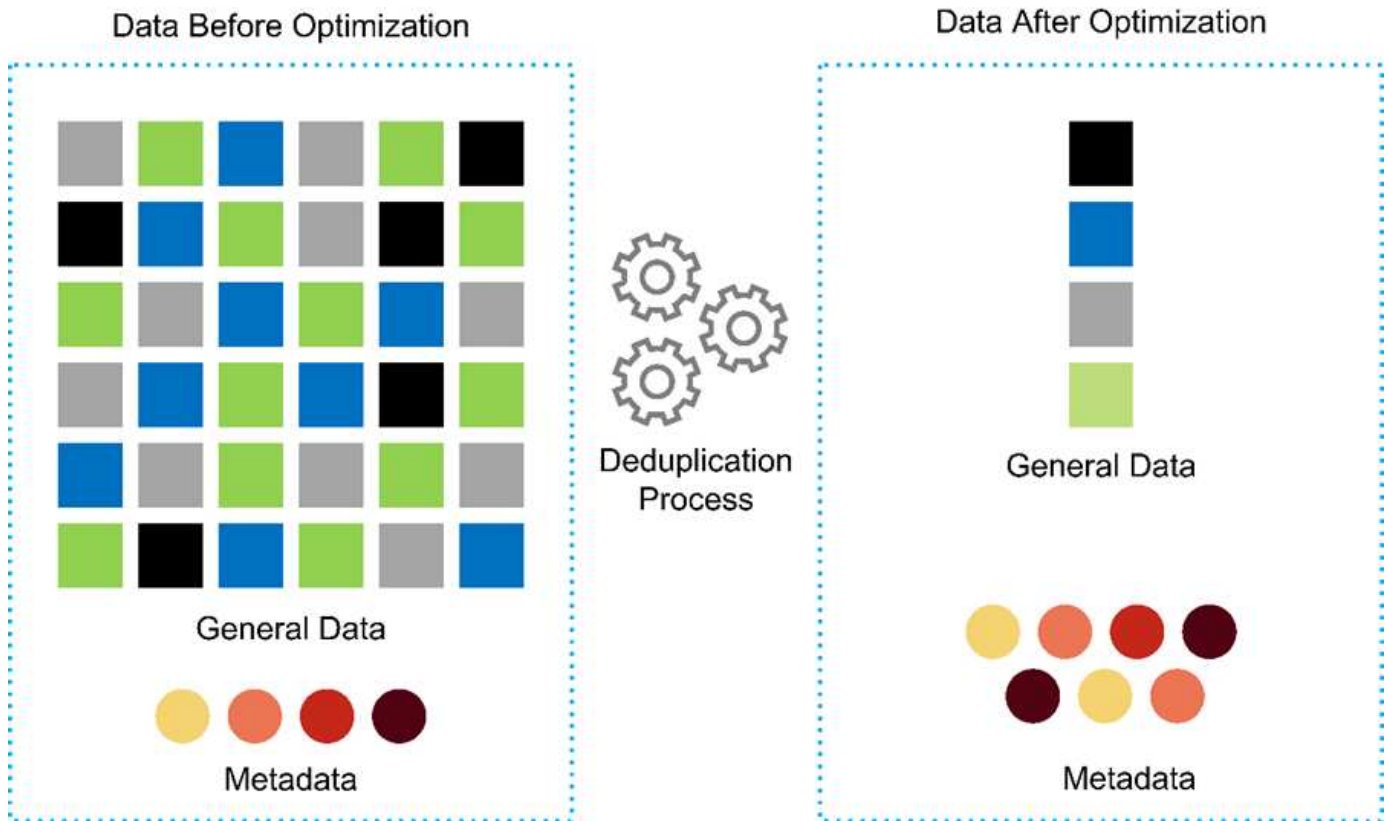
Durch die Nutzung von Intersight Management und der leicht skalierbaren Architektur von FlexPod können Unternehmen mehrere Genom-Workloads auf einer einzigen FlexPod Plattform ausführen. Dies steigert die Auslastung und senkt die Gesamtbetriebskosten (TCO). FlexPod ermöglicht die flexible Dimensionierung. Dabei stehen Ihnen die Wahl, beginnend mit unserem kleinen FlexPod Express, und die Skalierung zu großen FlexPod Datacenter-Implementierungen. Dank der in Cisco Intersight integrierten Funktionen zur rollenbasierten Zugriffssteuerung können Organisationen im Gesundheitswesen robuste Zugriffskontrollmechanismen implementieren, sodass keine separaten Infrastruktur-Stacks erforderlich sind. Mehrere Geschäftseinheiten im Gesundheitswesen können Genomik als wichtigste Kernkompetenzen nutzen.

FlexPod vereinfacht den IT-Betrieb und senkt die Betriebskosten. IT-Infrastrukturadministratoren können sich auf Aufgaben konzentrieren, die Klinikpersonal Innovationen ermöglichen, anstatt nur auf den Betrieb zu zugreifen.

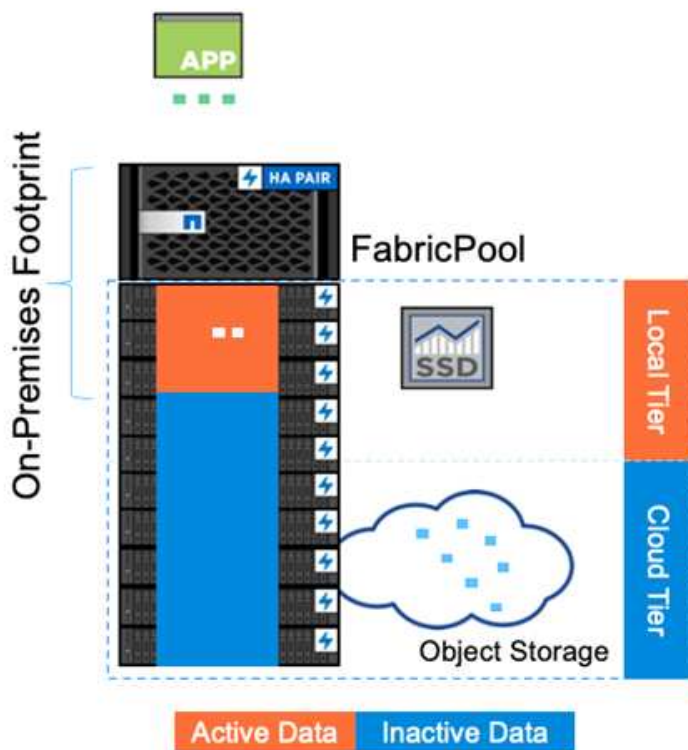
- **Validiertes Design und garantierte Ergebnisse.** Design- und Implementierungsleitfäden für FlexPod werden wiederholbar validiert. Sie umfassen umfassende Konfigurationsdetails und branchenspezifische Best Practices, die erforderlich sind, um ein FlexPod zuverlässig zu implementieren. Validierte Designrichtlinien, Implementierungsleitfäden und Architekturen von Cisco und NetApp helfen Ihrem Unternehmen im Gesundheitswesen oder Life Science, bei der Implementierung einer validierten und bewährten Plattform von Anfang an keine Unsicherheiten mehr zu machen. Mit FlexPod können Sie Implementierungszeiten verkürzen und Kosten, Komplexität und Risiken senken. Durch validierte Designs und Implementierungsleitfäden von FlexPod wird FlexPod als ideale Plattform für eine Vielzahl von Genomik-Workloads etabliert.
- **Innovation und Agilität.** FlexPod wird als ideale Plattform von EHRs wie Epic, Cerner, Meditech und

Imaging Systemen wie Agfa, GE, Philips empfohlen. Finden Sie weitere Informationen zu ["Epische Ehrenrolle"](#) Und Zielplattform-Architektur finden Sie im Epic userweb. Genomik auf Basis ["FlexPod"](#) Ermöglicht es Organisationen im Gesundheitswesen, den Weg der Innovationen flexibel fortzusetzen. Mit FlexPod kommt es auf die natürliche Weise, Veränderungen im Unternehmen umzusetzen. Wenn Unternehmen eine FlexPod-Plattform nutzen, können IT-Experten im Gesundheitswesen Zeit, Aufwand und Ressourcen für Innovationen bereitstellen und so agil sein wie die Anforderungen des Ökosystems.

- **Data befreit.** mit der konvergenten Infrastrukturplattform FlexPod und einem NetApp ONTAP Storage-System können Genomikdaten über eine einzige Plattform zur Verfügung gestellt und zugänglich gemacht werden. FlexPod mit NetApp ONTAP bietet eine einfache, intuitive und leistungsstarke Hybrid-Cloud-Plattform. Die Data Fabric von NetApp ONTAP verknüpft Daten über Standorte, physische Grenzen und Applikationen hinweg. Ihre Data Fabric wurde für Unternehmen in einer datenorientierten Welt entwickelt. Daten werden an zahlreichen Orten erstellt und verwendet. Oft werden sie auch an mehreren Orten sowie in mehreren Applikationen und Infrastrukturen gleichzeitig genutzt. Daher benötigen Sie eine einheitliche und integrierte Lösung für das Management. Mit FlexPod hat Ihr IT-Team die Kontrolle und vereinfacht die ständig zunehmende Komplexität IM IT-BEREICH.
- **Sichere Mandantenfähigkeit.** FlexPod verwendet FIPS 140-2-2-konforme Kryptografiemodule, die es Unternehmen ermöglichen, Sicherheit als Grundelement und nicht als Nachdenken zu implementieren. FlexPod ermöglicht es Unternehmen, sichere Mandantenfähigkeit von einer einzigen konvergenten Infrastrukturplattform aus zu implementieren, unabhängig von der Größe der Plattform. FlexPod mit sicherer Mandantenfähigkeit und QoS unterstützt die Trennung von Workloads und maximiert die Auslastung. Dadurch vermeiden Sie, dass Investitionen in spezialisierte Plattformen gebunden sind, die möglicherweise nicht ausgelastet sind und über spezielle Fachkenntnisse für das Management erforderlich sind.
- **Storage-Effizienz.** Genomics erfordert, dass der zugrunde liegende Storage über branchenführende Storage-Effizienz-Funktionen verfügt. NetApp Storage-Effizienzfunktionen wie Deduplizierung (inline und On-Demand), Datenkomprimierung und Data-Compaction (senken die Storage-Kosten ["ref"](#)). NetApp Deduplizierung bietet Deduplizierung auf Blockebene in einem FlexVol Volume. Im Wesentlichen werden durch Deduplizierung Blockduplikate entfernt und somit nur eindeutige Blöcke im FlexVol Volume gespeichert. Die Deduplizierung arbeitet mit einer hohen Granularität und wird auf dem aktiven File-System des FlexVol Volume betrieben. Die folgende Abbildung zeigt die Funktionsweise der NetApp Deduplizierung. Deduplizierung ist applikationsunabhängig. Somit können auch Daten von beliebigen Applikationen dedupliziert werden, die das NetApp System nutzen. Sie können die Volume-Deduplizierung als Inline-Prozess und als Hintergrundprozess ausführen. Sie können die Funktion so konfigurieren, dass sie automatisch ausgeführt, geplant oder manuell über die CLI, den NetApp ONTAP System Manager oder NetApp Active IQ Unified Manager gestartet wird.



- Genomik-Interoperabilität ermöglichen.** ONTAP FlexCache ist eine Remote-Caching-Funktion, die Dateiverteilung vereinfacht, WAN-Latenz reduziert und die Kosten für die WAN-Bandbreite senkt (["ref"](#)). Eine der wichtigsten Aktivitäten bei der Identifizierung von Genomvariationen und bei der Annotation ist die Zusammenarbeit zwischen Ärzten. Die ONTAP FlexCache-Technologie erhöht den Datendurchsatz, selbst wenn mehrere Kliniker an verschiedenen geografischen Standorten zusammenarbeiten. Angesichts der typischen Größe einer *.BAM-Datei (1 GB bis 100 GB) ist es von großer Bedeutung, dass die zugrunde liegende Plattform Dateien für Kliniker an verschiedenen geografischen Standorten verfügbar machen kann. Mit FlexPod mit ONTAP FlexCache sind genomische Daten und Applikationen vollständig auf mehrere Standorte vorbereitet. Somit wird die Zusammenarbeit zwischen den weltweiten Forschern nahtlos ermöglicht, da sie eine geringe Latenz und einen hohen Durchsatz bieten. Medizinische Einrichtungen, die Genomapplikationen auf mehreren Standorten ausführen, können mit der Data-Fabric-Architektur horizontal skaliert werden, um das Management zu einem ausgewogenen Verhältnis zwischen Kosten und Geschwindigkeit zu halten.
- Intelligente Nutzung der Speicherplattform.** FlexPod mit ONTAP Auto-Tiering und NetApp Fabric Pool Technologie vereinfacht das Datenmanagement. FabricPool senkt die Storage-Kosten, ohne dabei Einbußen bei Performance, Effizienz, Sicherheit oder Sicherung hinnehmen zu müssen. FabricPool ist transparent für Enterprise-Applikationen und nutzt die Cloud-Effizienz weiter, indem die Storage-TCO gesenkt werden, ohne dass die Applikationsinfrastruktur umgestaltet werden muss. FlexPod bietet die Storage Tiering-Funktionen von FabricPool für eine effizientere Nutzung von ONTAP Flash Storage. Weitere Informationen finden Sie unter ["FlexPod mit FabricPool"](#). Das folgende Diagramm bietet einen allgemeinen Überblick über FabricPool und seine Vorteile.



Automatic tiering
Zero-touch management
Preserves file system
Lower cost of ownership
Choice of object tier locations



- **Schnellere Variantenanalyse und -Annotation.** die FlexPod-Plattform ist schneller bereitzustellen und zu operationalisieren. Die FlexPod Plattform ermöglicht klinische Zusammenarbeit, da die Daten in großen Umgebungen mit niedriger Latenz und höherem Durchsatz verfügbar sind. Eine bessere Interoperabilität ermöglicht Innovationen. Medizinische Einrichtungen können nebeneinander ihre genomischen und nicht genomischen Workloads ausführen. Das bedeutet, dass Unternehmen für den Übergang zur Genomik keine speziellen Plattformen benötigen.

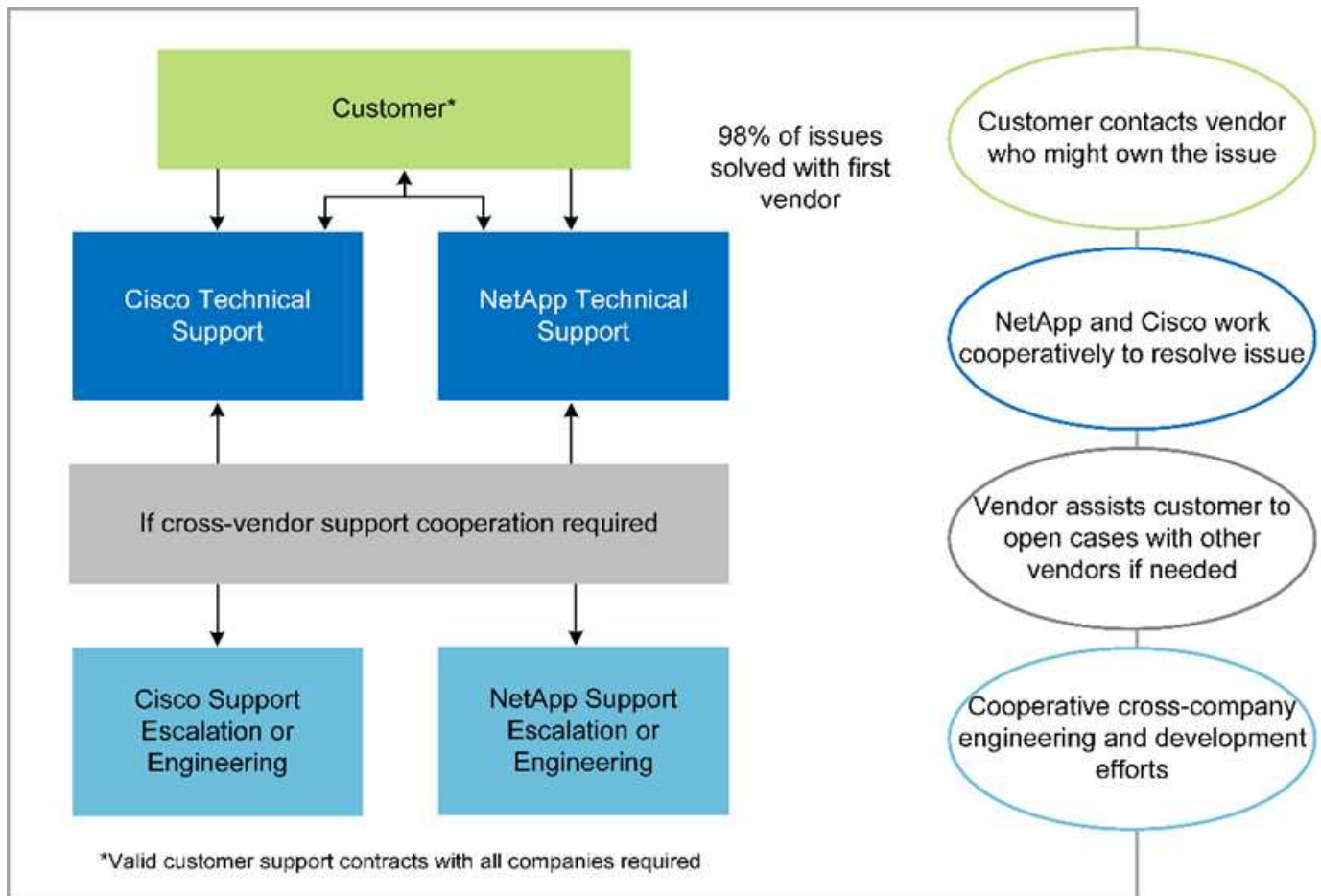
FlexPod ONTAP erweitert die Storage-Plattform routinemäßig auf die neuesten Funktionen. FlexPod Datacenter ist die optimale Grundlage für die Implementierung von FC- NVMe-Storage für hochperformanten Storage-Zugriff auf Applikationen, die sie benötigen. Da FC- NVMe mit hoher Verfügbarkeit, Multipathing und zusätzlicher Unterstützung von Betriebssystemen einhergeht, eignet sich FlexPod hervorragend als bevorzugte Plattform und bietet die Skalierbarkeit und Zuverlässigkeit, die zur Unterstützung dieser Funktionen erforderlich sind. ONTAP mit schnelleren I/O-Vorgängen und End-to-End-NVMe ermöglicht Analysen der Genomik schneller ("ref").

Die sequenzierten RAW-Genomdaten erzeugen große Dateigrößen, und es ist wichtig, dass diese Dateien den Variantenanalysatoren zur Verfügung gestellt werden, um die Gesamtzeit von der Probensammlung bis zur Variantenbeschriftung zu reduzieren. Wenn NVMe (Nonvolatile Memory Express) als Storage-Zugriffs- und Datenübertragungsprotokoll verwendet wird, bietet das Unternehmen einen beispiellosen Durchsatz und die schnellsten Reaktionszeiten. FlexPod implementiert das NVMe-Protokoll und greift über den PCI Express Bus (PCIe) auf Flash-Storage zu. PCIe ermöglicht die Implementierung von Zehntausenden von Befehlswarteschlangen, wodurch sich die Parallelisierung und der Durchsatz erhöhen. Ein einziges Protokoll von der Storage- bis zum Speicher sorgt für schnellen Datenzugriff.

- **Agilität für die klinische Forschung von Grund auf.** Dank flexibler, erweiterbarer Speicherkapazität und Performance können Forschungsunternehmen im Gesundheitswesen die Umgebung flexibel oder just-in-time (JIT) optimieren. Durch die Entkopplung der Storage-Systeme von der Computing- und Netzwerkinfrastruktur lässt sich die FlexPod Plattform unterbrechungsfrei vertikal und horizontal skalieren. Mithilfe von Cisco Intersight lässt sich die FlexPod Plattform sowohl mit integrierten als auch mit

benutzerdefinierten automatisierten Workflows managen. Durch die Cisco Intersight Workflows können Organisationen im Gesundheitswesen die Lebenszyklusmanagement-Zeiten von Anwendungen reduzieren. Wenn ein akademisches medizinisches Zentrum verlangt, dass Patientendaten anonymisiert und ihrem Zentrum für Forschungsinformatik bzw. Datacenter in Bezug auf Qualität zur Verfügung gestellt werden, kann die IT-Abteilung Cisco Intersight FlexPod Workflows nutzen, um sichere Daten-Backups, Klone und die Wiederherstellung in nur wenigen Sekunden statt Stunden durchzuführen. Mit NetApp Trident und Kubernetes können IT-Abteilungen neue Data Scientists bereitstellen und klinische Daten für die Modellentwicklung in wenigen Minuten – manchmal sogar in Sekunden – zur Verfügung stellen.

- **Schutz von Genomdaten.** NetApp SnapLock bietet ein speziell zu Zweck geuniversell einsetzbares Volume, in dem Dateien gespeichert und in einen nicht löschbaren, nicht wiederbeschreibbaren Zustand versetzt werden können. Die Produktionsdaten des Benutzers, die sich in einem FlexVol Volume befinden, können mithilfe von NetApp SnapMirror oder SnapVault gespiegelt oder in ein SnapLock Volume archiviert werden. Die Dateien im SnapLock Volume, das Volume selbst und das Hosting-Aggregat können bis zum Ende der Aufbewahrungsdauer nicht gelöscht werden. ONTAP FPolicy Software verhindert Ransomware-Angriffe, indem sie auf Dateien mit bestimmten Erweiterungen distanziert. Ein FPolicy-Ereignis kann für bestimmte Dateivorgänge ausgelöst werden. Das Ereignis ist mit einer Richtlinie verknüpft, die die Engine aufruft, die es verwenden muss. Sie können eine Richtlinie mit einer Reihe von Dateierweiterungen konfigurieren, die möglicherweise Ransomware enthalten könnten. Wenn eine Datei mit einer nicht zulässigen Erweiterung versucht, einen nicht autorisierten Vorgang auszuführen, verhindert FPolicy die Ausführung dieses Vorgangs ("ref").
- **Kooperativer Support für FlexPod** NetApp und Cisco haben ein solides, skalierbares und flexibles Support-Modell für den FlexPod entwickelt, das die individuellen Support-Anforderungen der konvergenten FlexPod Infrastruktur erfüllt. Dieses Modell greift auf die Erfahrungswerte, Ressourcen und das Know-how des technischen Supports von NetApp und Cisco zurück, um unabhängig von der Ursache des Problems einen optimierten Prozess der Identifizierung und Behebung von FlexPod Support-Problemen zu bieten. Die folgende Abbildung bietet einen Überblick über das kooperative Support-Modell für FlexPod. Der Kunde kontaktiert den Anbieter, der möglicherweise für das Problem zuständig ist, und Cisco und NetApp arbeiten gemeinsam an einer Lösung. Cisco und NetApp verfügen über unternehmensübergreifende Engineering- und Entwicklungsteams, die Hand in Hand arbeiten, um Probleme zu lösen. Dieses Support-Modell reduziert den Verlust von Informationen während der Übersetzung, sorgt für Vertrauen und reduziert Ausfallzeiten.



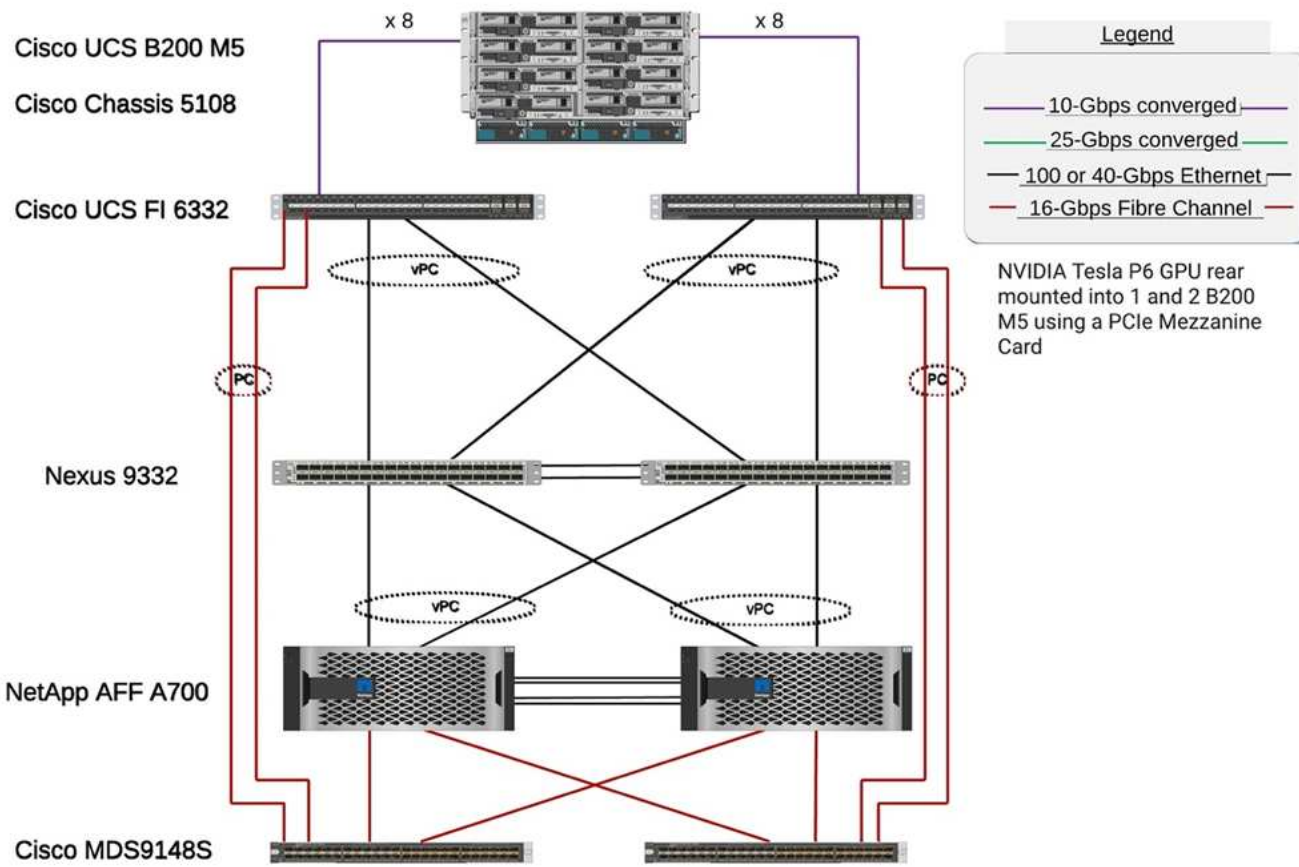
"Als Nächstes: Hardware- und Softwarekomponenten in der Lösungsinfrastruktur."

Hardware- und Softwarekomponenten der Lösungsinfrastruktur

"Previous: Vorteile der Implementierung genomische Workloads auf FlexPod."

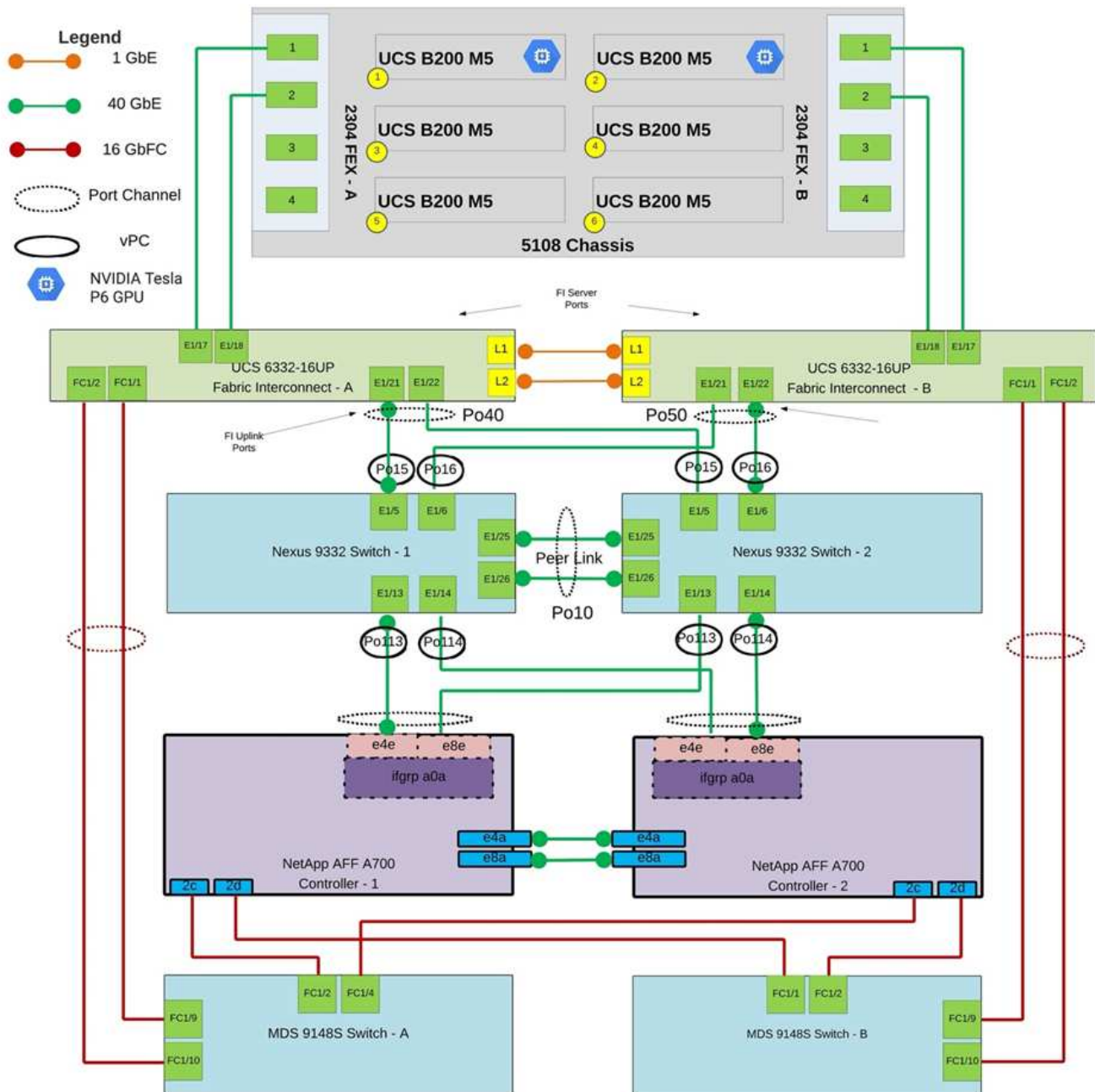
Die folgende Abbildung zeigt das FlexPod-System, das für die Einrichtung und Validierung von GATK verwendet wird. Wir haben genutzt "[FlexPod Datacenter mit VMware vSphere 7.0 und NetApp ONTAP 9.7 Cisco Validated Design \(CVD\)](#)" Während des Setups.

FlexPod for Genomics



Im folgenden Diagramm sind die Details zur FlexPod-Verkabelung dargestellt.

FlexPod for Genomics



In der folgenden Tabelle sind die während der GATK-Tests verwendeten Hardwarekomponenten auf einem FlexPod aufgeführt. Hier ist der "[NetApp Interoperabilitäts-Matrix-Tool](#)" (IMT) und "[Cisco Hardware Compatibility List](#) (HCL)".

Schicht	Produktfamilie	Menge und Modell	Details
Computing	Cisco UCS 5108 Chassis	1 oder 2	
	Cisco UCS Blade Server	6 B200 M5	Jeweils mit 2x 20 Cores, 2,7 GHz und 128 bis 384 GB RAM

Schicht	Produktfamilie	Menge und Modell	Details
	Cisco UCS Virtual Interface Card (VIC)	Cisco UCS 1440	Siehe
	2 Cisco UCS Fabric Interconnects	6332	-
Netzwerk	Cisco Nexus Switches	2 x Cisco Nexus 9332	-
Datennetzwerk Storage-Netzwerk	IP-Netzwerk für Storage-Zugriff über SMB-/CIFS-, NFS- oder iSCSI-Protokolle	Gleiche Netzwerk-Switches wie oben	-
	Storage-Zugriff über FC	2 x Cisco MDS 9148S	-
Storage	NetApp AFF A700 All-Flash-Storage-System	1 Cluster	Cluster mit zwei Nodes
	Festplatten-Shelf	Ein DS224C oder NS224 Festplatten-Shelf	Vollständig mit 24 Laufwerken bestückt
	SSD	24, 1,2 TB oder höhere Kapazität	-

In dieser Tabelle ist die Infrastruktursoftware aufgeführt.

Software	Produktfamilie	Version/Release	Details
Verschiedene	Linux	RHEL 8.3	-
	Windows	Windows Server 2012 R2 (64-Bit)	-
	NetApp ONTAP	ONTAP 9.8 oder höher	-
	Cisco UCS Fabric Interconnect	Cisco UCS Manager 4.1 oder höher	-
	Cisco Switches der Ethernet-Serie 3000 oder 9000	Für 9000-Serie, 7.0(3)I7(7) oder höher für 3000-Serie, 9.2(4) oder höher	-
	Cisco FC: Cisco MDS 9132T	8.4(1a) oder höher	-
	Hypervisor	VMware vSphere ESXi 7.0	-
Storage	Hypervisor-Managementsystem	VMware vCenter Server 7.0 (vCSA) oder höher	-
Netzwerk	NetApp Virtual Storage Console (VSC)	VSC 9.7 oder höher	-
	NetApp SnapCenter	SnapCenter 4.3 oder höher	-
	Cisco UCS Manager	4.1(3c) oder höher	-

Software	Produktfamilie	Version/Release	Details
Hypervisor	ESXi		
Vereinfachtes	Hypervisor- ManagementsystemVMware vCenter Server 7.0 (vCSA) oder höher		
	NetApp Virtual Storage Console (VSC)	VSC 9.7 oder höher	
	NetApp SnapCenter	SnapCenter 4.3 oder höher	
	Cisco UCS Manager	4.1(3c) oder höher	

"Weiter: [Genomik - GATK Einrichtung und Ausführung.](#)"

Genomik - GATK Einrichtung und Ausführung

"Früher: [Hardware- und Softwarekomponenten der Lösungsinfrastruktur.](#)"

Laut dem National Human Genome Research Institute (["NHGRI"](#) „Genomics ist die Untersuchung aller Gene einer Person (das Genom), einschließlich der Wechselwirkungen dieser Gene miteinander und mit der Umwelt einer Person. „

Laut dem ["NHGRI"](#), "Deoxyribonukleinsäure (DNA) ist die chemische Verbindung, die die notwendigen Anweisungen enthält, um die Aktivitäten von fast allen lebenden Organismen zu entwickeln und zu leiten. DNA-Moleküle bestehen aus zwei verdrehenden, paarweise angeordneten Strängen, die oft als Doppelhelix bezeichnet werden.“ „Der gesamte DNA-Satz eines Organismus wird sein Genom genannt.“

Sequenzierung ist der Prozess der Bestimmung der genauen Reihenfolge der Basen in einem Strang der DNA. Eine der häufigsten Sequenzierungsarten, die heute verwendet werden, nennt man Sequenzierung durch Synthese. Diese Technik verwendet die Emission von fluoreszierenden Signalen, um die Grundlagen zu bestellen. Forscher können mit Hilfe der DNA-Sequenzierung nach genetischen Variationen und Mutationen suchen, die bei der Entwicklung oder dem Fortschreiten einer Krankheit eine Rolle spielen könnten, während sich eine Person noch im embryonalen Stadium befindet.

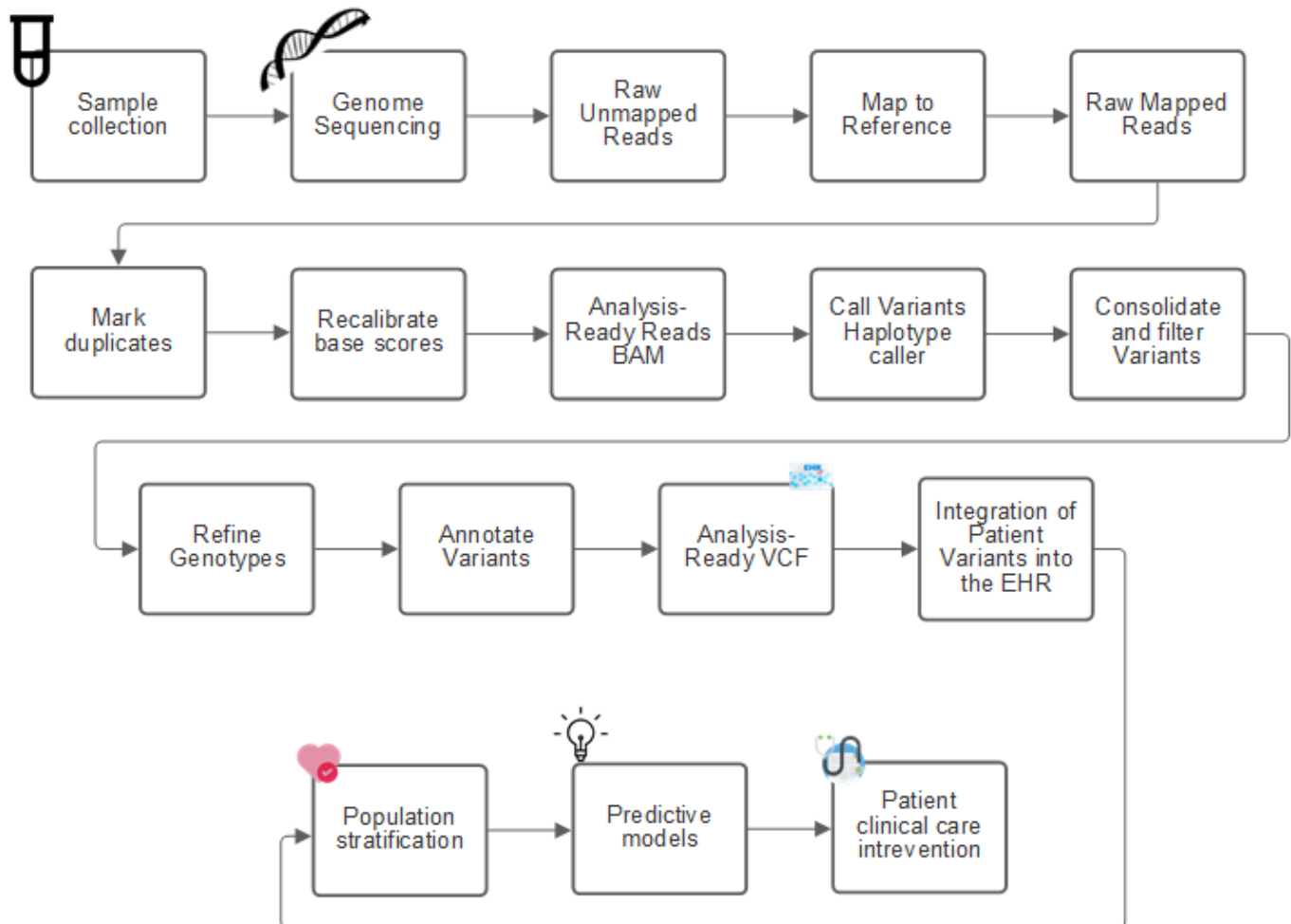
Von der Probe bis zur Variantenidentifikation, Anmerkung und Vorhersage

Genomik kann im allgemeinen zu den folgenden Schritten eingeteilt werden. Dies ist keine umfassende Liste:

1. Probenentnahme.
2. ["Genom-Sequenzierung"](#) Verwenden eines Sequenzers zum Generieren der Rohdaten.
3. Vorverarbeitung: Beispiel: ["Deduplizierung"](#) Wird verwendet ["Picard"](#).
4. Genomanalyse:
 - a. Wird einem Referenzgenom zugeordnet.
 - b. ["Variante"](#) Identifizierung und Beschriftung, die in der Regel mit GATK und ähnlichen Tools durchgeführt werden.
5. Integration in das Electronic Health Record-System (EHR).

6. "Bevölkerungsstratifizierung" Und Identifizierung der genetischen Variation über geografische Lage und ethnische Herkunft.
7. "Prädiktive Modelle" Verwendung von signifikanter Single-Nukleotid-Polymorphismus.
8. "Validierung".

Die folgende Abbildung zeigt den Prozess von der Probenahme bis zur Variantenidentifikation, Anmerkung und Vorhersage.



Das Human Genome Projekt wurde im April 2003 abgeschlossen und das Projekt stellte eine sehr hochwertige Simulation der menschlichen Genom-Sequenz dar, die in der Öffentlichkeit zur Verfügung stand. Das Referenzgenom initiierte eine Explosion der Forschung und Entwicklung von Genomfunktionen. Praktisch jede menschliche Krankheit hat eine Signatur in den Genen des Menschen. Bis vor kurzem nutzen Ärzte Gene, um Geburtsfehler wie Sichelzellenanämie vorherzusagen und zu bestimmen, die durch ein bestimmtes Erbmuster verursacht wird, das durch eine Änderung in einem einzelnen Gen verursacht wird. Die Schatzkammer der Daten, die das Humangenom-Projekt zur Verfügung gestellt wurde, führte zu dem Beginn des aktuellen Status der Genomfunktionen.

Die Genomik bietet zahlreiche Vorteile. Hier ein kleiner Satz von Vorteilen in den Bereichen Gesundheitswesen und Life Sciences:

- Bessere Diagnose am Point of Care
- Bessere Prognose
- Präzisionsmedizin

- Personalisierte Behandlungspläne
- Bessere Krankheitsüberwachung
- Verringerung unerwünschter Ereignisse
- Besserer Zugang zu Therapien
- Verbesserte Krankheitsüberwachung
- Effektive Teilnahme an klinischen Studien und bessere Auswahl von Patienten für klinische Studien auf Basis von Genotypen.

Genomik ist eine "[Vierköpfige](#)," Aufgrund der Computing-Anforderungen für den gesamten Lebenszyklus eines Datensatzes, zu Erfassung, Storage, Verteilung und Analyse

Genom Analysis Toolkit (GATK)

GATK wurde als Datenwissenschaftsplattform am entwickelt "[Broad Institute](#)". GATK ist eine Reihe von Open-Source-Tools, die Genomanalysen ermöglichen, insbesondere Variantenerkennung, Identifizierung, Annotation und Genotyping. Einer der Vorteile von GATK besteht darin, dass der Satz von Tools und Befehlen zu einem kompletten Workflow gekettet werden kann. Die Hauptprobleme, mit denen sich das breite Institut befasst, sind:

- Die Ursachen und biologischen Mechanismen von Krankheiten verstehen.
- Identifizieren Sie therapeutische Interventionen, die auf die grundlegende Ursache einer Krankheit wirken.
- Verstehen Sie die Sichtlinie von Varianten bis zur Funktion in der menschlichen Physiologie.
- Standards und Richtlinien erstellen "[Frameworks](#)" Für die Darstellung von Genomdaten, Speicherung, Analysen, Sicherheit usw.
- Standardisieren und Sozialisieren interoperabler Genom Aggregation Datenbanken (gnomAD).
- Genom-basierte Überwachung, Diagnose und Behandlung von Patienten mit größerer Präzision.
- Helfen Sie bei der Implementierung von Tools, mit denen Krankheiten schon lange vorhergesagt werden, bevor Symptome auftreten.
- Schaffen und stärken Sie eine Gemeinschaft von interdisziplinären Mitarbeitern, um die schwierigsten und wichtigsten Probleme in der Biomedizin zu lösen.

Nach Angaben des GATK und des breiten Instituts sollte die Genomsequenzierung in einem Pathologielabor als Protokoll behandelt werden; jede Aufgabe ist gut dokumentiert, optimiert, reproduzierbar und konsistent über Proben und Experimente hinweg. Im Folgenden finden Sie eine Reihe von Schritten, die vom Broad Institute empfohlen werden. Weitere Informationen finden Sie im "[GATK-Website](#)".

Einrichtung von FlexPod

Für Genomik-Workloads wurde eine FlexPod Infrastrukturplattform von Grund auf neu eingerichtet. Die FlexPod Plattform ist hochverfügbar und lässt sich unabhängig skalieren, beispielsweise Netzwerk, Storage und Compute unabhängig voneinander skalieren. Wir verwendeten den folgenden Cisco Validated Design Leitfaden als Referenzarchitekturdokument zur Einrichtung der FlexPod Umgebung: "[FlexPod Datacenter with VMware vSphere 7.0 and NetApp ONTAP 9.7](#)". Sehen Sie sich die folgenden FlexPod Plattform-Highlights an:

Um die FlexPod Lab-Einrichtung durchzuführen, gehen Sie wie folgt vor:

1. Zur Einrichtung und Validierung von FlexPod kommen die folgenden IP4-Reservierungen und -VLANs zum Einsatz.

IP Reservations

VLAN	IP Range	Subnet Mask	Purpose
3281	172.21.25 /24	255.255.255.0	IB-MGMT
3282	172.21.26 /24	255.255.255.0	vMotion
3283	172.21.27 /24	255.255.255.0	VM
3284	172.21.28 /24	255.255.255.0	NFS
3285	172.21.29 /24	255.255.255.0	iSCSI-A
3286	172.21.30 /24	255.255.255.0	iSCSI-B

2. Konfigurieren Sie iSCSI-basierte Boot-LUNs auf der ONTAP SVM.

ONTAP System Manager

LUNs

[+ Add](#)

	Name	Storage VM
▼	ESXi_Boot_Lun_1	Healthcare_SVM
▼	ESXi_Boot_Lun_2	Healthcare_SVM
▼	ESXi_Boot_Lun_3	Healthcare_SVM
▼	ESXi_Boot_Lun_4	Healthcare_SVM
▼	ESXi_Boot_Lun_5	Healthcare_SVM
▼	ESXi_Boot_Lun_6	Healthcare_SVM

3. Zuordnen von LUNs zu iSCSI-Initiatorgruppen

	Name	Storage VM	Volume	Size	IOPS	Latency (ms)	Throughput (MB/s)
▲	ESXi_Boot_Lun_1	Healthcare_SVM	ESXi_Boot_Vol	20 GB	3	0.16	0.01

STATUS	VOLUME	DESCRIPTION	SNAPSHOT COPIES (LOCAL)	SNAPMIRROR (LOCAL OR REMOTE)
Online	ESXi_Boot_Vol	-	Protected	Unprotected

SERIAL NUMBER	QOS POLICY GROUP	MAPPED TO INITIATORS	ID
80A4X+R8rAhP	-	GenomicsESXi_1 (1) iqn.1992-08.com.cisco:ucs-...	0

CAPACITY (AVAILABLE % TOTAL)	LUN FORMAT	SNAPSHOT POLICY
95% 20 GB	VMware	default

PATH
/vol/ESXi_Boot_Vol/ESXi_Boot_Lun_1

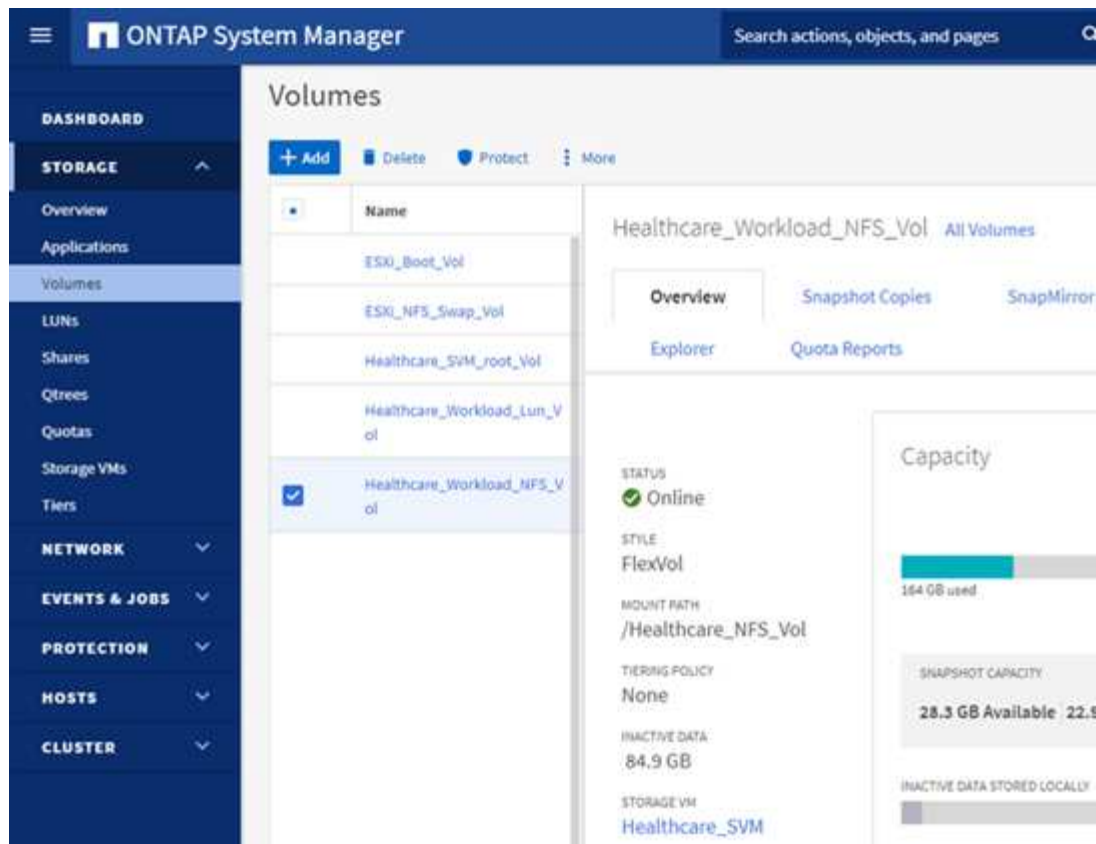
	Name	Storage VM	Volume	Size	IOPS	Latency (ms)	Throughput (MB/s)
▼	ESXi_Boot_Lun_1	Healthcare_SVM	ESXi_Boot_Vol	20 GB	1	0.25	0.01
▲	ESXi_Boot_Lun_2	Healthcare_SVM	ESXi_Boot_Vol	20 GB	4	0.18	0.02

STATUS ✔ Online	VOLUME ESXi_Boot_Vol	DESCRIPTION -	SNAPSHOT COPIES (LOCAL) STATUS ✔ Protected	SNAPMIRROR (LOCAL OR REMOTE) STATUS Unprotected
SERIAL NUMBER 80A4X+R8rAhU	QOS POLICY GROUP -	MAPPED TO INITIATORS ^ GenomicsESXi_2 (1) iqn.1992-08.com.cisco:ucs-...	ID 0	SNAPSHOT POLICY default
CAPACITY (AVAILABLE % TOTAL) 96% 20 GB	LUN FORMAT VMware			

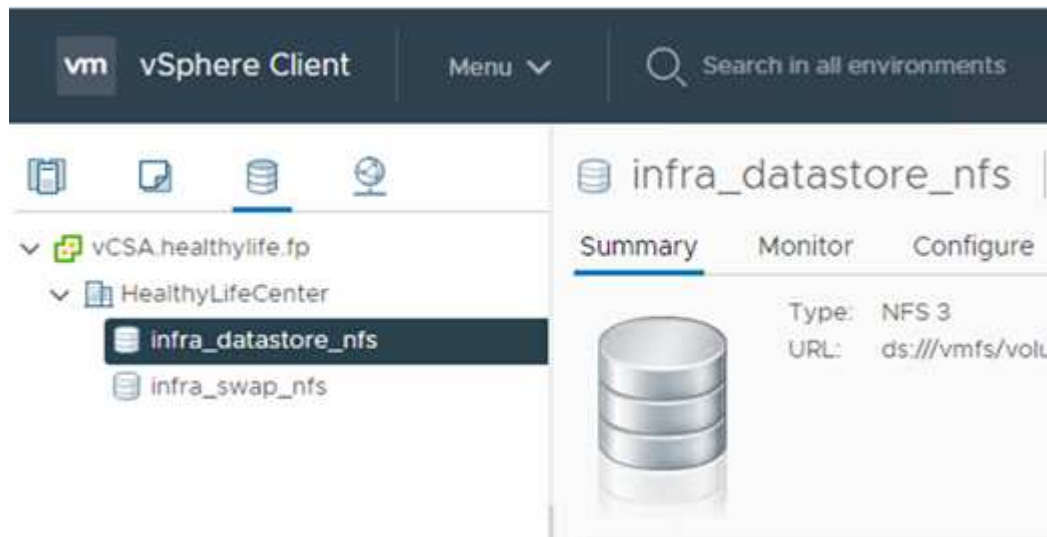
4. Installation von vSphere 7.0 mit iSCSI Boot
5. Registrieren Sie ESXi-Hosts mit vCenter.



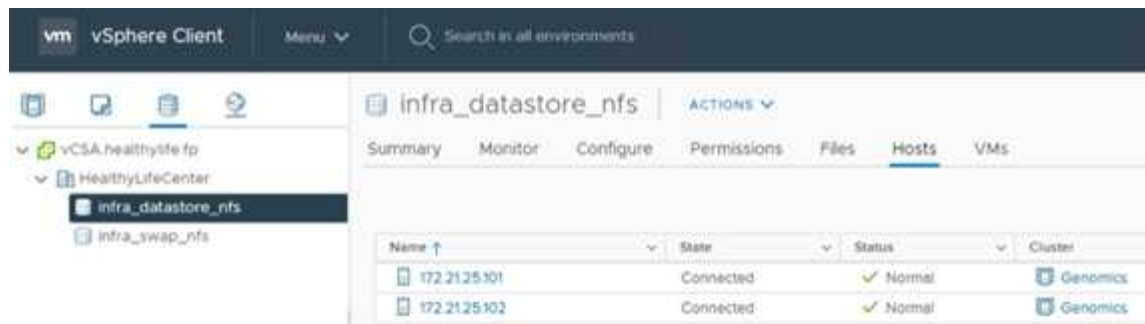
6. Bereitstellung eines NFS-Datenspeichers infra_datastore_nfs Auf dem ONTAP Storage.



7. Fügen Sie den Datastore zum vCenter hinzu.



8. Fügen Sie mithilfe von vCenter einen NFS-Datenspeicher zu den ESXi Hosts hinzu.



9. Erstellen Sie mithilfe von vCenter eine VM mit Red hat Enterprise Linux (RHEL) 8.3 zur Ausführung von GATK.
10. Ein NFS-Datstore wird der VM präsentiert und bei gemountet /mnt/genomics, Die zum Speichern von ausführbaren GATK-Dateien, Skripten, BAM-Dateien (Binary Alignment Map), Referenzdateien, Indexdateien, Wörterbuchdateien und Ausrufdateien für Variantenaufufe verwendet wird.

```
[root@genomics1 genomics]# df | grep genomics
/dev/sdb                308587328  5699492 287142812   2% /mnt/genomics
[root@genomics1 genomics]#
```

GATK-Einrichtung und -Ausführung

Installieren Sie die folgenden Voraussetzungen auf der RedHat Enterprise 8.3 Linux VM:

- Java 8 oder SDK 1.8 oder höher
- GATK 4.2.0.0 vom Broad Institute herunterladen "[GitHub-Website](#)". Genom-Sequenzdaten werden in der Regel in Form einer Reihe von tabulatorgetrennte ASCII-Spalten gespeichert. ASCII beansprucht jedoch zu viel Platz zum Speichern. Daher wurde ein neuer Standard entwickelt, der als BAM (*.bam)-Datei bezeichnet wird. Eine BAM-Datei speichert die Sequenzdaten in komprimierter, indizierter und binärer Form. Wir "[Heruntergeladen](#)" Eine Reihe öffentlich verfügbarer BAM-Dateien für die GATK-Ausführung vom "[Öffentliche Domäne](#)". Wir haben auch Indexdateien (*.bai), Wörterbuchdateien (*) heruntergeladen. Dict) und Referenzdatendateien (*. fasta) von der gleichen öffentlichen Domäne.

Nach dem Download verfügt das GATK-Tool-Kit über eine JAR-Datei und eine Reihe von Supportskripten.

- gatk-package-4.2.0.0-local.jar Ausführbar
- gatk Skriptdatei.

Wir haben die BAM-Dateien und die entsprechenden Index-, Wörterbuch- und Referenzgenom-Dateien für eine Familie heruntergeladen, die aus Vater-, Mutter- und Son *.bam-Dateien bestand.

Cromwell-Motor

Cromwell ist eine Open-Source-Engine, die auf wissenschaftliche Workflows ausgerichtet ist und Workflow-Management ermöglicht. Der Cromwell Motor kann in zwei laufen "[Modi](#)", Servermodus oder ein Einzelworkflowmodus. Das Verhalten des Cromwell-Motors kann mit dem gesteuert werden "[Cromwell Engine-Konfigurationsdatei](#)".

- **Servermodus.** aktiviert "[Rest-konforme](#)" Ausführung von Workflows in Cromwell Engine.
- **Run-Modus.** der Run-Modus eignet sich am besten zur Ausführung einzelner Workflows in Cromwell, "[ref](#)" Für einen vollständigen Satz verfügbarer Optionen im Run-Modus.

Wir nutzen die Cromwell Engine, um die Workflows und Pipelines nach Bedarf auszuführen. Die Cromwell Engine verwendet eine benutzerfreundliche "[Sprache für die Workflow-Beschreibung](#)" (WDL)-basierte Skriptsprache. Cromwell unterstützt auch einen zweiten Workflow-Skriptstandard, der als Common Workflow Language (CWL) bezeichnet wird. In diesem technischen Bericht wurde WDL verwendet. WDL wurde ursprünglich vom Broad Institute for Genome Analysis Pipelines entwickelt. Mithilfe der WDL-Workflows können verschiedene Strategien implementiert werden, darunter:

- **Linear Chaining.** wie der Name schon sagt, wird die Ausgabe von Task#1 als Eingabe an Task #2 gesendet.
- **Multi-in/out.** Dies ist ähnlich wie bei linearer Verkettung, da jede Aufgabe mehrere Ausgänge als Eingang zu nachfolgenden Aufgaben haben kann.
- **Scatter-Gather.** Dies ist eine der leistungsstärksten EAI-Strategien (Enterprise Application Integration), die zur Verfügung stehen, insbesondere bei ereignisgesteuerter Architektur. Jede Aufgabe wird entkoppelt ausgeführt, und die Ausgabe für jede Aufgabe wird in die Endausgabe konsolidiert.

Es gibt drei Schritte, wenn WDL zum Ausführen von GATK im Standalone-Modus verwendet wird:

1. Syntax validieren mit `womtool.jar`.

```
[root@genomics1 ~]# java -jar womtool.jar validate ghplo.wdl
```

2. Eingabe JSON generieren.

```
[root@genomics1 ~]# java -jar womtool.jar inputs ghplo.wdl > ghplo.json
```

3. Führen Sie den Workflow mit der Cromwell Engine und aus `Cromwell.jar`.

```
[root@genomics1 ~]# java -jar cromwell.jar run ghplo.wdl --inputs ghplo.json
```

Das GATK kann mit mehreren Methoden ausgeführt werden; dieses Dokument untersucht drei dieser Methoden.

Ausführung von GATK mit der JAR-Datei

Schauen wir uns eine einzelne Variante Call Pipeline-Ausführung unter Verwendung des haplotype Variant Caller an.

```
[root@genomics1 ~]# java -Dsamjdk.use_async_io_read_samtools=false \
-Dsamjdk.use_async_io_write_samtools=true \
-Dsamjdk.use_async_io_write_tribble=false \
-Dsamjdk.compression_level=2 \
-jar /mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-local.jar \
HaplotypeCaller \
--input /mnt/genomics/GATK/TEST\ DATA/bam/workshop_1906_2-
germline_bams_father.bam \
--output workshop_1906_2-germline_bams_father.validation.vcf \
--reference /mnt/genomics/GATK/TEST\ DATA/ref/workshop_1906_2-
germline_ref_ref.fasta
```

Bei dieser Methode der Ausführung verwenden wir die lokale GATK-Ausführungs-JAR-Datei, wir verwenden einen einzigen java-Befehl, um die JAR-Datei aufzurufen, und wir übergeben mehrere Parameter an den Befehl.

1. Dieser Parameter gibt an, dass wir den aufrufen HaplotypeCaller Variant Caller Pipeline.
2. -- input Gibt die Eingabe-BAM-Datei an.
3. --output Gibt die Variant-Ausgabedatei im Variantenaufrufformat (*.vcf) an ("ref").
4. Mit dem --reference Parameter, geben wir ein Referenzgenom weiter.

Nach der Ausführung sind die Ausgabedetails im Abschnitt zu finden ["Ausgabe zur Ausführung des GATK unter Verwendung der JAR-Datei."](#)

Ausführung von GATK mit ./gatk-Skript

Das GATK-Werkzeugkit kann mit dem ausgeführt werden ./gatk Skript: Untersuchen wir nun den folgenden Befehl:

```
[root@genomics1 execution]# ./gatk \
--java-options "-Xmx4G" \
HaplotypeCaller \
-I /mnt/genomics/GATK/TEST\ DATA/bam/workshop_1906_2-
germline_bams_father.bam \
-R /mnt/genomics/GATK/TEST\ DATA/ref/workshop_1906_2-
germline_ref_ref.fasta \
-O /mnt/genomics/GATK/TEST\ DATA/variants.vcf
```

Wir übergeben mehrere Parameter an den Befehl.

- Dieser Parameter gibt an, dass wir den aufrufen HaplotypeCaller Variant Caller Pipeline.
- -I Gibt die Eingabe-BAM-Datei an.
- -O Gibt die Variant-Ausgabedatei im Variantenaufrufformat (*.vcf) an ("ref").
- Mit dem -R Parameter, geben wir ein Referenzgenom weiter.

Nach der Ausführung sind die Ausgabedetails im Abschnitt zu finden

Ausführung von GATK mit Cromwell Engine

Wir verwenden die Cromwell-Engine, um die Ausführung des GATK zu verwalten. Schauen wir uns die Kommandozeile und ihre Parameter an.

```
[root@genomics1 genomics]# java -jar cromwell-65.jar \  
run /mnt/genomics/GATK/seq/ghplo.wdl \  
--inputs /mnt/genomics/GATK/seq/ghplo.json
```

Hier rufen wir den Java-Befehl auf, indem wir den übergeben `-jar` Parameter, um anzugeben, dass wir eine JAR-Datei ausführen möchten, z. B. `Cromwell-65.jar`. Der nächste Parameter wurde übergeben (`run`) Zeigt an, dass die Cromwell-Engine im Run-Modus läuft, die andere mögliche Option ist der Server-Modus. Der nächste Parameter lautet `*.wdl` Dass der Run-Modus zum Ausführen der Pipelines verwendet werden soll. Der nächste Parameter ist der Satz von Eingabeparametern für die ausgeführten Workflows.

Hier ist der Inhalt der `ghplo.wdl` Datei wie folgt aussehen:

```
[root@genomics1 seq]# cat ghplo.wdl  
workflow helloHaplotypeCaller {  
  call haplotypeCaller  
}  
task haplotypeCaller {  
  File GATK  
  File RefFasta  
  File RefIndex  
  File RefDict  
  String sampleName  
  File inputBAM  
  File bamIndex  
  command {  
    java -jar ${GATK} \  
      HaplotypeCaller \  
      -R ${RefFasta} \  
      -I ${inputBAM} \  
      -O ${sampleName}.raw.indels.snps.vcf  
  }  
  output {  
    File rawVCF = "${sampleName}.raw.indels.snps.vcf"  
  }  
}  
[root@genomics1 seq]#
```

Hier ist die entsprechende JSON-Datei mit den Eingaben zur Cromwell Engine.


```
[root@genomics1 seq]# cat ghplo.json
{
  "helloHaplotypeCaller.haplotypeCaller.GATK": "/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-local.jar",
  "helloHaplotypeCaller.haplotypeCaller.RefFasta": "/mnt/genomics/GATK/TEST DATA/ref/workshop_1906_2-germline_ref_ref.fasta",
  "helloHaplotypeCaller.haplotypeCaller.RefIndex": "/mnt/genomics/GATK/TEST DATA/ref/workshop_1906_2-germline_ref_ref.fasta.fai",
  "helloHaplotypeCaller.haplotypeCaller.RefDict": "/mnt/genomics/GATK/TEST DATA/ref/workshop_1906_2-germline_ref_ref.dict",
  "helloHaplotypeCaller.haplotypeCaller.sampleName": "fatherbam",
  "helloHaplotypeCaller.haplotypeCaller.inputBAM": "/mnt/genomics/GATK/TEST DATA/bam/workshop_1906_2-germline_bams_father.bam",
  "helloHaplotypeCaller.haplotypeCaller.bamIndex": "/mnt/genomics/GATK/TEST DATA/bam/workshop_1906_2-germline_bams_father.bai"
}
[root@genomics1 seq]#
```

Bitte beachten Sie, dass Cromwell für die Ausführung eine in-Memory-Datenbank verwendet. Nach der Ausführung ist das Ausgabungsprotokoll im Abschnitt zu sehen ["Ausgabe zur Ausführung von GATK mit Cromwell Engine."](#)

Eine umfassende Reihe von Schritten zur Ausführung des GATK finden Sie im ["GATK-Dokumentation"](#).

["Weiter: Ausgabe für die Ausführung von GATK mit der JAR-Datei."](#)

Ausgabe zur Ausführung des GATK unter Verwendung der JAR-Datei

["Früher: Genomik - GATK Einrichtung und Ausführung."](#)

Die Ausführung von GATK unter Verwendung der JAR-Datei hat folgende Probenausgabe erzeugt.

```
[root@genomics1 execution]# java -Dsamjdk.use_async_io_read_samtools=false \
-Dsamjdk.use_async_io_write_samtools=true \
-Dsamjdk.use_async_io_write_tribble=false \
-Dsamjdk.compression_level=2 \
-jar /mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-local.jar \
HaplotypeCaller \
--input /mnt/genomics/GATK/TEST\ DATA/bam/workshop_1906_2-germline_bams_father.bam \
--output workshop_1906_2-germline_bams_father.validation.vcf \
--reference /mnt/genomics/GATK/TEST\ DATA/ref/workshop_1906_2-
```

```

germline_ref_ref.fasta \
22:52:58.430 INFO NativeLibraryLoader - Loading libgkl_compression.so
from jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
Aug 17, 2021 10:52:58 PM
shaded.cloud_nio.com.google.auth.oauth2.ComputeEngineCredentials
runningOnComputeEngine
INFO: Failed to detect whether we are running on Google Compute Engine.
22:52:58.541 INFO HaplotypeCaller -
-----
22:52:58.542 INFO HaplotypeCaller - The Genome Analysis Toolkit (GATK)
v4.2.0.0
22:52:58.542 INFO HaplotypeCaller - For support and documentation go to
https://software.broadinstitute.org/gatk/
22:52:58.542 INFO HaplotypeCaller - Executing as
root@genomics1.healthylife.fp on Linux v4.18.0-305.3.1.el8_4.x86_64 amd64
22:52:58.542 INFO HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server
VM v1.8.0_302-b08
22:52:58.542 INFO HaplotypeCaller - Start Date/Time: August 17, 2021
10:52:58 PM EDT
22:52:58.542 INFO HaplotypeCaller -
-----
22:52:58.542 INFO HaplotypeCaller -
-----
22:52:58.542 INFO HaplotypeCaller - HTSJDK Version: 2.24.0
22:52:58.542 INFO HaplotypeCaller - Picard Version: 2.25.0
22:52:58.542 INFO HaplotypeCaller - Built for Spark Version: 2.4.5
22:52:58.542 INFO HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
22:52:58.543 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
22:52:58.543 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
22:52:58.543 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
22:52:58.543 INFO HaplotypeCaller - Deflater: IntelDeflater
22:52:58.543 INFO HaplotypeCaller - Inflater: IntelInflater
22:52:58.543 INFO HaplotypeCaller - GCS max retries/reopens: 20
22:52:58.543 INFO HaplotypeCaller - Requester pays: disabled
22:52:58.543 INFO HaplotypeCaller - Initializing engine
22:52:58.804 INFO HaplotypeCaller - Done initializing engine
22:52:58.809 INFO HaplotypeCallerEngine - Disabling physical phasing,
which is supported only for reference-model confidence output
22:52:58.820 INFO NativeLibraryLoader - Loading libgkl_utils.so from
jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_utils.so
22:52:58.821 INFO NativeLibraryLoader - Loading libgkl_pairhmm_omp.so

```

```

from jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_pairhmm_omp.so
22:52:58.854 INFO   IntelPairHmm - Using CPU-supported AVX-512 instructions
22:52:58.854 INFO   IntelPairHmm - Flush-to-zero (FTZ) is enabled when
running PairHMM
22:52:58.854 INFO   IntelPairHmm - Available threads: 16
22:52:58.854 INFO   IntelPairHmm - Requested threads: 4
22:52:58.854 INFO   PairHMM - Using the OpenMP multi-threaded AVX-
accelerated native PairHMM implementation
22:52:58.872 INFO   ProgressMeter - Starting traversal
22:52:58.873 INFO   ProgressMeter -           Current Locus   Elapsed Minutes
Regions Processed   Regions/Minute
22:53:00.733 WARN   InbreedingCoeff - InbreedingCoeff will not be
calculated at position 20:9999900 and possibly subsequent; at least 10
samples must have called genotypes
22:53:08.873 INFO   ProgressMeter -           20:17538652           0.2
58900              353400.0
22:53:17.681 INFO   HaplotypeCaller - 405 read(s) filtered by:
MappingQualityReadFilter
0 read(s) filtered by: MappingQualityAvailableReadFilter
0 read(s) filtered by: MappedReadFilter
0 read(s) filtered by: NotSecondaryAlignmentReadFilter
6628 read(s) filtered by: NotDuplicateReadFilter
0 read(s) filtered by: PassesVendorQualityCheckReadFilter
0 read(s) filtered by: NonZeroReferenceLengthAlignmentReadFilter
0 read(s) filtered by: GoodCigarReadFilter
0 read(s) filtered by: WellformedReadFilter
7033 total reads filtered
22:53:17.681 INFO   ProgressMeter -           20:63024652           0.3
210522             671592.9
22:53:17.681 INFO   ProgressMeter - Traversal complete. Processed 210522
total regions in 0.3 minutes.
22:53:17.687 INFO   VectorLoglessPairHMM - Time spent in setup for JNI call
: 0.010347438
22:53:17.687 INFO   PairHMM - Total compute time in PairHMM
computeLogLikelihoods() : 0.259172573
22:53:17.687 INFO   SmithWatermanAligner - Total compute time in java
Smith-Waterman : 1.27 sec
22:53:17.687 INFO   HaplotypeCaller - Shutting down engine
[August 17, 2021 10:53:17 PM EDT]
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCalle
r done. Elapsed time: 0.32 minutes.
Runtime.totalMemory()=5561122816
[root@genomics1 execution]#

```

Beachten Sie, dass sich die Ausgabedatei an dem nach der Ausführung angegebenen Speicherort befindet.

Ausgabe zur Ausführung des GATK mit dem Skript ./gatk

"Zurück: Ausgabe für die Ausführung von GATK mit der JAR-Datei."

Die Ausführung des GATK unter Verwendung des ./gatk Skript hat die folgende Musterausgabe erzeugt.

```
[root@genomics1 gatk-4.2.0.0]# ./gatk --java-options "-Xmx4G" \
HaplotypeCaller \
-I /mnt/genomics/GATK/TEST\ DATA/bam/workshop_1906_2-
germline_bams_father.bam \
-R /mnt/genomics/GATK/TEST\ DATA/ref/workshop_1906_2-
germline_ref_ref.fasta \
-O /mnt/genomics/GATK/TEST\ DATA/variants.vcf
Using GATK jar /mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false
-Dsamjdk.use_async_io_write_samtools=true
-Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2
-Xmx4G -jar /mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-local.jar
HaplotypeCaller -I /mnt/genomics/GATK/TEST DATA/bam/workshop_1906_2-
germline_bams_father.bam -R /mnt/genomics/GATK/TEST
DATA/ref/workshop_1906_2-germline_ref_ref.fasta -O /mnt/genomics/GATK/TEST
DATA/variants.vcf
23:29:45.553 INFO  NativeLibraryLoader - Loading libgkl_compression.so
from jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
Aug 17, 2021 11:29:45 PM
shaded.cloud_nio.com.google.auth.oauth2.ComputeEngineCredentials
runningOnComputeEngine
INFO: Failed to detect whether we are running on Google Compute Engine.
23:29:45.686 INFO  HaplotypeCaller -
-----
23:29:45.686 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK)
v4.2.0.0
23:29:45.686 INFO  HaplotypeCaller - For support and documentation go to
https://software.broadinstitute.org/gatk/
23:29:45.687 INFO  HaplotypeCaller - Executing as
root@genomics1.healthyliife.fp on Linux v4.18.0-305.3.1.el8_4.x86_64 amd64
23:29:45.687 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server
VM v11.0.12+7-LTS
23:29:45.687 INFO  HaplotypeCaller - Start Date/Time: August 17, 2021 at
11:29:45 PM EDT
23:29:45.687 INFO  HaplotypeCaller -
-----
```

```

23:29:45.687 INFO HaplotypeCaller -
-----
23:29:45.687 INFO HaplotypeCaller - HTSJDK Version: 2.24.0
23:29:45.687 INFO HaplotypeCaller - Picard Version: 2.25.0
23:29:45.687 INFO HaplotypeCaller - Built for Spark Version: 2.4.5
23:29:45.688 INFO HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
23:29:45.688 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
23:29:45.688 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
23:29:45.688 INFO HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
23:29:45.688 INFO HaplotypeCaller - Deflater: IntelDeflater
23:29:45.688 INFO HaplotypeCaller - Inflater: IntelInflater
23:29:45.688 INFO HaplotypeCaller - GCS max retries/reopens: 20
23:29:45.688 INFO HaplotypeCaller - Requester pays: disabled
23:29:45.688 INFO HaplotypeCaller - Initializing engine
23:29:45.804 INFO HaplotypeCaller - Done initializing engine
23:29:45.809 INFO HaplotypeCallerEngine - Disabling physical phasing,
which is supported only for reference-model confidence output
23:29:45.818 INFO NativeLibraryLoader - Loading libgkl_utils.so from
jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_utils.so
23:29:45.819 INFO NativeLibraryLoader - Loading libgkl_pairhmm_omp.so
from jar:file:/mnt/genomics/GATK/gatk-4.2.0.0/gatk-package-4.2.0.0-
local.jar!/com/intel/gkl/native/libgkl_pairhmm_omp.so
23:29:45.852 INFO IntelPairHmm - Using CPU-supported AVX-512 instructions
23:29:45.852 INFO IntelPairHmm - Flush-to-zero (FTZ) is enabled when
running PairHMM
23:29:45.852 INFO IntelPairHmm - Available threads: 16
23:29:45.852 INFO IntelPairHmm - Requested threads: 4
23:29:45.852 INFO PairHMM - Using the OpenMP multi-threaded AVX-
accelerated native PairHMM implementation
23:29:45.868 INFO ProgressMeter - Starting traversal
23:29:45.868 INFO ProgressMeter - Current Locus Elapsed Minutes
Regions Processed Regions/Minute
23:29:47.772 WARN InbreedingCoeff - InbreedingCoeff will not be
calculated at position 20:9999900 and possibly subsequent; at least 10
samples must have called genotypes
23:29:55.868 INFO ProgressMeter - 20:18885652 0.2
63390 380340.0
23:30:04.389 INFO HaplotypeCaller - 405 read(s) filtered by:
MappingQualityReadFilter
0 read(s) filtered by: MappingQualityAvailableReadFilter
0 read(s) filtered by: MappedReadFilter
0 read(s) filtered by: NotSecondaryAlignmentReadFilter

```

```

6628 read(s) filtered by: NotDuplicateReadFilter
0 read(s) filtered by: PassesVendorQualityCheckReadFilter
0 read(s) filtered by: NonZeroReferenceLengthAlignmentReadFilter
0 read(s) filtered by: GoodCigarReadFilter
0 read(s) filtered by: WellformedReadFilter
7033 total reads filtered
23:30:04.389 INFO ProgressMeter - 20:63024652 0.3
210522 681999.9
23:30:04.389 INFO ProgressMeter - Traversal complete. Processed 210522
total regions in 0.3 minutes.
23:30:04.395 INFO VectorLoglessPairHMM - Time spent in setup for JNI call
: 0.0121292030000000002
23:30:04.395 INFO PairHMM - Total compute time in PairHMM
computeLogLikelihoods() : 0.267345217
23:30:04.395 INFO SmithWatermanAligner - Total compute time in java
Smith-Waterman : 1.23 sec
23:30:04.395 INFO HaplotypeCaller - Shutting down engine
[August 17, 2021 at 11:30:04 PM EDT]
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCalle
r done. Elapsed time: 0.31 minutes.
Runtime.totalMemory()=2111832064
[root@genomics1 gatk-4.2.0.0]#

```

Beachten Sie, dass sich die Ausgabedatei an dem nach der Ausführung angegebenen Speicherort befindet.

["Weiter: Ausgabe für die Ausführung von GATK mit der Cromwell-Engine."](#)

Ausgabe zur Ausführung von GATK mit Cromwell Engine

Die Ausführung von GATK mit der Cromwell-Engine hat die folgende Probenausgabe erzeugt.

```

[root@genomics1 genomics]# java -jar cromwell-65.jar run
/mnt/genomics/GATK/seq/ghplo.wdl --inputs
/mnt/genomics/GATK/seq/ghplo.json
[2021-08-18 17:10:50,78] [info] Running with database db.url =
jdbc:hsqldb:mem:856a1f0d-9a0d-42e5-9199-
5e6c1d0f72dd;shutdown=false;hsqldb.tx=mvcc
[2021-08-18 17:10:57,74] [info] Running migration
RenameWorkflowOptionsInMetadata with a read batch size of 100000 and a
write batch size of 100000
[2021-08-18 17:10:57,75] [info] [RenameWorkflowOptionsInMetadata] 100%
[2021-08-18 17:10:57,83] [info] Running with database db.url =
jdbc:hsqldb:mem:6afe0252-2dc9-4e57-8674-
ce63c67aa142;shutdown=false;hsqldb.tx=mvcc

```

```

[2021-08-18 17:10:58,17] [info] Slf4jLogger started
[2021-08-18 17:10:58,33] [info] Workflow heartbeat configuration:
{
  "cromwellId" : "cromid-41b7e30",
  "heartbeatInterval" : "2 minutes",
  "ttl" : "10 minutes",
  "failureShutdownDuration" : "5 minutes",
  "writeBatchSize" : 10000,
  "writeThreshold" : 10000
}
[2021-08-18 17:10:58,38] [info] Metadata summary refreshing every 1
second.
[2021-08-18 17:10:58,38] [info] No metadata archiver defined in config
[2021-08-18 17:10:58,38] [info] No metadata deleter defined in config
[2021-08-18 17:10:58,40] [info] KvWriteActor configured to flush with
batch size 200 and process rate 5 seconds.
[2021-08-18 17:10:58,40] [info] WriteMetadataActor configured to flush
with batch size 200 and process rate 5 seconds.
[2021-08-18 17:10:58,44] [info] CallCacheWriteActor configured to flush
with batch size 100 and process rate 3 seconds.
[2021-08-18 17:10:58,44] [warn] 'docker.hash-lookup.gcr-api-queries-per-
100-seconds' is being deprecated, use 'docker.hash-lookup.gcr.throttle'
instead (see reference.conf)
[2021-08-18 17:10:58,54] [info] JobExecutionTokenDispenser - Distribution
rate: 50 per 1 seconds.
[2021-08-18 17:10:58,58] [info] SingleWorkflowRunnerActor: Version 65
[2021-08-18 17:10:58,58] [info] SingleWorkflowRunnerActor: Submitting
workflow
[2021-08-18 17:10:58,64] [info] Unspecified type (Unspecified version)
workflow 3e246147-b1a9-41dc-8679-319f81b7701e submitted
[2021-08-18 17:10:58,66] [info] SingleWorkflowRunnerActor: Workflow
submitted 3e246147-b1a9-41dc-8679-319f81b7701e
[2021-08-18 17:10:58,66] [info] 1 new workflows fetched by cromid-41b7e30:
3e246147-b1a9-41dc-8679-319f81b7701e
[2021-08-18 17:10:58,67] [info] WorkflowManagerActor: Starting workflow
3e246147-b1a9-41dc-8679-319f81b7701e
[2021-08-18 17:10:58,68] [info] WorkflowManagerActor: Successfully started
WorkflowActor-3e246147-b1a9-41dc-8679-319f81b7701e
[2021-08-18 17:10:58,68] [info] Retrieved 1 workflows from the
WorkflowStoreActor
[2021-08-18 17:10:58,70] [info] WorkflowStoreHeartbeatWriteActor
configured to flush with batch size 10000 and process rate 2 minutes.
[2021-08-18 17:10:58,76] [info] MaterializeWorkflowDescriptorActor
[3e246147]: Parsing workflow as WDL draft-2
[2021-08-18 17:10:59,34] [info] MaterializeWorkflowDescriptorActor
[3e246147]: Call-to-Backend assignments:

```

```

helloHaplotypeCaller.haplotypeCaller -> Local
[2021-08-18 17:11:00,54] [info] WorkflowExecutionActor-3e246147-b1a9-41dc-8679-319f81b7701e [3e246147]: Starting
helloHaplotypeCaller.haplotypeCaller
[2021-08-18 17:11:01,56] [info] Assigned new job execution tokens to the following groups: 3e246147: 1
[2021-08-18 17:11:01,70] [info] BackgroundConfigAsyncJobExecutionActor [3e246147helloHaplotypeCaller.haplotypeCaller:NA:1]: java -jar /mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-41dc-8679-319f81b7701e/call-haplotypeCaller/inputs/-179397211/gatk-package-4.2.0.0-local.jar \
    HaplotypeCaller \
    -R /mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-41dc-8679-319f81b7701e/call-haplotypeCaller/inputs/604632695/workshop_1906_2-germline_ref_ref.fasta \
    -I /mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-41dc-8679-319f81b7701e/call-haplotypeCaller/inputs/604617202/workshop_1906_2-germline_bams_father.bam \
    -O fatherbam.raw.indels.snps.vcf
[2021-08-18 17:11:01,72] [info] BackgroundConfigAsyncJobExecutionActor [3e246147helloHaplotypeCaller.haplotypeCaller:NA:1]: executing: /bin/bash /mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-41dc-8679-319f81b7701e/call-haplotypeCaller/execution/script
[2021-08-18 17:11:03,49] [info] BackgroundConfigAsyncJobExecutionActor [3e246147helloHaplotypeCaller.haplotypeCaller:NA:1]: job id: 26867
[2021-08-18 17:11:03,53] [info] BackgroundConfigAsyncJobExecutionActor [3e246147helloHaplotypeCaller.haplotypeCaller:NA:1]: Status change from - to WaitingForReturnCode
[2021-08-18 17:11:03,54] [info] Not triggering log of token queue status. Effective log interval = None
[2021-08-18 17:11:23,65] [info] BackgroundConfigAsyncJobExecutionActor [3e246147helloHaplotypeCaller.haplotypeCaller:NA:1]: Status change from WaitingForReturnCode to Done
[2021-08-18 17:11:25,04] [info] WorkflowExecutionActor-3e246147-b1a9-41dc-8679-319f81b7701e [3e246147]: Workflow helloHaplotypeCaller complete.
Final Outputs:
{
  "helloHaplotypeCaller.haplotypeCaller.rawVCF": "/mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-41dc-8679-319f81b7701e/call-haplotypeCaller/execution/fatherbam.raw.indels.snps.vcf"
}
[2021-08-18 17:11:28,43] [info] WorkflowManagerActor: Workflow actor for 3e246147-b1a9-41dc-8679-319f81b7701e completed with status 'Succeeded'. The workflow will be removed from the workflow store.
[2021-08-18 17:11:32,24] [info] SingleWorkflowRunnerActor workflow

```



```

finished with status 'Succeeded'.
{
  "outputs": {
    "helloHaplotypeCaller.haplotypeCaller.rawVCF":
"/mnt/genomics/cromwell-executions/helloHaplotypeCaller/3e246147-b1a9-
41dc-8679-319f81b7701e/call-
haplotypeCaller/execution/fatherbam.raw.indels.snps.vcf"
  },
  "id": "3e246147-b1a9-41dc-8679-319f81b7701e"
}
[2021-08-18 17:11:33,45] [info] Workflow polling stopped
[2021-08-18 17:11:33,46] [info] 0 workflows released by cromid-41b7e30
[2021-08-18 17:11:33,46] [info] Shutting down WorkflowStoreActor - Timeout
= 5 seconds
[2021-08-18 17:11:33,46] [info] Shutting down WorkflowLogCopyRouter -
Timeout = 5 seconds
[2021-08-18 17:11:33,46] [info] Shutting down JobExecutionTokenDispenser -
Timeout = 5 seconds
[2021-08-18 17:11:33,46] [info] Aborting all running workflows.
[2021-08-18 17:11:33,46] [info] JobExecutionTokenDispenser stopped
[2021-08-18 17:11:33,46] [info] WorkflowStoreActor stopped
[2021-08-18 17:11:33,47] [info] WorkflowLogCopyRouter stopped
[2021-08-18 17:11:33,47] [info] Shutting down WorkflowManagerActor -
Timeout = 3600 seconds
[2021-08-18 17:11:33,47] [info] WorkflowManagerActor: All workflows
finished
[2021-08-18 17:11:33,47] [info] WorkflowManagerActor stopped
[2021-08-18 17:11:33,64] [info] Connection pools shut down
[2021-08-18 17:11:33,64] [info] Shutting down SubWorkflowStoreActor -
Timeout = 1800 seconds
[2021-08-18 17:11:33,64] [info] Shutting down JobStoreActor - Timeout =
1800 seconds
[2021-08-18 17:11:33,64] [info] Shutting down CallCacheWriteActor -
Timeout = 1800 seconds
[2021-08-18 17:11:33,64] [info] SubWorkflowStoreActor stopped
[2021-08-18 17:11:33,64] [info] Shutting down ServiceRegistryActor -
Timeout = 1800 seconds
[2021-08-18 17:11:33,64] [info] Shutting down DockerHashActor - Timeout =
1800 seconds
[2021-08-18 17:11:33,64] [info] Shutting down IoProxy - Timeout = 1800
seconds
[2021-08-18 17:11:33,64] [info] CallCacheWriteActor Shutting down: 0
queued messages to process
[2021-08-18 17:11:33,64] [info] JobStoreActor stopped
[2021-08-18 17:11:33,64] [info] CallCacheWriteActor stopped
[2021-08-18 17:11:33,64] [info] KvWriteActor Shutting down: 0 queued

```

```

messages to process
[2021-08-18 17:11:33,64] [info] IoProxy stopped
[2021-08-18 17:11:33,64] [info] WriteMetadataActor Shutting down: 0 queued
messages to process
[2021-08-18 17:11:33,65] [info] ServiceRegistryActor stopped
[2021-08-18 17:11:33,65] [info] DockerHashActor stopped
[2021-08-18 17:11:33,67] [info] Database closed
[2021-08-18 17:11:33,67] [info] Stream materializer shut down
[2021-08-18 17:11:33,67] [info] WDL HTTP import resolver closed
[root@genomics1 genomics]#

```

"Als Nächstes: GPU-Setup."

GPU-Einrichtung

"Zurück: Ausgabe für die Ausführung von GATK mit der Cromwell-Engine."

Zum Zeitpunkt der Veröffentlichung bietet das GATK-Tool keine native Unterstützung für die GPU-basierte Ausführung vor Ort. Die folgende Einrichtung und Anleitung werden bereitgestellt, damit die Leser verstehen können, wie einfach es ist, FlexPod mit einer an der Rückseite montierten NVIDIA Tesla P6 GPU mit einer PCIe Mezzanine-Karte für GATK zu verwenden.

Zur Einrichtung der FlexPod Umgebung haben wir folgendes Cisco Validated Design (CVD) als Referenzarchitektur und Best Practice Guide verwendet. Damit können wir Applikationen ausführen, die GPUs verwenden.

- ["FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for Deep Learning"](#)

Kernpunkte dieses Setups:

1. Wir haben eine PCIe NVIDIA Tesla P6 GPU in einem Mezzanine-Steckplatz in den UCS B200 M5 Servern verwendet.

Equipment / Chassis / Chassis 1 / Servers / Server 1

< General **Inventory** Virtual Machines Installed Firmware CIMC Sessions SEL Logs VIF Paths Health >

< Motherboard CIMC CPUs **GPUs** Memory Adapters HBAs NICs iSCSI vNICs Security >

Advanced Filter Export Print

Name	ID	Model	Serial	Mode
Graphics Card 2	2	UCSB-GPU-P6-R	FCH212373V7	Compute

Equipment / Chassis / Chassis 1 / Servers / Server 2

< General **Inventory** Virtual Machines Installed Firmware CIMC Sessions SEL Logs VIF Paths Health >

< Motherboard CIMC CPUs **GPUs** Memory Adapters HBAs NICs iSCSI vNICs Security >

Advanced Filter Export Print

Name	ID	Model	Serial	Mode
Graphics Card 2	2	UCSB-GPU-P6-R	FCH212373Y1	Compute

- Wir haben uns für dieses Setup im NVIDIA Partner-Portal registriert und eine Evaluierungslizenz (auch als Berechtigung bekannt) erhalten, die GPUs im Compute-Modus verwenden kann.
- Wir haben die erforderliche NVIDIA vGPU-Software von der NVIDIA-Partner-Website heruntergeladen.
- Wir haben die Berechtigung heruntergeladen *.bin Datei von der NVIDIA-Partner-Website.
- Wir installierten einen NVIDIA vGPU-Lizenzserver und fügten die Berechtigungen unter Verwendung von dem auf dem Lizenzserver hinzu *.bin Datei wird von der NVIDIA-Partnerwebsite heruntergeladen.
- Stellen Sie sicher, dass Sie die richtige NVIDIA vGPU-Softwareversion für Ihre Implementierung im NVIDIA-Partnerportal auswählen. Für dieses Setup haben wir Treiberversion 460.73.02 verwendet.
- Mit diesem Befehl wird der installiert **"NVIDIA vGPU Manager"** In ESXi.

```
[root@localhost:~] esxcli software vib install -v
/vmfs/volumes/infra_datastore_nfs/nvidia/vib/NVIDIA_bootbank_NVIDIA-
VMware_ESXi_7.0_Host_Driver_460.73.02-1OEM.700.0.0.15525992.vib
Installation Result
Message: Operation finished successfully.
Reboot Required: false
VIBs Installed: NVIDIA_bootbank_NVIDIA-
VMware_ESXi_7.0_Host_Driver_460.73.02-1OEM.700.0.0.15525992
VIBs Removed:
VIBs Skipped:
```

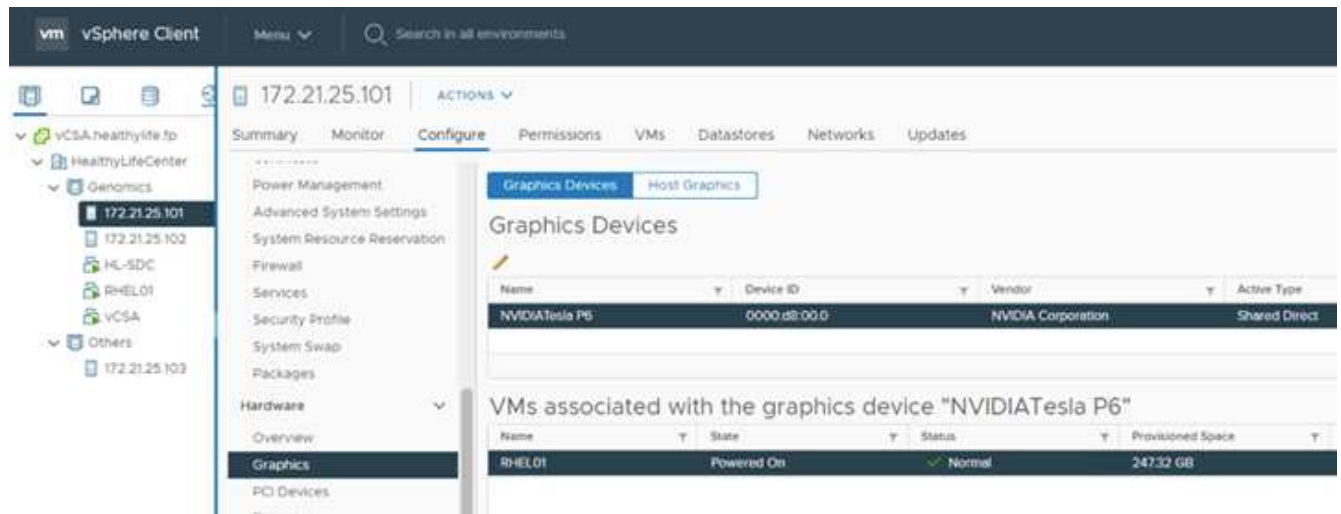
- Führen Sie nach dem Neubooten des ESXi-Servers den folgenden Befehl aus, um die Installation zu validieren und den Zustand der GPUs zu überprüfen.

```

[root@localhost:~] nvidia-smi
Wed Aug 18 21:37:19 2021
+-----+
+-----+
| NVIDIA-SMI 460.73.02      Driver Version: 460.73.02      CUDA Version: N/A
|
|-----+-----+
+-----+
| GPU  Name           Persistence-M| Bus-Id        Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage | GPU-Util
Compute M. |
|
|                               |
MIG M. |
|=====+=====+=====+
=====|
|   0  Tesla P6             On   | 00000000:D8:00.0 Off |
0 |
| N/A   35C    P8      9W /  90W | 15208MiB / 15359MiB |      0%
Default |
|
|                               |
N/A |
+-----+-----+
+-----+
+-----+
+-----+
+-----+
| Processes:
|
| GPU    GI    CI          PID    Type    Process name          GPU
Memory |
|        ID    ID                 |          Usage
|
|=====+=====+=====+
=====|
|   0    N/A   N/A     2812553    C+G    RHEL01
15168MiB |
+-----+-----+
+-----+
[root@localhost:~]

```

9. Mit vCenter "[Konfigurieren](#)" Die Einstellungen des Grafikgeräts auf „Shared Direct“.



10. Vergewissern Sie sich, dass der sichere Boot für die RedHat-VM deaktiviert ist.
11. Stellen Sie sicher, dass die Firmware für VM-Startoptionen auf EFI (gesetzt ist "ref").

Edit Settings
RHEL01

Virtual Hardware
VM Options

> General Options	VM Name: RHEL01
> VMware Remote Console Options	<input type="checkbox"/> Lock the guest operating system when the last remote user disconnects
> Encryption	Expand for encryption settings
> Power management	Expand for power management settings
> VMware Tools	Expand for VMware Tools settings
> Boot Options	
Firmware	EFI (recommended) ▼
Secure Boot	<input type="checkbox"/> Enabled
Boot Delay	When powering on or resetting, delay boot order by 0 milliseconds
Force EFI setup	<input type="checkbox"/> During the next boot, force entry into the EFI setup screen
Failed Boot Recovery	<input type="checkbox"/> If the VM fails to find boot device, automatically retry after 10 seconds
> Advanced	Expand for advanced settings
> Fibre Channel NPIV	Expand for Fibre Channel NPIV settings

CANCEL
OK

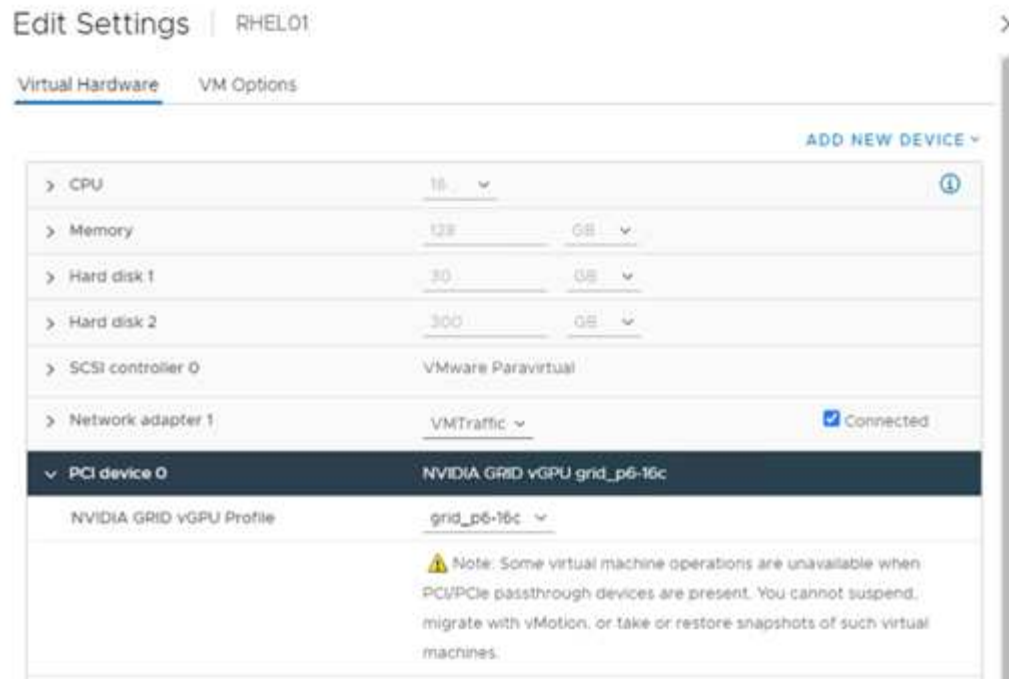
12. Stellen Sie sicher, dass die folgenden PARAMS zur erweiterten Konfiguration der VM-Optionen hinzugefügt werden. Der Wert des `pciPassthru.64bitMMIOSizeGB` Parameter hängt vom GPU-Speicher und der Anzahl der der VM zugewiesenen GPUs ab. Beispiel:
- Wenn einer VM 4 x 32-GB-V100-GPUs zugewiesen sind, sollte dieser Wert 128 sein.
 - Wenn einer VM 4 x 16-GB-P6-GPUs zugewiesen sind, sollte dieser Wert 64 sein.

✕

X

Name	Value
pciPassthru.64bitMMIOSizeGB	64
pciPassthru.use64bitMMIO	TRUE

13. Wenn Sie der virtuellen Maschine in vCenter vGPUs als neues PCI-Gerät hinzufügen, stellen Sie sicher, dass Sie NVIDIA GRID vGPU als PCI-Gerätetyp auswählen.
14. Wählen Sie das richtige GPU-Profil aus, das die verwendete GPU, den GPU-Speicher und den Nutzungszweck anführt, z. B. Grafik oder Rechner.



15. Auf der RedHat Linux VM können NVIDIA-Treiber installiert werden, indem Sie den folgenden Befehl ausführen:

```
[root@genomics1 genomics]# sh NVIDIA-Linux-x86_64-460.73.01-grid.run
```

16. Überprüfen Sie, ob das richtige vGPU-Profil angegeben wird, indem Sie den folgenden Befehl ausführen:

```
[root@genomics1 genomics]# nvidia-smi -query-gpu=gpu_name
-format=csv,noheader -id=0 | sed -e 's/ /-/g'
GRID-P6-16C
[root@genomics1 genomics]#
```

17. Überprüfen Sie nach dem Neubooten, ob die richtigen NVIDIA vGPU-Versionen zusammen mit den Treiberversionen gemeldet werden.


```

[root@genomics1 genomics]# nvidia-smi
Wed Aug 18 20:30:56 2021
+-----+
+-----+
| NVIDIA-SMI 460.73.01      Driver Version: 460.73.01      CUDA Version:
11.2      |
|-----+-----+
+-----+
| GPU  Name                Persistence-M| Bus-Id        Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage | GPU-Util
Compute M. |
|
MIG M. |
|=====+=====+=====
=====|
|   0  GRID P6-16C          On    | 00000000:02:02.0 Off |
N/A |
| N/A   N/A    P8      N/A /  N/A |   2205MiB / 16384MiB |      0%
Default |
|
N/A |
+-----+-----+
+-----+
+-----+
+-----+
+-----+
| Processes:
|
| GPU    GI    CI          PID    Type    Process name                        GPU
Memory |
|          ID    ID                                   Usage
|
|=====+=====+=====
=====|
|   0    N/A   N/A         8604      G    /usr/libexec/Xorg
13MiB |
+-----+-----+
+-----+
[root@genomics1 genomics]#

```

18. Stellen Sie sicher, dass die IP-Adresse des Lizenzservers auf der VM in der vGPU-Grid-Konfigurationsdatei konfiguriert ist.

a. Kopieren Sie die Vorlage.

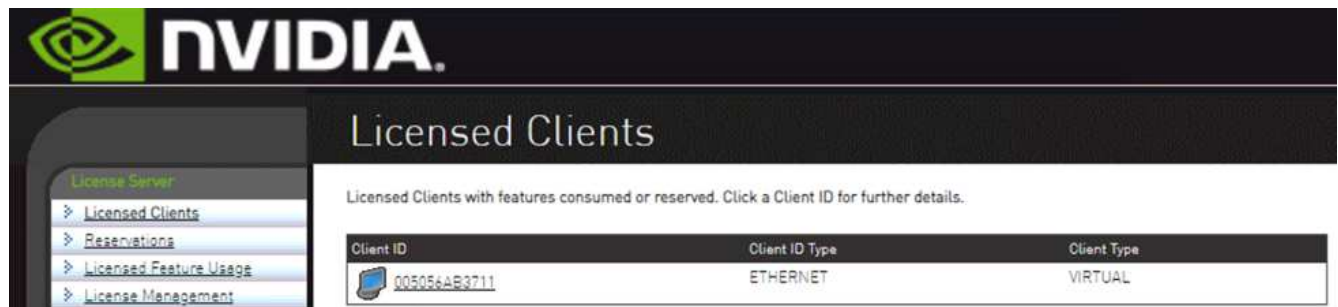
```
[root@genomics1 genomics]# cp /etc/nvidia/gridd.conf.template  
/etc/nvidia/gridd.conf
```

- b. Bearbeiten Sie die Datei `/etc/nvidia/rid.conf`, Fügen Sie die IP-Adresse des Lizenzservers hinzu und setzen Sie den Funktionstyp auf 1.

```
ServerAddress=192.168.169.10
```

```
FeatureType=1
```

19. Nach dem Neustart der VM sollten Sie einen Eintrag unter lizenzierte Clients im Lizenzserver sehen, wie unten gezeigt.



20. Weitere Informationen zum Herunterladen der Software GATK und Cromwell finden Sie im Abschnitt Solutions Setup.
21. Nachdem GATK GPUs vor Ort, die Workflow-Beschreibungssprache, verwenden kann *.wdl Enthält die Laufzeitattribute wie unten dargestellt.

```

task ValidateBAM {
  input {
    # Command parameters
    File input_bam
    String output_basename
    String? validation_mode
    String gatk_path
    # Runtime parameters
    String docker
    Int machine_mem_gb = 4
    Int additional_disk_space_gb = 50
  }
  Int disk_size = ceil(size(input_bam, "GB")) + additional_disk_space_gb
  String output_name = "${output_basename}_${validation_mode}.txt"
  command {
    ${gatk_path} \
      ValidateSamFile \
      --INPUT ${input_bam} \
      --OUTPUT ${output_name} \
      --MODE ${default="SUMMARY" validation_mode}
  }
  runtime {
    gpuCount: 1
    gpuType: "nvidia-tesla-p6"
    docker: docker
    memory: machine_mem_gb + " GB"
    disks: "local-disk " + disk_size + " HDD"
  }
  output {
    File validation_report = "${output_name}"
  }
}

```

["Weiter: Fazit."](#)

Schlussfolgerung

["Zurück: GPU-Setup."](#)

Viele Gesundheitseinrichtungen auf der ganzen Welt haben FlexPod als gemeinsame Plattform standardisiert. Mit FlexPod können Sie Funktionen im Gesundheitswesen zuverlässig implementieren. FlexPod mit NetApp ONTAP wird standardmäßig mit der Möglichkeit geliefert, eine sofort einsatzbereite Reihe branchenführender Protokolle zu implementieren. Unabhängig vom Ursprung der Anforderung, Genomik eines Patienten zu betreiben, verfügen Interoperabilität, Zugänglichkeit, Verfügbarkeit und Skalierbarkeit

standardmäßig über eine FlexPod-Plattform. Wenn sie auf einer FlexPod-Plattform standardisiert ist, wird die Innovationskultur ansteckend.

Wo Sie weitere Informationen finden

Sehen Sie sich die folgenden Dokumente und Websites an, um mehr über die in diesem Dokument beschriebenen Daten zu erfahren:

- FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for Deep Learning

["https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_480ml_aiml_deployement.pdf"](https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_480ml_aiml_deployement.pdf)

- FlexPod Datacenter with VMware vSphere 7.0 and NetApp ONTAP 9.7

["https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/fp_vmware_vsphere_7_0_ontap_9_7.html"](https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/fp_vmware_vsphere_7_0_ontap_9_7.html)

- ONTAP 9 Dokumentationszentrum

["http://docs.netapp.com"](http://docs.netapp.com)

- Agil und effizient – wie FlexPod die Modernisierung des Datacenters fördert

["https://www.flexpod.com/idc-white-paper/"](https://www.flexpod.com/idc-white-paper/)

- KI im Gesundheitswesen

["https://www.netapp.com/us/media/na-369.pdf"](https://www.netapp.com/us/media/na-369.pdf)

- FlexPod für das Gesundheitswesen vereinfachen den Wandel

["https://flexpod.com/solutions/verticals/healthcare/"](https://flexpod.com/solutions/verticals/healthcare/)

- FlexPod von Cisco und NetApp

["https://flexpod.com/"](https://flexpod.com/)

- KI- und Analyselösungen für das Gesundheitswesen (NetApp)

["https://www.netapp.com/us/artificial-intelligence/healthcare-ai-analytics/index.aspx"](https://www.netapp.com/us/artificial-intelligence/healthcare-ai-analytics/index.aspx)

- KI im Gesundheitswesen Intelligente Infrastrukturauswahlen führen zum Erfolg

<https://www.netapp.com/pdf.html?item=/media/7410-wp-7314.pdf>

- FlexPod Datacenter with ONTAP 9.8, ONTAP Storage Connector for Cisco Intersight und Cisco Intersight Managed Mode.

<https://www.netapp.com/pdf.html?item=/media/25001-tr-4883.pdf>

- FlexPod-Datacenter mit Red hat Enterprise Linux OpenStack Platform

["https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_openstack_osp6.html"](https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_openstack_osp6.html)

Versionsverlauf

Version	Datum	Versionsverlauf des Dokuments
Version 1.0	November 2021	Erste Version.

Copyright-Informationen

Copyright © 2025 NetApp. Alle Rechte vorbehalten. Gedruckt in den USA. Dieses urheberrechtlich geschützte Dokument darf ohne die vorherige schriftliche Genehmigung des Urheberrechtsinhabers in keiner Form und durch keine Mittel – weder grafische noch elektronische oder mechanische, einschließlich Fotokopieren, Aufnehmen oder Speichern in einem elektronischen Abrufsystem – auch nicht in Teilen, vervielfältigt werden.

Software, die von urheberrechtlich geschütztem NetApp Material abgeleitet wird, unterliegt der folgenden Lizenz und dem folgenden Haftungsausschluss:

DIE VORLIEGENDE SOFTWARE WIRD IN DER VORLIEGENDEN FORM VON NETAPP ZUR VERFÜGUNG GESTELLT, D. H. OHNE JEGLICHE EXPLIZITE ODER IMPLIZITE GEWÄHRLEISTUNG, EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE STILLSCHWEIGENDE GEWÄHRLEISTUNG DER MARKTGÄNGIGKEIT UND EIGNUNG FÜR EINEN BESTIMMTEN ZWECK, DIE HIERMIT AUSGESCHLOSSEN WERDEN. NETAPP ÜBERNIMMT KEINERLEI HAFTUNG FÜR DIREKTE, INDIREKTE, ZUFÄLLIGE, BESONDERE, BEISPIELHAFTE SCHÄDEN ODER FOLGESCHÄDEN (EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE BESCHAFFUNG VON ERSATZWAREN ODER -DIENSTLEISTUNGEN, NUTZUNGS-, DATEN- ODER GEWINNVERLUSTE ODER UNTERBRECHUNG DES GESCHÄFTSBETRIEBS), UNABHÄNGIG DAVON, WIE SIE VERURSACHT WURDEN UND AUF WELCHER HAFTUNGSTHEORIE SIE BERUHEN, OB AUS VERTRAGLICH FESTGELEGTER HAFTUNG, VERSCHULDENSUNABHÄNGIGER HAFTUNG ODER DELIKTSHAFTUNG (EINSCHLIESSLICH FAHRLÄSSIGKEIT ODER AUF ANDEREM WEGE), DIE IN IRGEND EINER WEISE AUS DER NUTZUNG DIESER SOFTWARE RESULTIEREN, SELBST WENN AUF DIE MÖGLICHKEIT DERARTIGER SCHÄDEN HINGEWIESEN WURDE.

NetApp behält sich das Recht vor, die hierin beschriebenen Produkte jederzeit und ohne Vorankündigung zu ändern. NetApp übernimmt keine Verantwortung oder Haftung, die sich aus der Verwendung der hier beschriebenen Produkte ergibt, es sei denn, NetApp hat dem ausdrücklich in schriftlicher Form zugestimmt. Die Verwendung oder der Erwerb dieses Produkts stellt keine Lizenzierung im Rahmen eines Patentrechts, Markenrechts oder eines anderen Rechts an geistigem Eigentum von NetApp dar.

Das in diesem Dokument beschriebene Produkt kann durch ein oder mehrere US-amerikanische Patente, ausländische Patente oder anhängige Patentanmeldungen geschützt sein.

ERLÄUTERUNG ZU „RESTRICTED RIGHTS“: Nutzung, Vervielfältigung oder Offenlegung durch die US-Regierung unterliegt den Einschränkungen gemäß Unterabschnitt (b)(3) der Klausel „Rights in Technical Data – Noncommercial Items“ in DFARS 252.227-7013 (Februar 2014) und FAR 52.227-19 (Dezember 2007).

Die hierin enthaltenen Daten beziehen sich auf ein kommerzielles Produkt und/oder einen kommerziellen Service (wie in FAR 2.101 definiert) und sind Eigentum von NetApp, Inc. Alle technischen Daten und die Computersoftware von NetApp, die unter diesem Vertrag bereitgestellt werden, sind gewerblicher Natur und wurden ausschließlich unter Verwendung privater Mittel entwickelt. Die US-Regierung besitzt eine nicht ausschließliche, nicht übertragbare, nicht unterlizenzierbare, weltweite, limitierte unwiderrufliche Lizenz zur Nutzung der Daten nur in Verbindung mit und zur Unterstützung des Vertrags der US-Regierung, unter dem die Daten bereitgestellt wurden. Sofern in den vorliegenden Bedingungen nicht anders angegeben, dürfen die Daten ohne vorherige schriftliche Genehmigung von NetApp, Inc. nicht verwendet, offengelegt, vervielfältigt, geändert, aufgeführt oder angezeigt werden. Die Lizenzrechte der US-Regierung für das US-Verteidigungsministerium sind auf die in DFARS-Klausel 252.227-7015(b) (Februar 2014) genannten Rechte beschränkt.

Markeninformationen

NETAPP, das NETAPP Logo und die unter <http://www.netapp.com/TM> aufgeführten Marken sind Marken von NetApp, Inc. Andere Firmen und Produktnamen können Marken der jeweiligen Eigentümer sein.