FlexPod per la genomica

FlexPod

NetApp
March 25, 2024

Sommario

FlexPod per la genomica . . . 1
TR-4911: Genomica FlexPod . . . 1
Vantaggi dell’implementazione di workload genomici su FlexPod . . . 3
Componenti hardware e software dell’infrastruttura della soluzione . . . 8
Genomica - Installazione ed esecuzione di GATK . . . 12
Output per l’esecuzione di GATK utilizzando il file jar . . . 22
Output per l’esecuzione di GATK utilizzando lo script ./gatk . . . 25
Output per l’esecuzione di GATK utilizzando il motore Cromwell . . . 27
Configurazione della GPU . . . 31
Conclusione . . . 40

FlexPod per la genomica

TR-4911: Genomica FlexPod

JayaKishore Esanakula, NetApp

Ci sono pochi campi della medicina che sono più importanti della genomica per l'assistenza sanitaria e le scienze biologiche, e la genomica sta rapidamente diventando uno strumento clinico chiave per medici e infermieri. La genomica, se combinata con l'imaging medico e la patologia digitale, ci aiuta a capire in che modo i geni di un paziente potrebbero essere influenzati dai protocolli di trattamento. Il successo della genomica nel settore sanitario dipende sempre più dall'interoperabilità dei dati su larga scala. L'obiettivo finale è quello di dare un senso agli enormi volumi di dati genetici e identificare correlazioni e varianti clinicamente rilevanti che migliorano la diagnosi e rendono la medicina di precisione una realtà. La genomica ci aiuta a comprendere l'origine dei focolai di malattia, come evolvono le malattie e quali trattamenti e strategie potrebbero essere efficaci. Chiaramente, la genomica ha molti benefici che spaziano dalla prevenzione alla diagnosi e al trattamento. Le organizzazioni del settore sanitario si trovano ad affrontare diverse sfide, tra cui:

- Migliore qualità dell'assistenza
- Assistenza basata sul valore
- Esplosione dei dati
- Medicina di precisione
- Pandemie
- Dispositivi indossabili, monitoraggio remoto e assistenza
- Sicurezza informatica

Percorsi clinici e protocolli clinici standardizzati sono uno dei componenti critici della medicina moderna. Uno degli aspetti chiave della standardizzazione è l'interoperabilità tra gli operatori sanitari, non solo per le cartelle cliniche, ma anche per i dati genomici. La domanda principale è che le organizzazioni sanitarie cederanno la proprietà dei dati genomici al posto della proprietà dei pazienti dei dati personali di genomica e delle relative cartelle mediche?

L'interoperabilità dei dati dei pazienti è fondamentale per la medicina di precisione, una delle forze trainanti della recente esplosione della crescita dei dati. L'obiettivo della medicina di precisione è quello di rendere più efficaci e precise le soluzioni di manutenzione della salute, prevenzione delle malattie, diagnosi e trattamento.

Il tasso di crescita dei dati è stato esponenziale. All'inizio di febbraio 2021, i laboratori statunitensi hanno sequenziato circa 8,000 ceppi COVID-19 alla settimana. Il numero di genomi sequenziati era aumentato a 29,000 alla settimana entro aprile 2021. Ogni genoma umano completamente sequenziato ha una dimensione di circa 125 GB. Pertanto, con un tasso di 29,000 genomi sequenziati alla settimana, lo storage totale del genoma a riposo sarebbe superiore a 180 petabyte all'anno. Diversi paesi hanno impegnato risorse per l'epidemiologia genomica per migliorare la sorveglianza genomica e prepararsi alla prossima ondata di sfide sanitarie globali.

Il costo ridotto della ricerca genomica sta portando a test genetici e ricerca a un ritmo senza precedenti. I tre

PS si trovano a un punto di svolta: Potenza del computer, privacy dei dati e personalizzazione della medicina. Entro il 2025 i ricercatori stimano che 100 milioni fino a 2 miliardi di genomi umani saranno sequenziati. Affinché la genomica sia efficace e una proposta preziosa, le funzionalità di genomica devono essere parte integrante dei flussi di lavoro di cura; devono essere facilmente accessibili e utilizzabili durante la visita di un paziente. Inoltre, è altrettanto importante integrare i dati medici elettronici dei pazienti con i dati genomici dei pazienti. Con l'avvento di un'infrastruttura convergente all'avanguardia come FlexPod, le organizzazioni possono introdurre le proprie funzionalità di genomica nei flussi di lavoro quotidiani di medici, infermieri e responsabili delle cliniche. Per informazioni aggiornate sulla piattaforma FlexPod, consulta questa pagina "White paper su FlexPod Datacenter con Cisco UCS serie X.".

Per un medico, il vero valore della genomica include la medicina di precisione e piani di trattamento personalizzati in base ai dati genomici di un paziente. In passato, non c'è mai stata una tale sinergia tra medici e data scientist, e la genomica sta beneficiando delle innovazioni tecnologiche del recente passato, oltre a partnership reali tra le organizzazioni sanitarie e i leader tecnologici del settore.

I centri medici accademici e le altre organizzazioni di settore sanitario e delle scienze della vita sono sulla buona strada per stabilire il centro di eccellenza (COE) nella scienza del genoma. Secondo il Dr. Charlie Gersbach, Dr Greg Crawford e il Dr. Tim e Reddy della Duke University, "sappiamo che i geni non vengono attivati o disattivati da un semplice switch binario, ma sono invece il risultato di più switch di regolazione dei geni che funzionano insieme. " Hanno anche determinato che "nessuna di queste parti del genoma funziona in isolamento. Il genoma è un web molto complicato che l'evoluzione ha intessuto" ("rif").

NetApp e Cisco si sono adoperati per implementare miglioramenti incrementali nella piattaforma FlexPod da oltre 10 anni. Tutti i commenti dei clienti vengono ascoltati, valutati e legati ai flussi di valore e ai set di funzionalità di FlexPod. È questo continuo loop di feedback, collaborazioni, miglioramenti e festeggiamenti che contraddistingue FlexPod come una piattaforma di infrastruttura convergente affidabile in tutto il mondo. È stata semplificata e progettata da zero per essere la piattaforma più affidabile, robusta, versatile e agile per le organizzazioni sanitarie.

Scopo

La piattaforma di infrastruttura convergente FlexPod consente a un'organizzazione sanitaria di ospitare uno o più carichi di lavoro di genomica, insieme ad altre applicazioni sanitarie cliniche e non cliniche. Questo report tecnico utilizza uno strumento di genomica open-source e standard di settore chiamato GATK durante la convalida della piattaforma FlexPod. Tuttavia, una discussione più approfondita sulla genomica o sul GATK non rientra nell'ambito di questo documento.

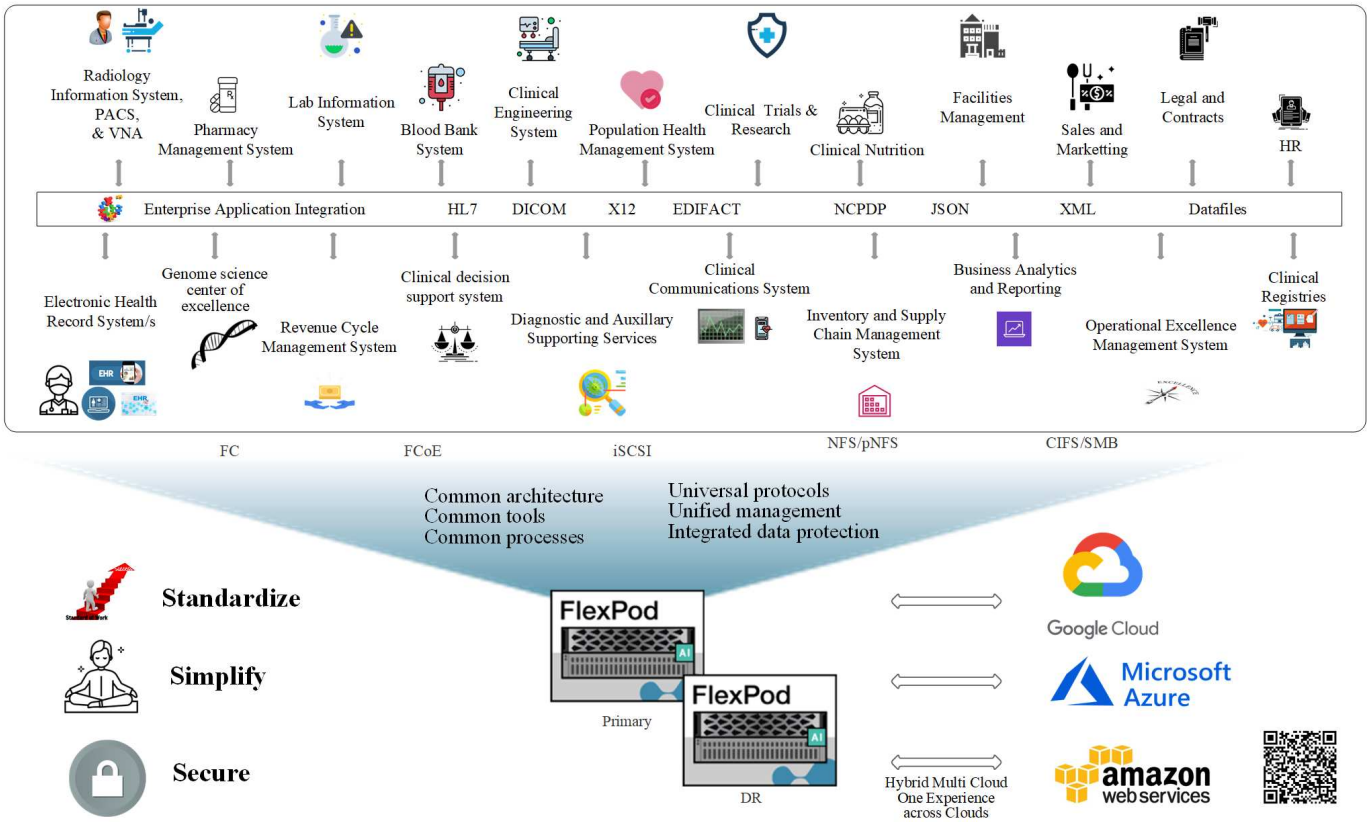
Pubblico

Il presente documento è destinato ai responsabili tecnici del settore sanitario, ai tecnici delle soluzioni partner Cisco e NetApp e al personale dei servizi professionali. NetApp presuppone che il lettore abbia una buona comprensione dei concetti di dimensionamento di calcolo e storage, nonché una familiarità tecnica con le minacce per il settore sanitario, la sicurezza sanitaria, i sistemi IT per il settore sanitario, Cisco UCS e i sistemi storage NetApp.

Funzionalità ospedaliere implementate su FlexPod

Un ospedale tipico dispone di un insieme diversificato di sistemi IT. La maggior parte di questi sistemi viene acquistata da un vendor, mentre pochissimi sono costruiti dal sistema ospedaliero in casa. Pertanto, il sistema ospedaliero deve gestire un ambiente di infrastruttura diversificato nei propri data center. Quando gli ospedali unificano i propri sistemi in una piattaforma di infrastruttura convergente come FlexPod, le organizzazioni possono standardizzare le operazioni del data center. Con FlexPod, le organizzazioni sanitarie possono implementare sistemi clinici e non clinici sulla stessa piattaforma, unificando in tal modo le operazioni del data center.

Hospital capabilities deployed on a FlexPod



"Avanti: Vantaggi dell'implementazione di workload genomici su FlexPod."

Vantaggi dell'implementazione di workload genomici su FlexPod

"Precedente: Introduzione."

Questa sezione fornisce un breve elenco dei vantaggi per l'esecuzione di un carico di lavoro genomico su una piattaforma di infrastruttura convergente FlexPod. Descriviamo rapidamente le funzionalità di un ospedale. La seguente vista dell'architettura di business mostra le funzionalità di un ospedale implementate su una piattaforma di infrastruttura convergente FlexPod ibrida-pronta per il cloud.

- **Evitare i silos nell'assistenza sanitaria.** i silos nell'assistenza sanitaria sono una preoccupazione molto reale. I reparti vengono spesso inseriti in silos nel proprio set di hardware e software, non per scelta, ma organicamente per evoluzione. Ad esempio, radiologia, cardiologia, EHR, genomica, analytics, ciclo di ricavi e altri reparti finiscono con il loro set individuale di software e hardware dedicati. Le organizzazioni del settore sanitario gestiscono un numero limitato di professionisti IT per gestire le proprie risorse hardware e software. Il punto di flessione si verifica quando si prevede che questo insieme di individui gestisca un insieme molto diversificato di hardware e software. L'eterogeneità è aggravata da un insieme incongruente di processi portati all'organizzazione sanitaria dai vendor.
- **Inizia con le piccole dimensioni e fai crescere.** il kit di tool GATK è ottimizzato per l'esecuzione della CPU, che offre le migliori suite di piattaforme come FlexPod. FlexPod consente una scalabilità indipendente di rete, calcolo e storage. Inizia con le dimensioni ridotte e scala man mano che le tue

capacità di genomica e l'ambiente crescono. Le organizzazioni del settore sanitario non devono investire in piattaforme specializzate per eseguire carichi di lavoro genomici. Le organizzazioni possono invece sfruttare piattaforme versatili come FlexPod per eseguire carichi di lavoro di genomica e non genomica sulla stessa piattaforma. Ad esempio, se il reparto di pediatria desidera implementare la funzionalità di genomica, la leadership IT può eseguire il provisioning di calcolo, storage e networking su un'istanza di FlexPod esistente. Man mano che la business unit genomica cresce, le organizzazioni sanitarie possono scalare la propria piattaforma FlexPod in base alle esigenze.

- **Pannello di controllo singolo e flessibilità senza pari.** Cisco Intersight semplifica significativamente le operazioni IT attraverso il bridging delle applicazioni con l'infrastruttura, fornendo visibilità e gestione da server bare-metal e hypervisor ad applicazioni senza server, riducendo così i costi e mitigando i rischi. Questa piattaforma SaaS unificata utilizza un design Open API unificato che si integra in modo nativo con piattaforme e tool di terze parti. Inoltre, consente la gestione del tuo team operativo del data center on-site o da qualsiasi luogo utilizzando un'app mobile.

  Gli utenti possono ottenere rapidamente valore tangibile nel proprio ambiente sfruttando Intersight come piattaforma di gestione. Grazie all'automazione per molte attività manuali quotidiane, Intersight elimina gli errori e semplifica le operazioni quotidiane. Inoltre, le funzionalità di supporto avanzate di Intersight consentono agli utenti di restare al passo con i problemi e accelerare la risoluzione dei problemi. In combinazione, le organizzazioni dedicano molto meno tempo e denaro alla propria infrastruttura applicativa e più tempo allo sviluppo del business principale.

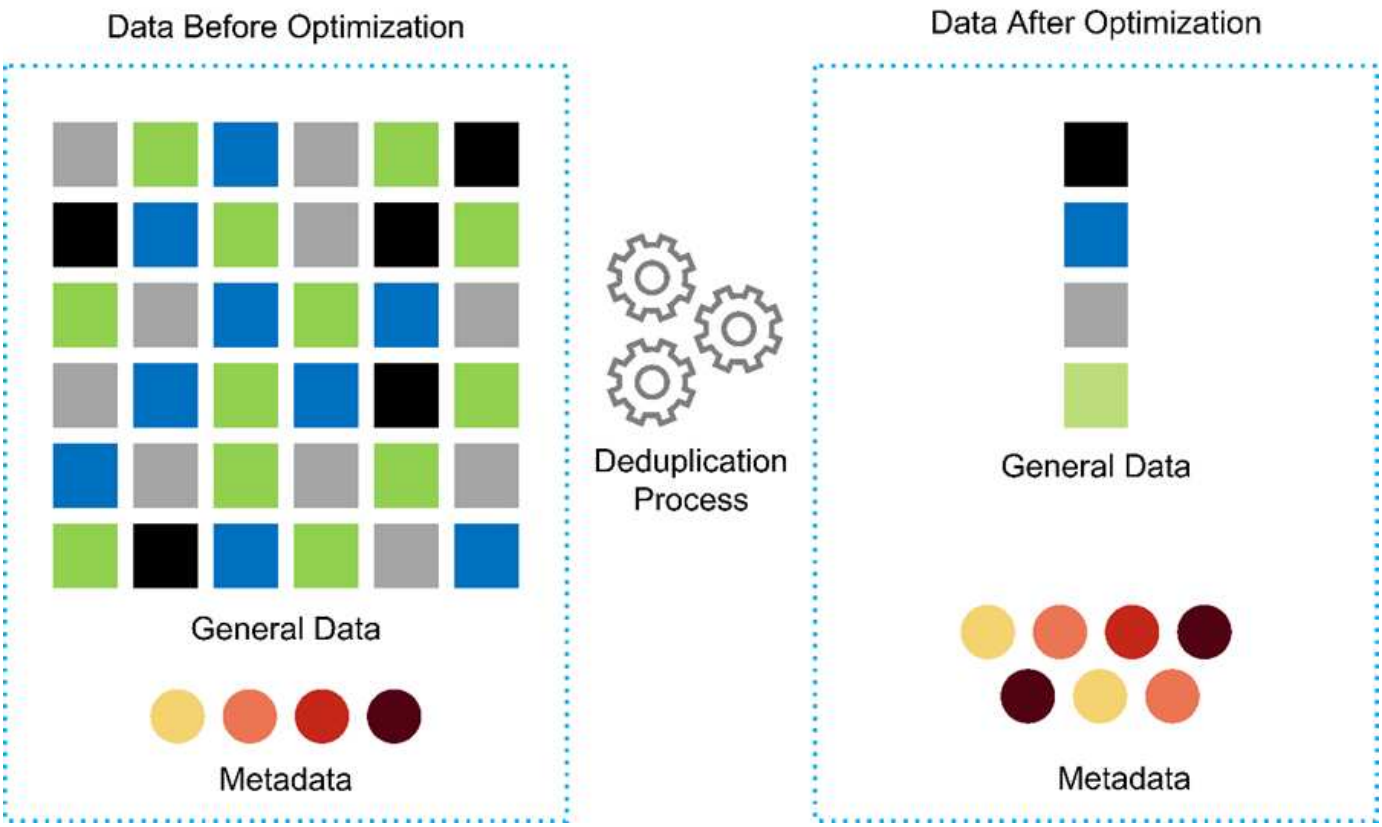
  Sfruttando la gestione di Intersight e l'architettura facilmente scalabile di FlexPod, le organizzazioni possono eseguire diversi carichi di lavoro genoma su una singola piattaforma FlexPod, aumentando l'utilizzo e riducendo il TCO (Total Cost of Ownership). FlexPod consente un dimensionamento flessibile, con scelte a partire dal nostro piccolo FlexPod Express e scalabilità in implementazioni di grandi dimensioni di FlexPod Datacenter. Grazie alle funzionalità di controllo degli accessi basate sui ruoli integrate in Cisco Intersight, le organizzazioni sanitarie possono implementare solidi meccanismi di controllo degli accessi, evitando la necessità di stack di infrastruttura separati. Più business unit all'interno dell'organizzazione sanitaria possono sfruttare la genomica come competenza chiave.

  In ultima analisi, FlexPod aiuta a semplificare le operazioni IT e a ridurre i costi operativi, consentendo agli amministratori dell'infrastruttura IT di concentrarsi su attività che aiutano i medici a innovare, anziché essere relegati per tenere le luci accese.

- **Progettazione validata e risultati garantiti.** le guide di progettazione e implementazione di FlexPod sono validate per essere ripetibili e coprono dettagli completi di configurazione e Best practice di settore che sono necessarie per implementare un FlexPod in tutta sicurezza. Le guide alla progettazione, le guide all'implementazione e le architetture validate di Cisco e NetApp aiutano la tua organizzazione sanitaria o di life science a eliminare ogni incertezza dall'implementazione di una piattaforma validata e affidabile fin dall'inizio. Con FlexPod, puoi accelerare i tempi di implementazione e ridurre costi, complessità e rischi. I design validati e le guide all'implementazione di FlexPod definiscono FlexPod come la piattaforma ideale per una vasta gamma di carichi di lavoro di genomica.
- **Innovazione e agilità.** FlexPod è una piattaforma ideale per gli EHR come Epic, Cerner, Meditech e sistemi di imaging come Agfa, GE, Philips. Per ulteriori informazioni su "EPIC Honor roll" E l'architettura della piattaforma di destinazione, vedi Epic userweb. Esecuzione di genomica su "FlexPod" consente alle organizzazioni del settore sanitario di continuare il proprio percorso di innovazione con agilità. Con FlexPod, l'implementazione del cambiamento organizzativo è naturale. Quando le organizzazioni si standardizzano su una piattaforma FlexPod, gli esperti IT del settore sanitario possono fornire tempo, impegno e risorse per innovare e quindi essere agili come richiesto dall'ecosistema.
- **Data liberated.** con la piattaforma di infrastruttura convergente FlexPod e un sistema storage NetApp ONTAP, i dati genomici possono essere resi disponibili e accessibili utilizzando un'ampia varietà di protocolli su larga scala da una singola piattaforma. FlexPod con NetApp ONTAP offre una piattaforma di cloud ibrido semplice, intuitiva e potente. Il data fabric basato su NetApp ONTAP consente di unire i dati tra

siti, oltre i confini fisici e tra applicazioni diverse. Il tuo data fabric è costruito per le aziende basate sui dati in un mondo incentrato sui dati. I dati vengono creati e utilizzati in più sedi e spesso devono essere sfruttati e condivisi con altre sedi, applicazioni e infrastrutture. Pertanto, è necessario un metodo coerente e integrato per gestirlo. FlexPod mette il tuo team IT sotto controllo e semplifica l'aumento della complessità DELL'IT.

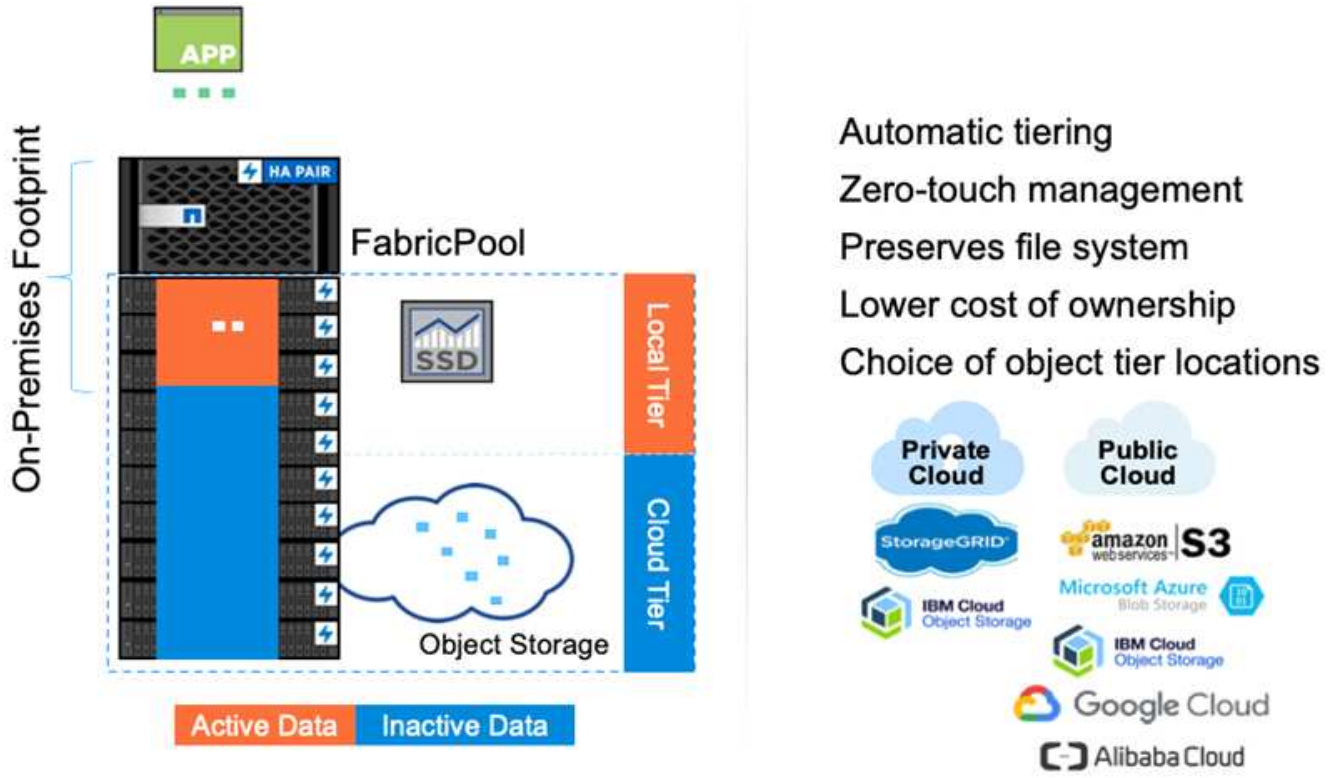
- **Multitenancy sicura.** FlexPod utilizza moduli crittografici conformi a FIPS 140-2, consentendo alle organizzazioni di implementare la sicurezza come elemento fondamentale, non come elemento secondario. FlexPod consente alle organizzazioni di implementare la multi-tenancy sicura da una singola piattaforma di infrastruttura convergente indipendentemente dalle dimensioni della piattaforma. FlexPod con multi-tenancy e QoS sicuri aiuta a separare i carichi di lavoro e a massimizzare l'utilizzo. In questo modo si evita che il capitale venga bloccato in piattaforme specializzate potenzialmente sottoutilizzate e che richiedono un set di competenze specialistiche da gestire.
- **Efficienza dello storage.** la genomica richiede che lo storage sottostante disponga di funzionalità di efficienza dello storage leader del settore. È possibile ridurre i costi dello storage con le funzionalità di efficienza dello storage NetApp come la deduplica (inline e on demand), la compressione dei dati e la compattazione dei dati ("rif"). La deduplica NetApp fornisce la deduplica a livello di blocco in un volume FlexVol. Essenzialmente, la deduplica rimuove i blocchi duplicati, memorizzando solo blocchi univoci nel volume FlexVol. La deduplica funziona con un elevato grado di granularità e opera sul file system attivo del volume FlexVol. La figura seguente mostra una panoramica del funzionamento della deduplica NetApp. La deduplica è trasparente per l'applicazione. Pertanto, può essere utilizzato per deduplicare i dati provenienti da qualsiasi applicazione che utilizzi il sistema NetApp. È possibile eseguire la deduplica del volume come processo inline e come processo in background. È possibile configurarlo in modo che venga eseguito automaticamente, pianificato o eseguito manualmente tramite CLI, Gestore di sistema NetApp ONTAP o NetApp Active IQ Unified Manager.



- **Abilitare l'interoperabilità genomica.** ONTAP FlexCache è una funzionalità di caching remoto che semplifica la distribuzione dei file, riduce la latenza della WAN e riduce i costi di larghezza di banda della WAN (. "rif"). Una delle attività chiave durante l'identificazione e l'annotazione delle varianti genomiche è la collaborazione tra i medici. La tecnologia ONTAP FlexCache aumenta il throughput dei dati anche quando i

medici collaboratori si trovano in diverse aree geografiche. Data la dimensione tipica di un file *.BAM (da 1 GB a 100 GB), è fondamentale che la piattaforma sottostante possa rendere i file disponibili ai medici in diverse aree geografiche. FlexPod con ONTAP FlexCache rende i dati genomici e le applicazioni realmente multisito, il che rende perfetta la collaborazione tra ricercatori dislocati in tutto il mondo, con bassa latenza e throughput elevato. Le organizzazioni sanitarie che eseguono applicazioni di genomica in un ambiente multisito possono scalare in orizzontale utilizzando il data fabric per bilanciare gestibilità con costi e velocità.

- **Uso intelligente della piattaforma di storage.** FlexPod con il tiering automatico ONTAP e la tecnologia Fabric Pool di NetApp semplificano la gestione dei dati. FabricPool aiuta a ridurre i costi dello storage senza compromettere performance, efficienza, sicurezza o protezione. FabricPool è trasparente per le applicazioni aziendali e sfrutta l'efficienza del cloud riducendo il TCO dello storage senza la necessità di riprogettare l'infrastruttura applicativa. FlexPod può trarre vantaggio dalle funzionalità di tiering dello storage di FabricPool per un utilizzo più efficiente dello storage flash ONTAP. Per ulteriori informazioni, vedere "FlexPod con FabricPool". Il seguente diagramma fornisce una panoramica di alto livello di FabricPool e dei suoi vantaggi.



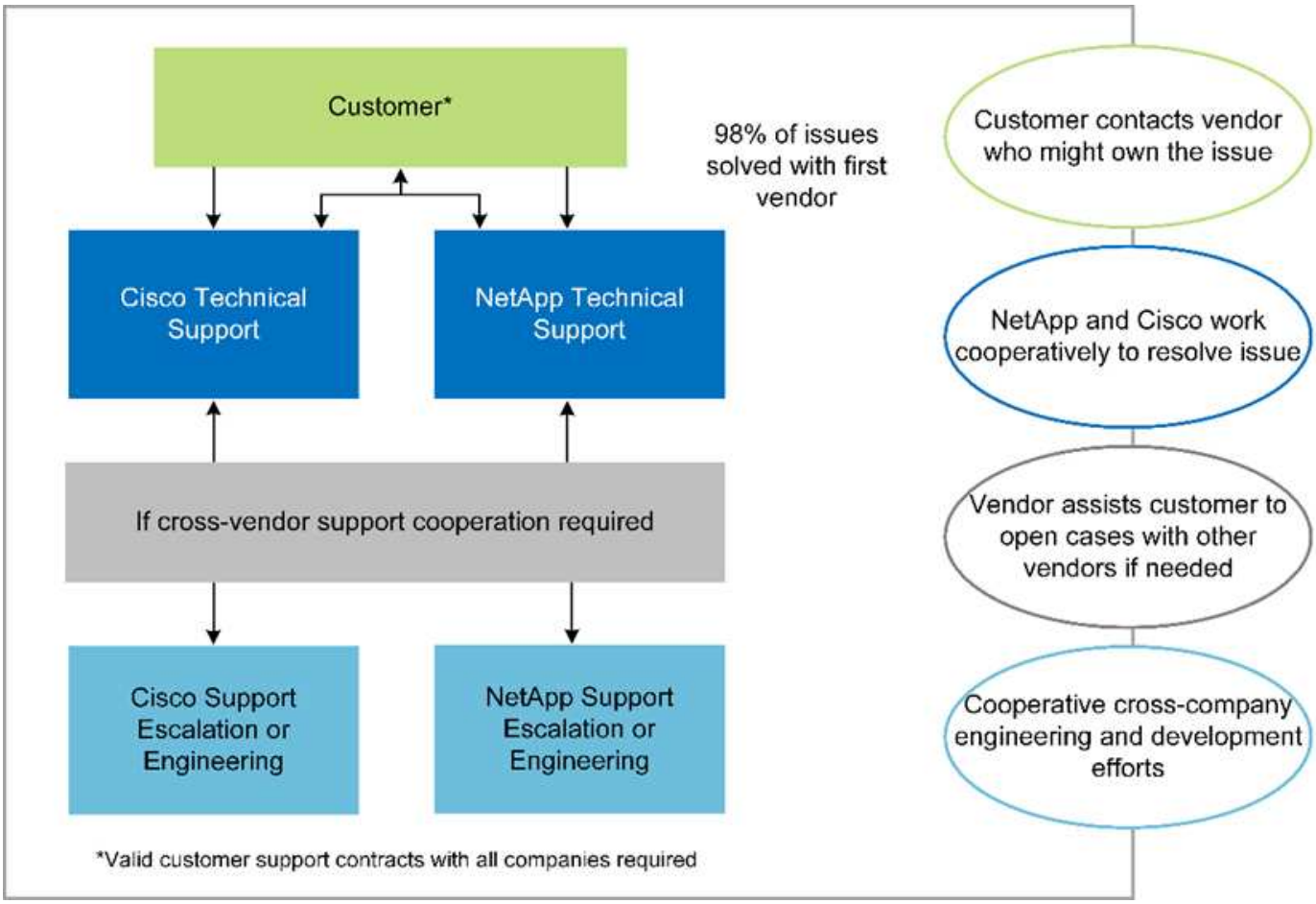
- **Analisi e annotazione delle varianti più rapide.** la piattaforma FlexPod è più veloce da implementare e da rendere operativa. La piattaforma FlexPod consente la collaborazione tra medici rendendo i dati disponibili su larga scala con bassa latenza e throughput aumentato. Una maggiore interoperabilità favorisce l'innovazione. Le organizzazioni del settore sanitario possono eseguire i workload genomici e non genomici in maniera affiancata, il che significa che le organizzazioni non hanno bisogno di piattaforme specializzate per iniziare il loro percorso di genomica.

  FlexPod ONTAP aggiunge regolarmente funzionalità all'avanguardia alla piattaforma di storage. FlexPod Datacenter è la base ottimale per l'implementazione di FC-NVMe per consentire l'accesso allo storage dalle performance elevate alle applicazioni che ne hanno bisogno. Poiché FC- NVMe si evolve per includere alta disponibilità, multipathing e supporto aggiuntivo del sistema operativo, FlexPod è la piattaforma scelta, fornendo la scalabilità e l'affidabilità necessarie per supportare queste funzionalità.

ONTAP con i/o più veloce con NVMe end-to-end consente di completare più rapidamente le analisi genomiche ("rif").

I dati del genoma raw in sequenza producono file di grandi dimensioni ed è importante che questi file siano resi disponibili agli analizzatori delle varianti per ridurre il tempo totale necessario dalla raccolta dei campioni all'annotazione delle varianti. NVMe (nonvolatile memory express), se utilizzato come protocollo di accesso allo storage e di trasporto dei dati, offre livelli di throughput senza precedenti e tempi di risposta più rapidi. FlexPod implementa il protocollo NVMe durante l'accesso allo storage flash tramite il bus PCI Express (PCIe). PCIe consente l'implementazione di decine di migliaia di code di comandi, aumentando la parallelizzazione e il throughput. Un singolo protocollo, dallo storage alla memoria, consente di accedere rapidamente ai dati.

- **L'agilità per la ricerca clinica da zero.** la capacità e le performance di storage flessibili ed espandibili consentono alle organizzazioni di ricerca nel settore sanitario di ottimizzare l'ambiente in modo elastico o JIT (Just-in-Time). Disaccoppiando lo storage dall'infrastruttura di calcolo e di rete, la piattaforma FlexPod può essere scalata verso l'alto e verso l'esterno senza interruzioni. Grazie a Cisco Intersight, la piattaforma FlexPod può essere gestita con flussi di lavoro automatizzati integrati e personalizzati. I flussi di lavoro di Cisco Intersight consentono alle organizzazioni sanitarie di ridurre i tempi di gestione del ciclo di vita delle applicazioni. Quando un centro medico accademico richiede che i dati dei pazienti siano anonimi e resi disponibili al proprio centro per la ricerca informatica e/o il centro per la qualità, l'organizzazione IT può sfruttare i flussi di lavoro di Cisco Intersight FlexPod per eseguire backup dei dati sicuri, clonare e ripristinare in pochi secondi, non ore. Con NetApp Trident e Kubernetes, le organizzazioni IT possono eseguire il provisioning di nuovi data scientist e rendere disponibili i dati clinici per lo sviluppo dei modelli in pochi minuti, talvolta anche in pochi secondi.
- **Protezione dei dati genoma.** NetApp SnapLock offre un volume speciale in cui i file possono essere memorizzati e impegnati in uno stato non cancellabile e non riscrivibile. I dati di produzione dell'utente che risiedono in un volume FlexVol possono essere mirrorati o archiviati in un volume SnapLock tramite la tecnologia NetApp SnapMirror o SnapVault. I file nel volume SnapLock, nel volume stesso e nel relativo aggregato di hosting non possono essere cancellati fino alla fine del periodo di conservazione. Utilizzando il software ONTAP FPolicy, le organizzazioni possono prevenire gli attacchi ransomware impedendo operazioni su file con estensioni specifiche. È possibile attivare un evento FPolicy per operazioni di file specifiche. L'evento è legato a una policy, che richiama il motore che deve utilizzare. È possibile configurare un criterio con una serie di estensioni di file che potrebbero contenere ransomware. Quando un file con un'estensione non consentita tenta di eseguire un'operazione non autorizzata, FPolicy impedisce l'esecuzione di tale operazione ("rif").
- **Supporto congiunto di FlexPod.** NetApp e Cisco hanno definito il supporto congiunto di FlexPod, un modello di supporto forte, scalabile e flessibile per soddisfare i requisiti di supporto esclusivi dell'infrastruttura convergente di FlexPod. Questo modello utilizza l'esperienza, le risorse e l'esperienza di supporto tecnico di NetApp e Cisco per offrire un processo ottimizzato per l'identificazione e la risoluzione dei problemi di supporto FlexPod, indipendentemente dalla posizione del problema. La figura seguente fornisce una panoramica del modello di supporto cooperativo FlexPod. Il cliente contatta il vendor che potrebbe essere responsabile del problema e Cisco e NetApp lavorano in collaborazione per risolverlo. Cisco e NetApp dispongono di team di sviluppo e progettazione multiazienda che lavorano insieme per risolvere i problemi. Questo modello di supporto riduce la perdita di informazioni durante la traduzione, garantisce fiducia e riduce i tempi di inattività.

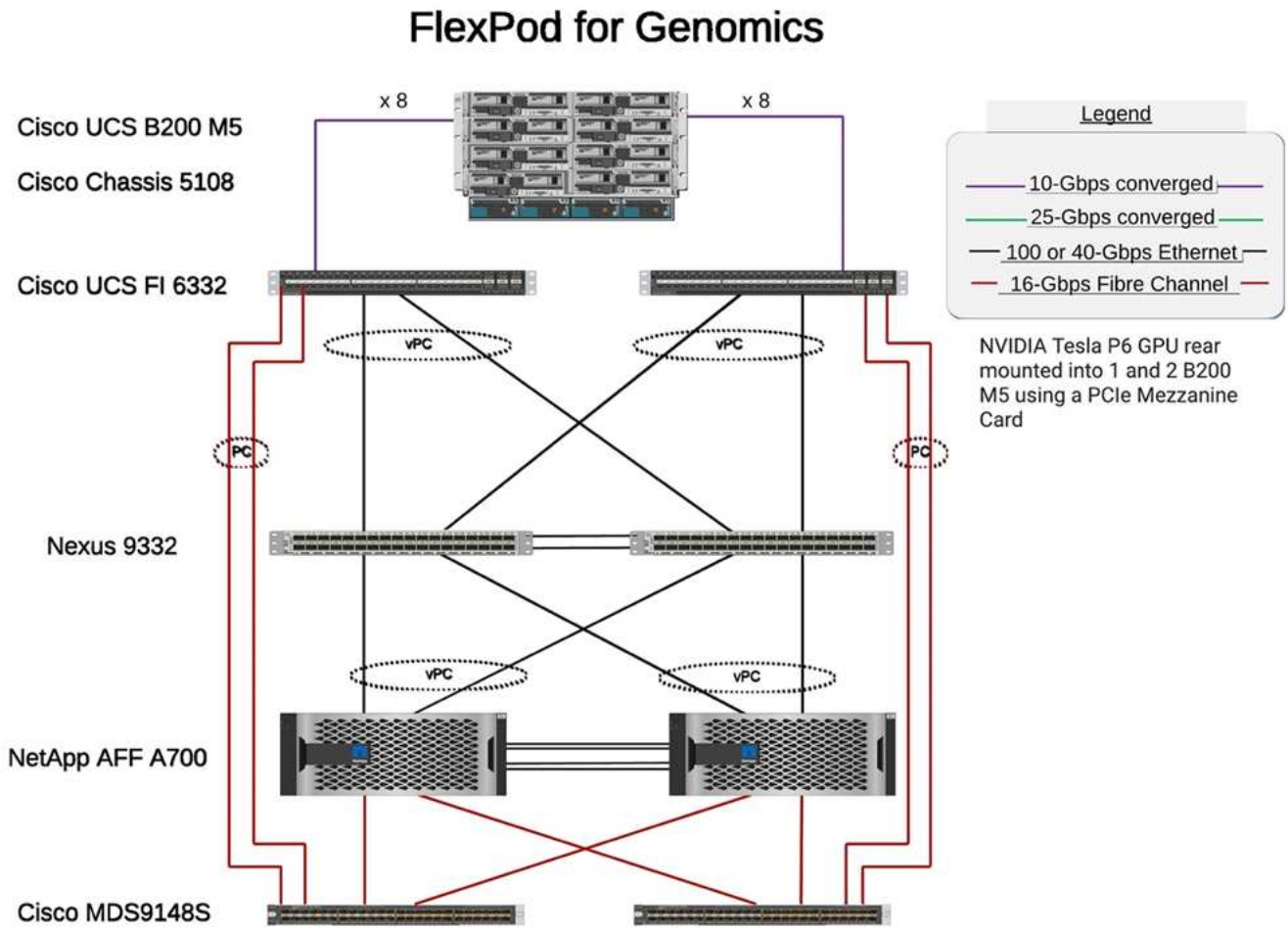


"Avanti: Componenti hardware e software dell'infrastruttura della soluzione."

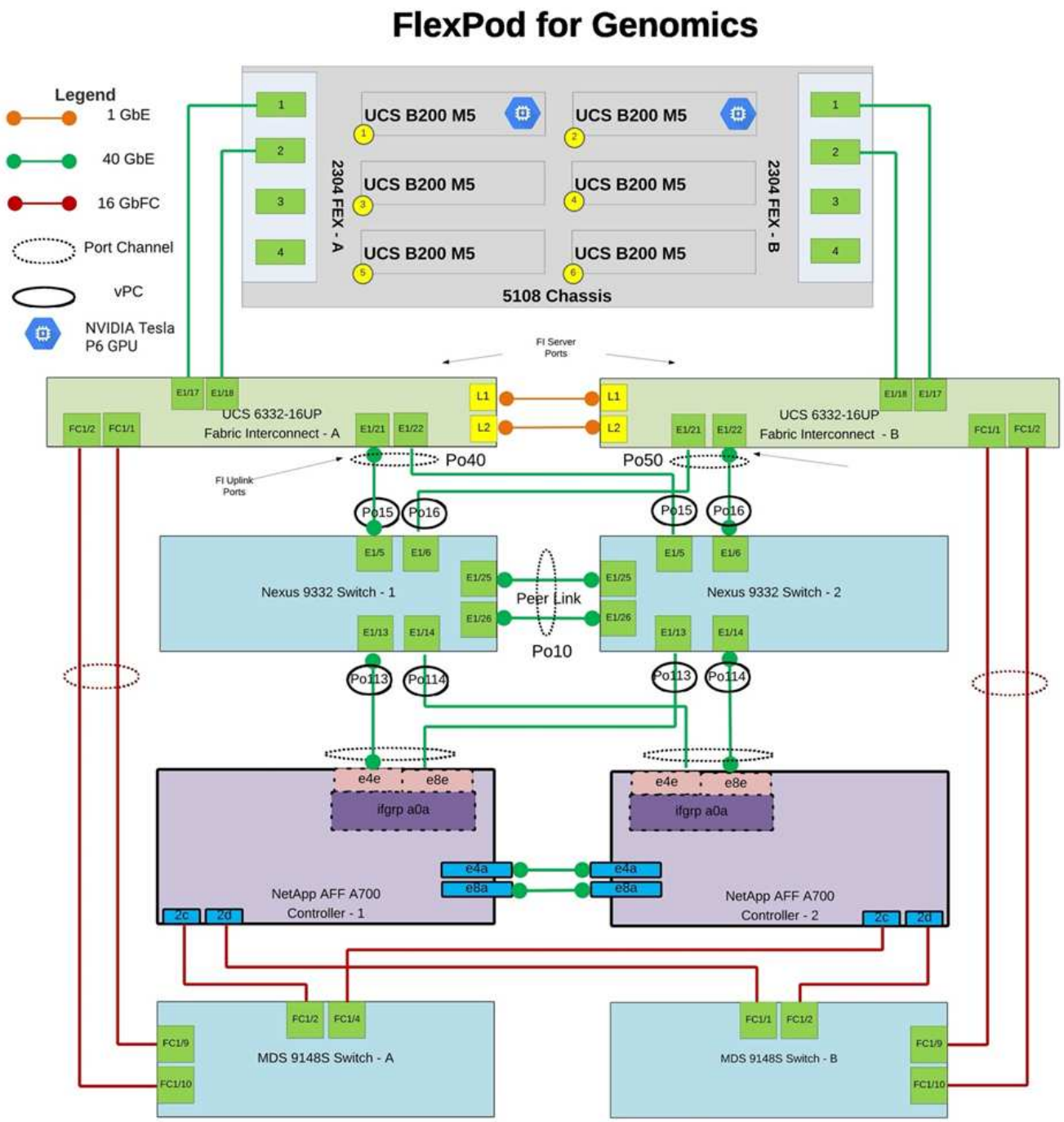
Componenti hardware e software dell'infrastruttura della soluzione

"Precedente: Vantaggi dell'implementazione di workload genomici su FlexPod."

La figura seguente mostra il sistema FlexPod utilizzato per la configurazione e la convalida di GATK. Abbiamo utilizzato "Data center FlexPod con VMware vSphere 7.0 e NetApp ONTAP 9.7 Cisco Validated Design (CVD)" durante il processo di configurazione.



Il seguente diagramma illustra i dettagli del cablaggio FlexPod.



La seguente tabella elenca i componenti hardware utilizzati durante il test GATK abilitando su un FlexPod. Ecco il "Tool di matrice di interoperabilità NetApp" (IMT) e. "Cisco hardware Compatibility List (HCL) (elenco compatibilità hardware Cisco)".

Layer	Famiglia di prodotti	Quantità e modello	Dettagli
Calcolo	Chassis Cisco UCS 5108	1 o 2	
	Blade server Cisco UCS	6 B200 M5	Ciascuno con 2 core da 20 o più, 2,7 GHz e 128 GB di RAM

Layer	Famiglia di prodotti	Quantità e modello	Dettagli
	Cisco UCS Virtual Interface Card (VIC)	Cisco UCS 1440	Vedere
	2 interconnessioni fabric Cisco UCS	6332	-
Rete	Switch Cisco Nexus	2 Cisco Nexus 9332	-
Rete di storage	Rete IP per l'accesso allo storage su protocolli SMB/CIFS, NFS o iSCSI	Stessi switch di rete come sopra	-
	Accesso allo storage tramite FC	2 Cisco MDS 9148S	-
Storage	Sistema storage all-flash NetApp AFF A700	1 cluster	Cluster con due nodi
	Shelf di dischi	Uno shelf di dischi DS224C o NS224	Completamente popolato con 24 dischi
	SSD	Capacità di 24, 1,2 TB o superiore	-

Questa tabella elenca il software dell'infrastruttura.

Software	Famiglia di prodotti	Versione o release	Dettagli
Vari	Linux	RHEL 8.3	-
	Windows	Windows Server 2012 R2 (64 bit)	-
	NetApp ONTAP	ONTAP 9.8 o versione successiva	-
	Cisco UCS Fabric Interconnect	Cisco UCS Manager 4.1 o versione successiva	-
	Switch Cisco Ethernet serie 3000 o 9000	Per la serie 9000, 7.0(3)I7(7) o versioni successive per la serie 3000, 9.2(4) o versioni successive	-
	Cisco FC: Cisco MDS 9132T	8.4(1a) o successiva	-
	Hypervisor	VMware vSphere ESXi 7.0	-
Storage	Sistema di gestione dell'hypervisor	VMware vCenter Server 7.0 (vCSA) o versione successiva	-
Rete	NetApp Virtual Storage Console (VSC)	VSC 9.7 o versione successiva	-

Software	Famiglia di prodotti	Versione o release	Dettagli
	NetApp SnapCenter	SnapCenter 4.3 o versione successiva	-
	Cisco UCS Manager	4.1(3c) o versione successiva	
Hypervisor	ESXi		
Gestione	Sistema di gestione dell'hypervisor VMware vCenter Server 7.0 (vCSA) o versione successiva		
	NetApp Virtual Storage Console (VSC)	VSC 9.7 o versione successiva	
	NetApp SnapCenter	SnapCenter 4.3 o versione successiva	
	Cisco UCS Manager	4.1(3c) o versione successiva	

"Next: Genomica - Installazione ed esecuzione di GATK."

Genomica - Installazione ed esecuzione di GATK

"Precedente: Componenti hardware e software dell'infrastruttura della soluzione."

Secondo il National Human Genome Research Institute ("NHGRI"), "la genomica è lo studio di tutti i geni di una persona (il genoma), comprese le interazioni di questi geni tra loro e con l'ambiente di una persona. "

In base a. "NHGRI"L'acido desossiribonucleico (DNA) è il composto chimico che contiene le istruzioni necessarie per sviluppare e dirigere le attività di quasi tutti gli organismi viventi. Le molecole di DNA sono costituite da due trefoli torcenti accoppiati, spesso indicati come a doppia elica". "Il set completo di DNA di un organismo è chiamato genoma".

Il sequenziamento è il processo di determinazione dell'ordine esatto delle basi in un filamento di DNA. Uno dei tipi più comuni di sequenziamento oggi utilizzato è chiamato sequenziamento per sintesi. Questa tecnica utilizza l'emissione di segnali fluorescenti per ordinare le basi. I ricercatori possono utilizzare il sequenziamento del DNA per cercare variazioni genetiche e qualsiasi mutazione che possa svolgere un ruolo nello sviluppo o nella progressione di una malattia mentre una persona è ancora nello stadio embrionale.

Dall'identificazione del campione alla variante, all'annotazione e alla previsione

Ad alto livello, la genomica può essere classificata nei seguenti passaggi. Questo non è un elenco completo:

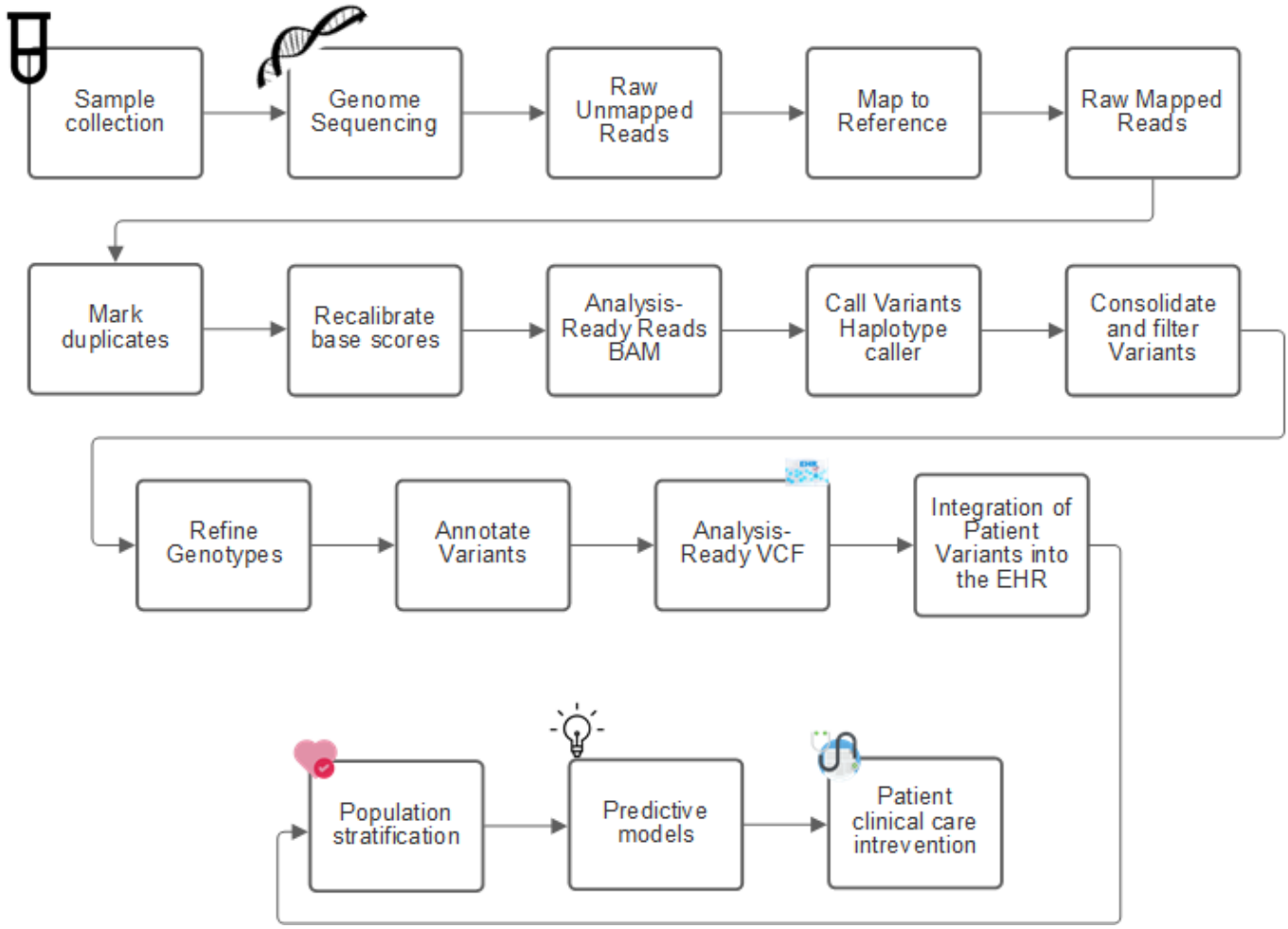
1. Raccolta dei campioni.
2. "Sequenziamento del genoma" utilizzo di un sequencer per generare i dati raw.
3. Pre-elaborazione. Ad esempio, "deduplica" utilizzo di "Picard".
4. Analisi genomica.

a. Mappatura a un genoma di riferimento.

b. "Variante" L'identificazione e l'annotazione vengono in genere eseguite utilizzando GATK e strumenti simili.

5. Integrazione nel sistema di cartelle cliniche elettroniche (EHR).
6. "Stratificazione della popolazione" e identificazione della variazione genetica attraverso la posizione geografica e il background etnico.
7. "Modelli predittivi" utilizzando un significativo polimorfismo a singolo nucleotide.
8. "Convalida".

La figura seguente mostra il processo che va dal campionamento all'identificazione della variante, all'annotazione e alla previsione.



Il progetto Human Genome è stato completato nell'aprile 2003 e il progetto ha realizzato una simulazione di altissima qualità della sequenza di genomi umani disponibile in pubblico dominio. Questo genoma di riferimento ha dato inizio a un'esplosione nella ricerca e nello sviluppo delle capacità genomiche. Praticamente ogni disturbo umano ha una firma nei geni di quell'essere umano. Fino a poco tempo fa, i medici utilizzavano i geni per predire e determinare i difetti congeniti come l'anemia falciforme, causata da un certo schema di ereditarietà causato da un cambiamento in un singolo gene. Il tesoro dei dati messi a disposizione dal progetto sul genoma umano ha portato all'avvento dello stato attuale delle capacità genomiche.

La genomica offre una vasta gamma di vantaggi. Ecco una piccola serie di vantaggi nei settori sanitario e delle scienze biologiche:

- Migliore diagnosi presso i punti di cura
- Migliore prognosi
- Medicina di precisione
- Piani di trattamento personalizzati
- Migliore monitoraggio delle malattie
- Riduzione degli eventi avversi
- Migliore accesso alle terapie
- Miglioramento del monitoraggio delle malattie
- Partecipazione efficace agli studi clinici e migliore selezione dei pazienti per gli studi clinici basati sui genotipi.

La genomica è un "bestia a quattro teste," a causa delle esigenze di calcolo per tutto il ciclo di vita di un set di dati: acquisizione, storage, distribuzione e analisi.

GATK (Genome Analysis Toolkit)

GATK è stata sviluppata come piattaforma per la data science presso "Broad Institute". GATK è un insieme di strumenti open-source che consentono l'analisi del genoma, in particolare rilevamento, identificazione, annotazione e genotipizzazione delle varianti. Uno dei vantaggi di GATK è che il set di strumenti e/o comandi può essere concatenato per formare un workflow completo. Le principali sfide affrontate da un ampio istituto sono le seguenti:

- Comprendere le cause alla radice e i meccanismi biologici delle malattie.
- Identificare gli interventi terapeutici che agiscono alla causa fondamentale di una malattia.
- Comprendere la linea di vista dalle varianti al funzionamento in fisiologia umana.
- Creare standard e policy "framework" per la rappresentazione dei dati genoma, lo storage, l'analisi, la sicurezza e così via.
- Standardizzare e socializzare database di aggregazione dei genomi interoperabili (gnomAD).
- Monitoraggio, diagnosi e trattamento dei pazienti basati sul genoma con maggiore precisione.
- Aiuta a implementare strumenti che prevedano le malattie ben prima che appaiano i sintomi.
- Crea e potenzia una community di collaboratori interdisciplinari per affrontare i problemi più difficili e importanti della biomedicina.

Secondo il GATK e l'ampio istituto, il sequenziamento del genoma deve essere trattato come un protocollo in un laboratorio di patologia; ogni attività è ben documentata, ottimizzata, riproducibile e coerente tra campioni ed esperimenti. Di seguito viene riportata una serie di procedure consigliate dal Broad Institute. Per ulteriori informazioni, vedere "Sito web di GATK".

Configurazione di FlexPod

La convalida del carico di lavoro di genomics include una configurazione da zero di una piattaforma di infrastruttura FlexPod. La piattaforma FlexPod è altamente disponibile e può essere scalata in modo indipendente; ad esempio, la rete, lo storage e il calcolo possono essere scalati in modo indipendente. Abbiamo utilizzato la seguente guida alla progettazione convalidata da Cisco come documento di riferimento sull'architettura per configurare l'ambiente FlexPod: "Data center FlexPod con VMware vSphere 7.0 e NetApp ONTAP 9.7". Scopri i seguenti punti salienti della configurazione della piattaforma FlexPod:

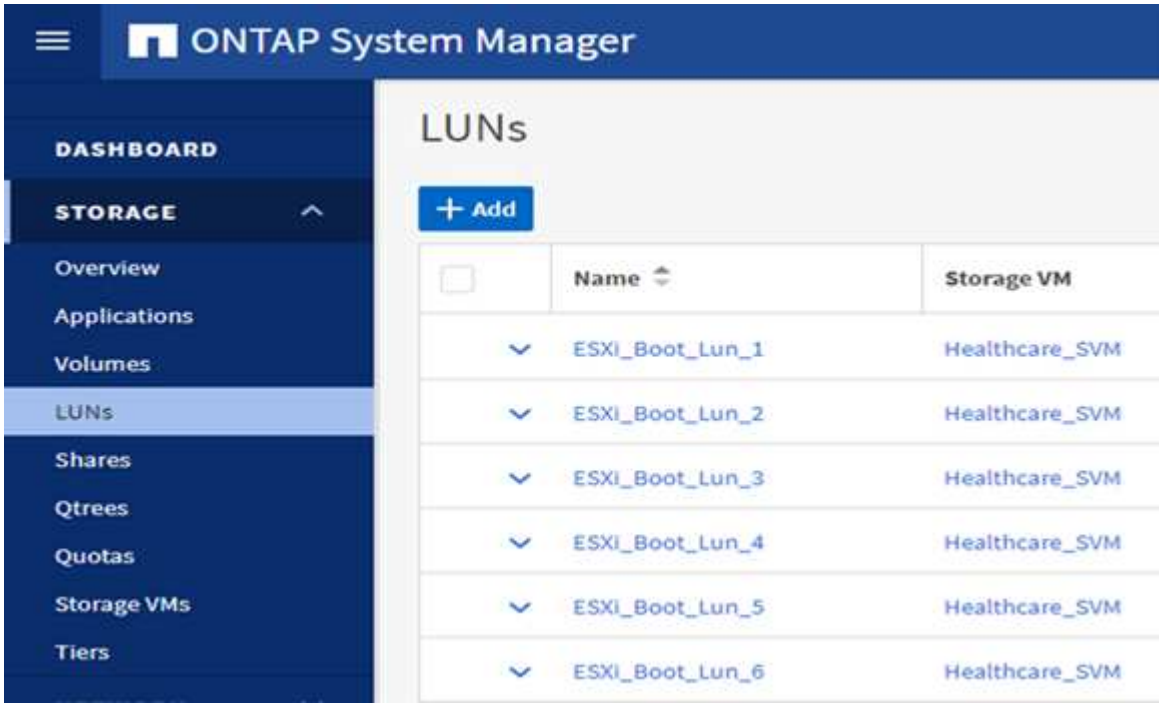
Per eseguire la configurazione del laboratorio FlexPod, attenersi alla seguente procedura:

1. La configurazione e la convalida del laboratorio FlexPod utilizza le seguenti prenotazioni IP4 e VLAN.

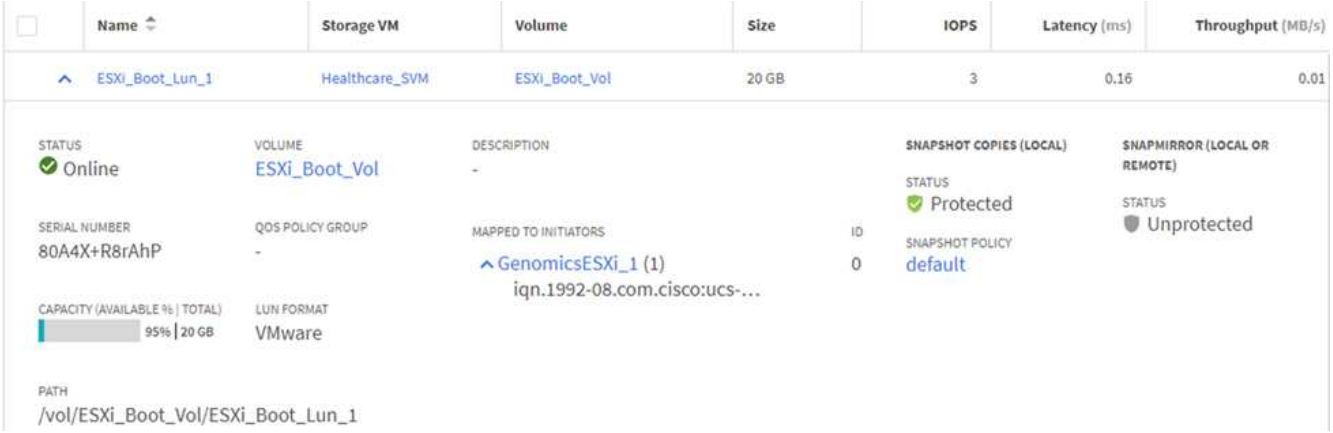
IP Reservations

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. Configurare le LUN di avvio basate su iSCSI sulla SVM ONTAP.

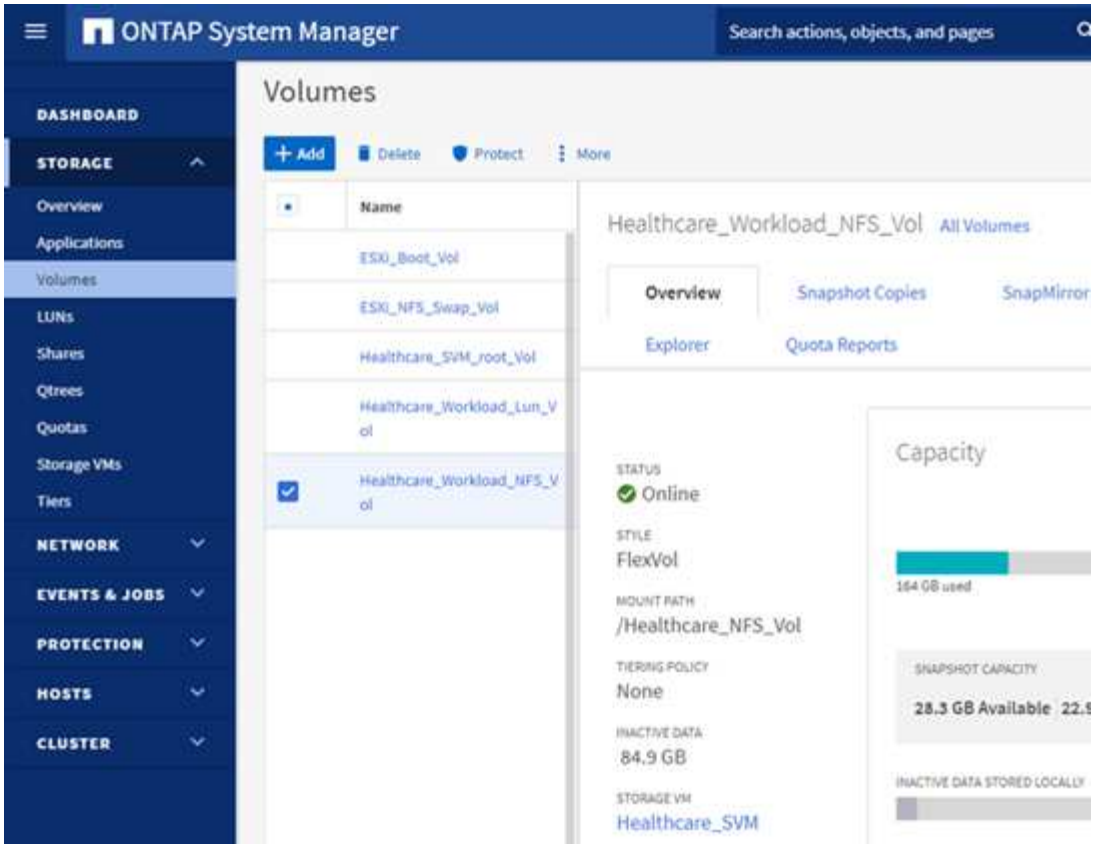


3. Mappare i LUN ai gruppi di iniziatori iSCSI.

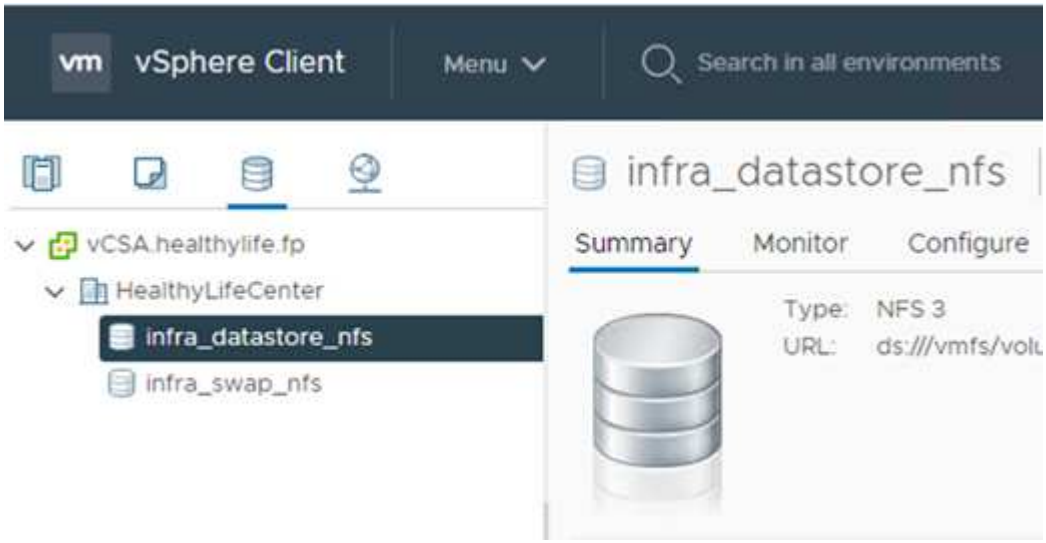


4. Installare vSphere 7.0 con l'avvio iSCSI.
5. Registrare gli host ESXi con vCenter.

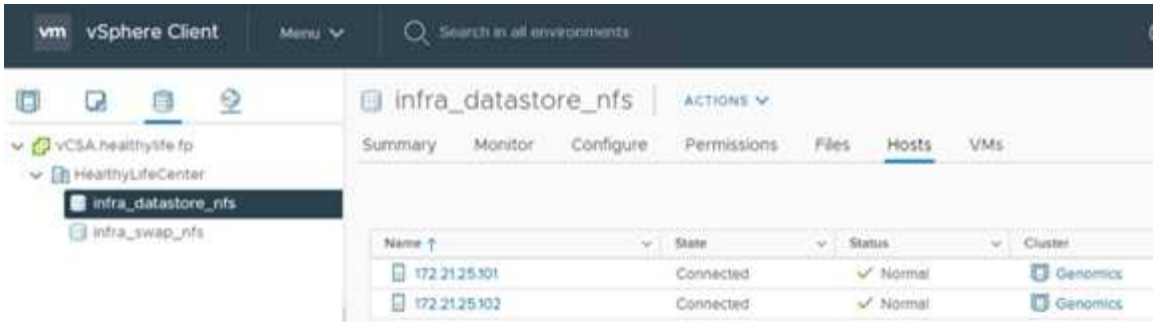
6. Eseguire il provisioning di un datastore NFS `infra_datastore_nfs` Sullo storage ONTAP.



7. Aggiungere il datastore al vCenter.



8. Utilizzando vCenter, aggiungere un datastore NFS agli host ESXi.



9. Utilizzando vCenter, creare una macchina virtuale Red Hat Enterprise Linux (RHEL) 8.3 per eseguire GATK.
10. Un datastore NFS viene presentato alla macchina virtuale e montato su `/mnt/genomics`, Utilizzato per memorizzare file eseguibili GATK, script, file BAM (Binary Alignment Map), file di riferimento, file di indice, file del dizionario e file out per la chiamata delle varianti.

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

Configurazione ed esecuzione di GATK

Installare i seguenti prerequisiti su RedHat Enterprise 8.3 Linux VM:

- Java 8 o SDK 1.8 o versione successiva
- Scarica GATK 4.2.0.0 dal Broad Institute "Sito GitHub". I dati della sequenza genoma sono generalmente memorizzati sotto forma di una serie di colonne ASCII delimitate da tabulazioni. Tuttavia, ASCII occupa troppo spazio per la memorizzazione. Pertanto, un nuovo standard evoluto chiamato file BAM (**.bam). Un file BAM memorizza i dati della sequenza in un formato compresso, indicizzato e binario. Noi "scaricato" Un insieme di file BAM disponibili pubblicamente per l'esecuzione di GATK da "di dominio pubblico". Abbiamo anche scaricato file di indice (**.bai), file di dizionario (**. dict) e file di dati di riferimento (**. fasta) dello stesso dominio pubblico.

Dopo il download, il kit di strumenti GATK ha un file jar e una serie di script di supporto.

- `gatk-package-4.2.0.0-local.jar` eseguibile
- `gatk` file di script.

Abbiamo scaricato i file BAM e i corrispondenti file di indice, dizionario e genoma di riferimento per una famiglia composta da file *.bam padre, madre e figlio.

Motore Cromwell

Cromwell è un motore open-source orientato ai flussi di lavoro scientifici che consente la gestione del workflow. Il motore Cromwell può essere eseguito in due "modalità", Server mode o Run mode a singolo flusso di lavoro. Il comportamento del motore Cromwell può essere controllato tramite "File di configurazione del motore Cromwell".

- **Server mode.** attiva "Riposante" Esecuzione dei flussi di lavoro nel motore Cromwell.
- **Run mode.** la modalità Run è più adatta per l'esecuzione di singoli flussi di lavoro in Cromwell, "rif" Per una serie completa di opzioni disponibili in modalità Run.

Utilizziamo il motore Cromwell per eseguire flussi di lavoro e pipeline su larga scala. Il motore Cromwell utilizza un sistema intuitivo "linguaggio di descrizione del workflow" Linguaggio di scripting basato su (WDL). Cromwell supporta anche un secondo standard di scripting per il workflow, denominato Common workflow Language (CWL). Nel corso di questo report tecnico, abbiamo utilizzato WDL. WDL è stato originariamente sviluppato dal Broad Institute for Genome analysis Pipeline. I flussi di lavoro WDL possono essere implementati utilizzando diverse strategie, tra cui:

- **Linear Chaining.** come suggerisce il nome, l'output dell'attività n. 1 viene inviato all'attività n. 2 come input.
- **Multi-in/out.** questo è simile al concatenamento lineare in quanto ogni task può avere più output inviati come input a task successivi.
- **Scatter-Gather.** si tratta di una delle strategie di integrazione applicativa aziendale (EAI) più potenti disponibili, soprattutto se utilizzata in un'architettura basata sugli eventi. Ogni task viene eseguito in modo disaccoppiato e l'output di ogni task viene consolidato nell'output finale.

Quando si utilizza WDL per eseguire GATK in una modalità standalone, sono disponibili tre passaggi:

1. Validare la sintassi utilizzando `womtool.jar`.

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

2. Generare input JSON.

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

3. Eseguire il flusso di lavoro utilizzando il motore Cromwell e. `Cromwell.jar`.

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

Il GATK può essere eseguito utilizzando diversi metodi; questo documento esplora tre di questi metodi.

Esecuzione di GATK utilizzando il file jar

Esaminiamo ora l'esecuzione di una singola pipeline di chiamate con il chiamante della variante haplotype.

```
[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
```

In questo metodo di esecuzione, utilizziamo il file jar di esecuzione locale di GATK, utilizziamo un singolo comando java per richiamare il file jar e passiamo diversi parametri al comando.

1. Questo parametro indica che stiamo richiamando `HaplotypeCaller` pipeline chiamante variante.
2. `-- input` Specifica il file BAM di input.
3. `--output` specifica il file di output della variante nel formato di chiamata della variante (*.vcf) ("rif").
4. Con `--reference` parametro, stiamo passando un genoma di riferimento.

Una volta eseguita l'operazione, i dettagli dell'output sono disponibili nella sezione "Output per l'esecuzione di GATK utilizzando il file jar."

Esecuzione di GATK utilizzando lo script ./gatk

Il kit di strumenti GATK può essere eseguito utilizzando `./gatk` script. Esaminiamo il seguente comando:

```
[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
```

Passiamo diversi parametri al comando.

- Questo parametro indica che stiamo richiamando `HaplotypeCaller` pipeline chiamante variante.
- `-I` Specifica il file BAM di input.
- `-O` specifica il file di output della variante nel formato di chiamata della variante (*.vcf) ("rif").
- Con `-R` parametro, stiamo passando un genoma di riferimento.

Una volta eseguita l'operazione, i dettagli dell'output sono disponibili nella sezione

Esecuzione di GATK utilizzando il motore Cromwell

Utilizziamo il motore Cromwell per gestire l'esecuzione di GATK. Esaminiamo la riga di comando e i relativi parametri.

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

In questo caso, viene richiamato il comando Java passando a. `-jar` parametro per indicare che si intende eseguire un file jar, ad esempio `Cromwell-65.jar`. Il parametro successivo è stato superato (`run`) Indica che il motore Cromwell è in esecuzione in modalità Run, mentre l'altra opzione possibile è la modalità Server. Il parametro successivo è `*.wdl` Che la modalità Run debba utilizzare per eseguire le pipeline. Il parametro successivo è l'insieme di parametri di input per i flussi di lavoro in esecuzione.

Di seguito sono elencati i contenuti di `ghplo.wdl` file simile a:

```
[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]#
```

Ecco il file JSON corrispondente con gli input al motore Cromwell.

```
[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]#
```

Tenere presente che Cromwell utilizza un database in-memory per l'esecuzione. Una volta eseguito, il log di output viene visualizzato nella sezione "Output per l'esecuzione di GATK utilizzando il motore Cromwell."

Per una serie completa di passaggi su come eseguire GATK, vedere "Documentazione GATK".

"Successivo: Output per l'esecuzione di GATK utilizzando il file jar."

Output per l'esecuzione di GATK utilizzando il file jar

"Precedente: Genomica - impostazione ed esecuzione di GATK."

L'esecuzione di GATK utilizzando il file jar ha prodotto il seguente output di esempio.

```
[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]#
```

Si noti che il file di output si trova nella posizione specificata dopo l’esecuzione.

Output per l'esecuzione di GATK utilizzando lo script ./gatk

"Precedente: Output per l'esecuzione di GATK utilizzando il file jar."

L'esecuzione di GATK utilizzando `./gatk` lo script ha prodotto il seguente output di esempio.

```
[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.healthylife.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
: 0.012129203000000002
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]#
```

Si noti che il file di output si trova nella posizione specificata dopo l'esecuzione.

"Avanti: Output per l'esecuzione di GATK utilizzando il motore Cromwell."

Output per l'esecuzione di GATK utilizzando il motore Cromwell

L'esecuzione di GATK utilizzando il motore Cromwell ha prodotto il seguente output di esempio.

```
[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]#
```

"Avanti: Configurazione della GPU."

Configurazione della GPU

"Precedente: Output per l'esecuzione di GATK utilizzando il motore Cromwell."

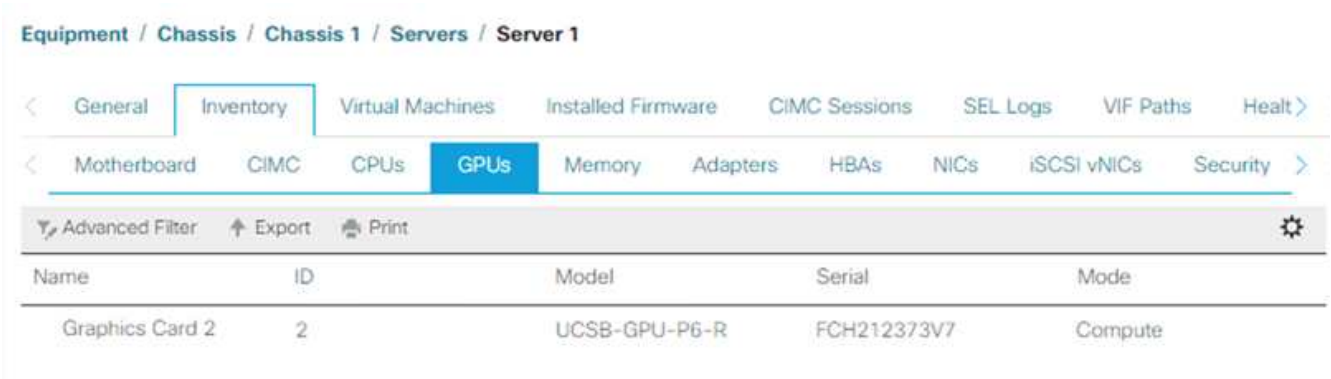
Al momento della pubblicazione, il tool GATK non supporta in modo nativo l'esecuzione on-premise basata su GPU. La seguente configurazione e guida consentono ai lettori di comprendere quanto sia semplice utilizzare FlexPod con una GPU NVIDIA Tesla P6 montata sul retro utilizzando una scheda mezzanine PCIe per GATK.

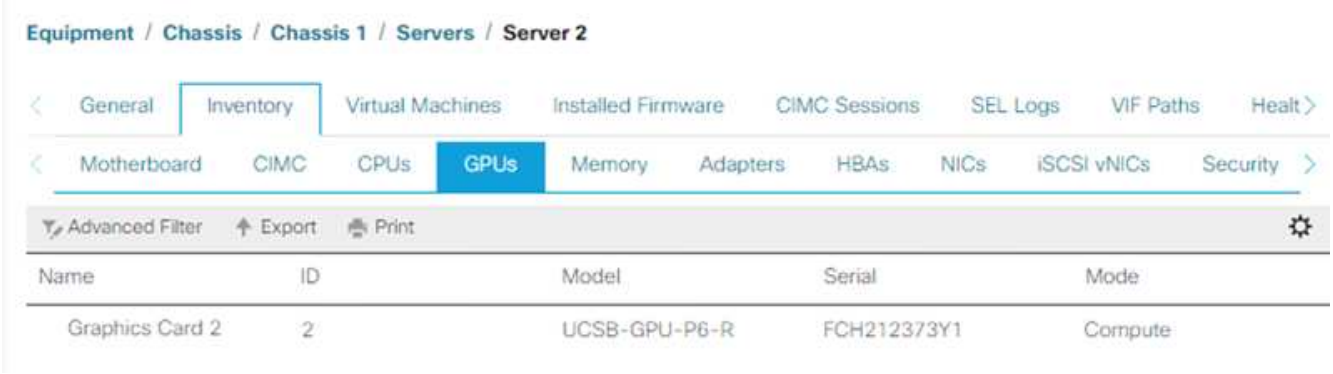
Abbiamo utilizzato il seguente progetto validato da Cisco (CVD) come architettura di riferimento e guida alle Best practice per configurare l'ambiente FlexPod in modo da poter eseguire applicazioni che utilizzano GPU.

- "Data center FlexPod per ai/ML con Cisco UCS 480 ML per l'apprendimento approfondito"

Ecco una serie di punti chiave durante questa configurazione:

1. Abbiamo utilizzato una GPU NVIDIA Tesla P6 PCIe in uno slot mezzanino nei server UCS B200 M5.





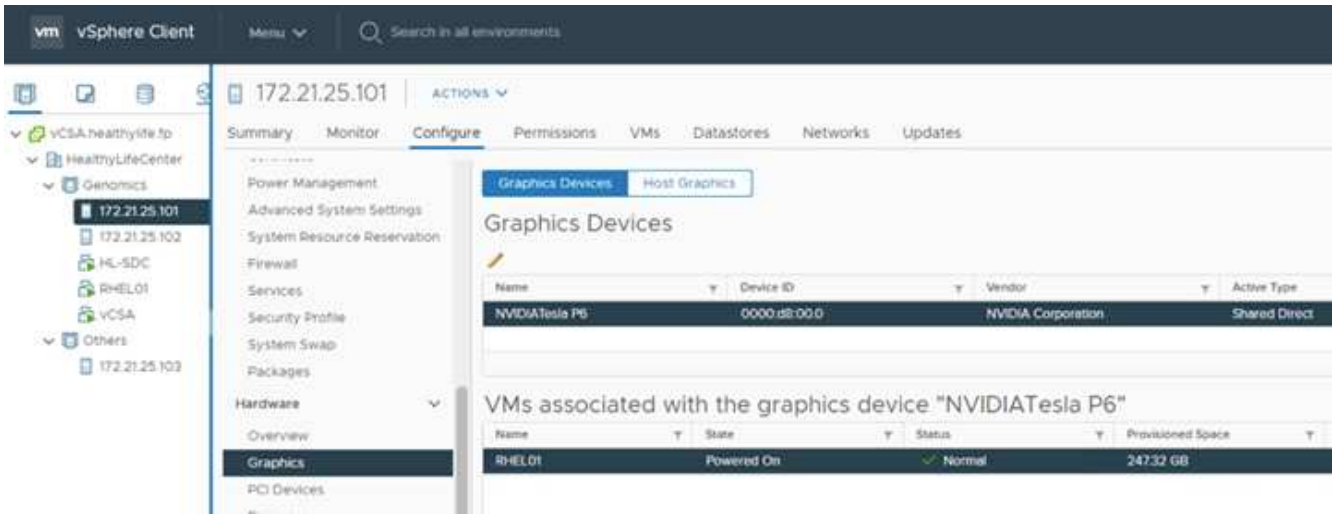
2. Per questa configurazione, ci siamo registrati sul portale partner NVIDIA e abbiamo ottenuto una licenza di valutazione (nota anche come diritto) per poter utilizzare le GPU in modalità di calcolo.
3. Il software NVIDIA vGPU richiesto è stato scaricato dal sito Web del partner NVIDIA.
4. Abbiamo scaricato i diritti `*.bin` Dal sito web del partner NVIDIA.
5. Abbiamo installato un server di licenza NVIDIA vGPU e aggiunto le autorizzazioni al server di licenza utilizzando `*.bin` File scaricato dal sito del partner NVIDIA.
6. Assicurarsi di scegliere la versione software NVIDIA vGPU corretta per l'implementazione sul portale dei partner NVIDIA. Per questa configurazione è stata utilizzata la versione del driver 460.73.02.
7. Questo comando installa "NVIDIA vGPU Manager" In ESXi.

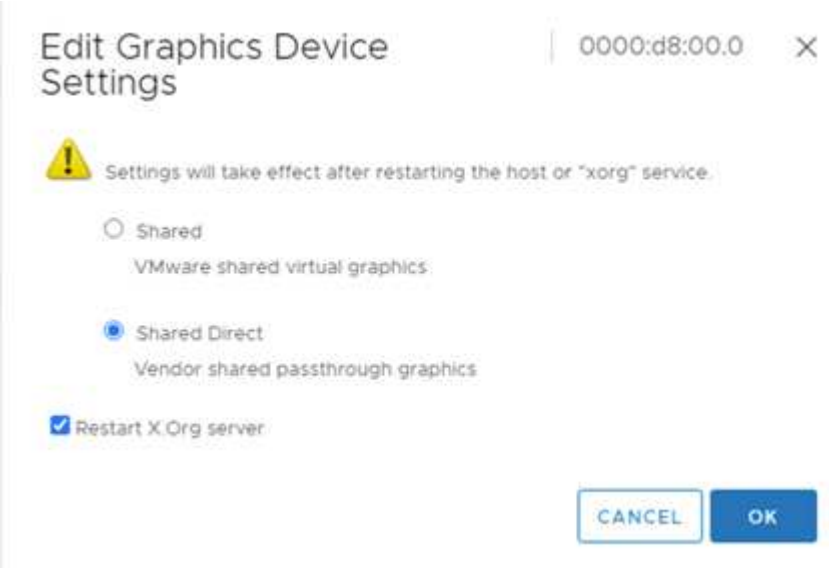
```
[root@localhost:~] esxcli software vib install -v
/vmfs/volumes/infra_datastore_nfs/nvidia/vib/NVIDIA_bootbank_NVIDIA-
VMware_ESXi_7.0_Host_Driver_460.73.02-1OEM.700.0.0.15525992.vib
Installation Result
Message: Operation finished successfully.
Reboot Required: false
VIBs Installed: NVIDIA_bootbank_NVIDIA-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

8. Dopo aver riavviato il server ESXi, eseguire il seguente comando per convalidare l'installazione e controllare lo stato delle GPU.

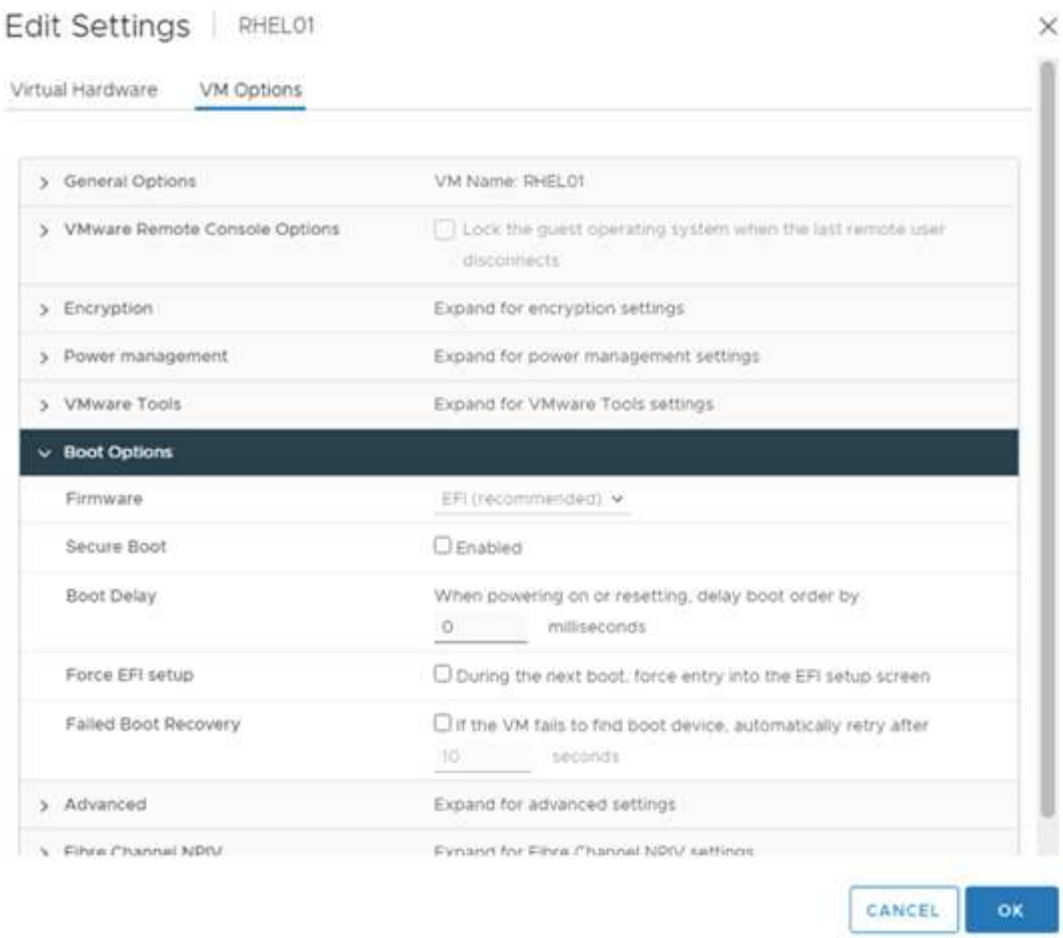
```
[root@localhost:~] nvidia-smi
Wed Aug 18 21:37:19 2021
+-----------------------------------------------------------------------------
------+
| NVIDIA-SMI 460.73.02    Driver Version: 460.73.02    CUDA Version: N/A
|
|-------------------------------+----------------------
+----------------------+
| 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. Utilizzando vCenter, "configurare" L'impostazione del dispositivo grafico è "Shared Direct".





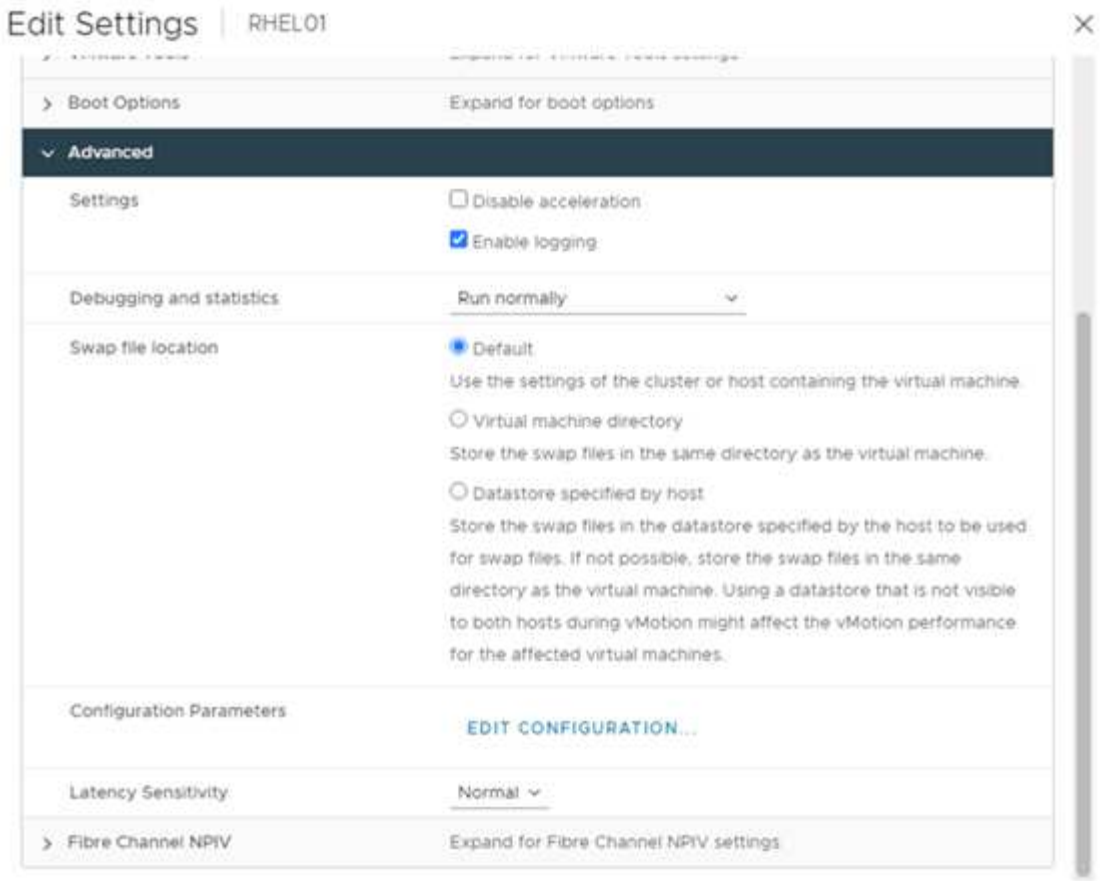
10. Assicurarsi che l'avvio sicuro sia disattivato per la macchina virtuale RedHat.
11. Assicurarsi che il firmware VM Boot Options sia impostato su EFI ("rif").

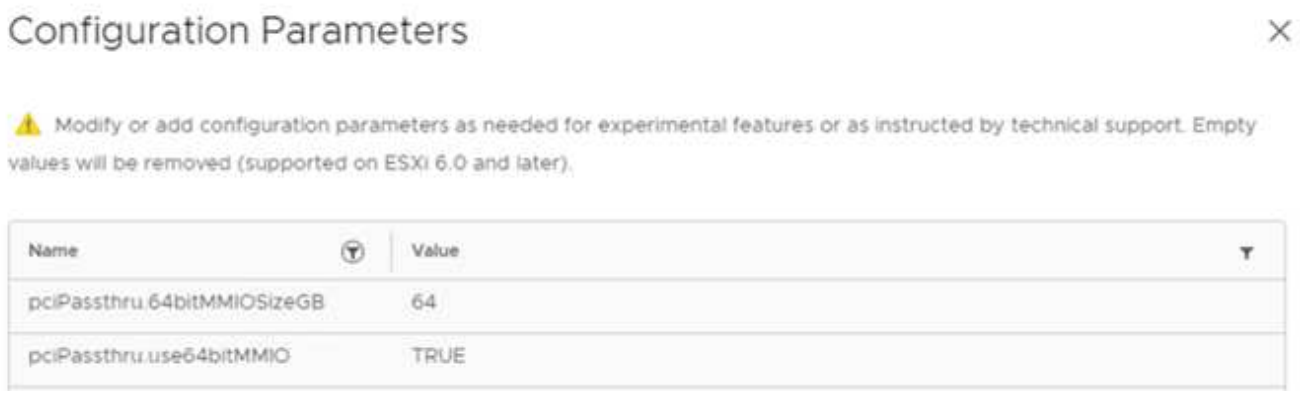


12. Assicurarsi che i SEGUENTI PARAMETRI siano aggiunti alla configurazione avanzata di modifica delle opzioni della macchina virtuale. Il valore di `pciPassthru.64bitMMIOSizeGB` Il parametro dipende dalla memoria della GPU e dal numero di GPU assegnate alla VM. Ad esempio:

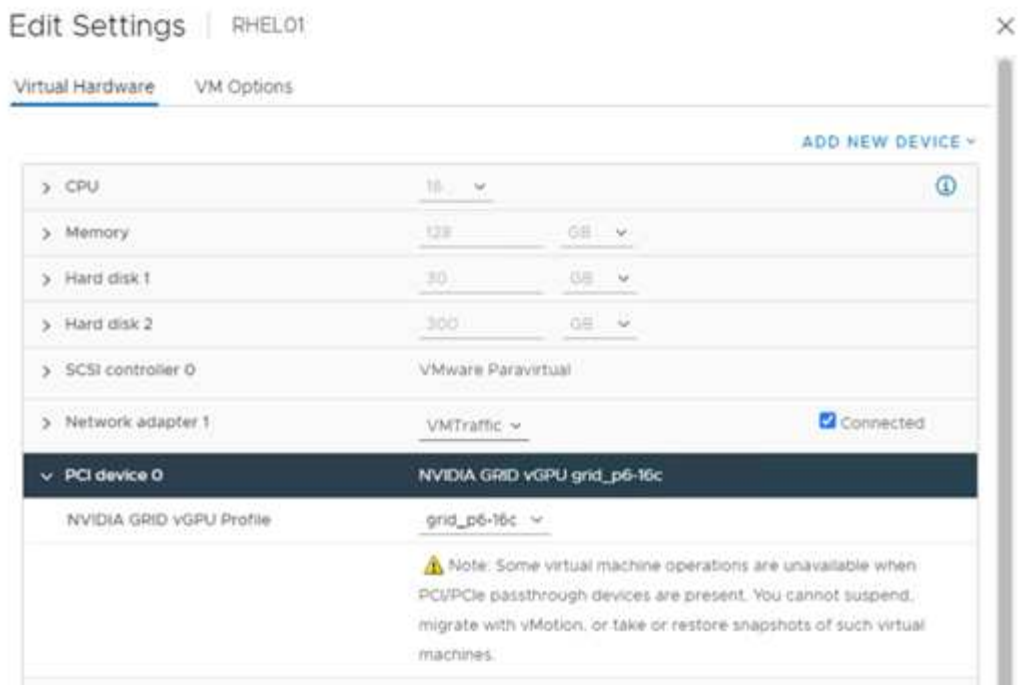
    a. Se a una macchina virtuale sono assegnate 4 GPU V100 da 32 GB, questo valore deve essere 128.

    b. Se a una macchina virtuale sono assegnate 4 GPU P6 da 16 GB, questo valore deve essere 64.





13. Quando si aggiungono vGPU come nuovo dispositivo PCI alla macchina virtuale in vCenter, assicurarsi di selezionare NVIDIA GRID vGPU come tipo di dispositivo PCI.
14. Scegliere il profilo GPU corretto che si adatta alla GPU utilizzata, alla memoria GPU e allo scopo di utilizzo: Ad esempio, grafica o calcolo.



15. Su RedHat Linux VM, i driver NVIDIA possono essere installati eseguendo il seguente comando:

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

16. Verificare che venga segnalato il profilo vGPU corretto eseguendo il seguente comando:

```
[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. Dopo il riavvio, verificare che la scheda NVIDIA vGPU corretta sia riportata insieme alle versioni dei driver.

```
[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. Assicurarsi che l'IP del server di licenza sia configurato sulla macchina virtuale nel file di configurazione della griglia vGPU.
    a. Copiare il modello.

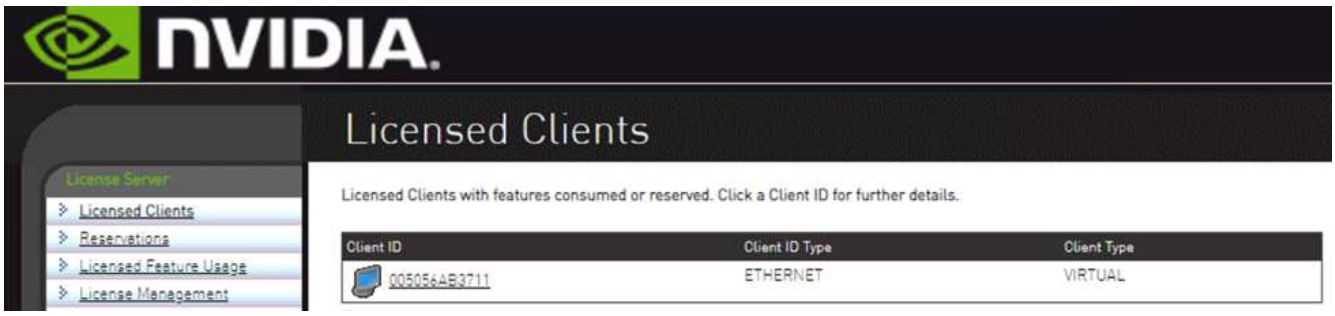
```
[root@genomics1 genomics]# cp /etc/nvidia/gridd.conf.template
/etc/nvidia/gridd.conf
```

b. Modificare il file `/etc/nvidia/rid.conf`, Aggiungere l'indirizzo IP del server di licenza e impostare il tipo di funzione su 1.

```
ServerAddress=192.168.169.10
```

```
FeatureType=1
```

19. Dopo aver riavviato la macchina virtuale, nel server di licenza viene visualizzata una voce sotto Licensed Clients (Client concessi in licenza), come mostrato di seguito.



20. Per ulteriori informazioni sul download del software GATK e Cromwell, consultare la sezione Solutions Setup.

21. Dopo che GATK può utilizzare le GPU on-premise, il linguaggio di descrizione del workflow `*. wdl` ha gli attributi di runtime come mostrato di seguito.

```
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 addtional_disk_space_gb = 50
  }
  Int disk_size = ceil(size(input_bam, "GB")) + addtional_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}"
  }
}
```

"Prossimo: Conclusione."

Conclusione

"Precedente: Configurazione della GPU."

Molte organizzazioni sanitarie di tutto il mondo hanno standardizzato FlexPod come piattaforma comune. Con FlexPod, puoi implementare le funzionalità del settore sanitario in tutta sicurezza. FlexPod con NetApp ONTAP è dotato di serie della capacità di implementare un set di protocolli leader del settore pronto all'uso. Indipendentemente dall'origine della richiesta di eseguire genomica di un dato paziente, interoperabilità, accessibilità, disponibilità e scalabilità sono standard con una piattaforma FlexPod. Se

standardizzato su una piattaforma FlexPod, la cultura dell'innovazione diventa contagiosa.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e siti Web:

- Data center FlexPod per ai/ML con Cisco UCS 480 ML per l'apprendimento approfondito

  "https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_480ml_aiml_deployment.pdf"

- Data center FlexPod con VMware vSphere 7.0 e 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"

- Centro documentazione di ONTAP 9

  "http://docs.netapp.com"

- Agile ed efficiente: Come FlexPod promuove la modernizzazione del data center

  "https://www.flexpod.com/idc-white-paper/"

- Ai nel settore sanitario

  "https://www.netapp.com/us/media/na-369.pdf"

- FlexPod per il settore sanitario semplifica la tua trasformazione

  "https://flexpod.com/solutions/verticals/healthcare/"

- FlexPod di Cisco e NetApp

  "https://flexpod.com/"

- Ai e Analytics per il settore sanitario (NetApp)

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

- Ai nel settore sanitario le scelte di infrastruttura intelligente aumentano il successo

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

- Data center FlexPod con ONTAP 9.8, connettore storage ONTAP per Cisco Intersight e modalità gestita Cisco Intersight.

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

- Data center FlexPod con piattaforma OpenStack Linux aziendale Red Hat

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

Cronologia delle versioni

Versione	Data	Cronologia delle versioni del documento
Versione 1.0	Novembre 2021	Release iniziale.