Lustre avec le système de stockage NetApp E-Series

NetApp artificial intelligence solutions

NetApp
August 31, 2026

Sommaire

Lustre avec le système de stockage NetApp E-Series

Lustre avec le stockage NetApp E-Series - Introduction

Lustre avec le stockage NetApp E-Series est une solution de système de fichiers parallèle validée pour les charges de travail d'IA et de HPC, construite à partir d'éléments de base répétables de paires de serveurs HA et de baies tout flash E-Series.

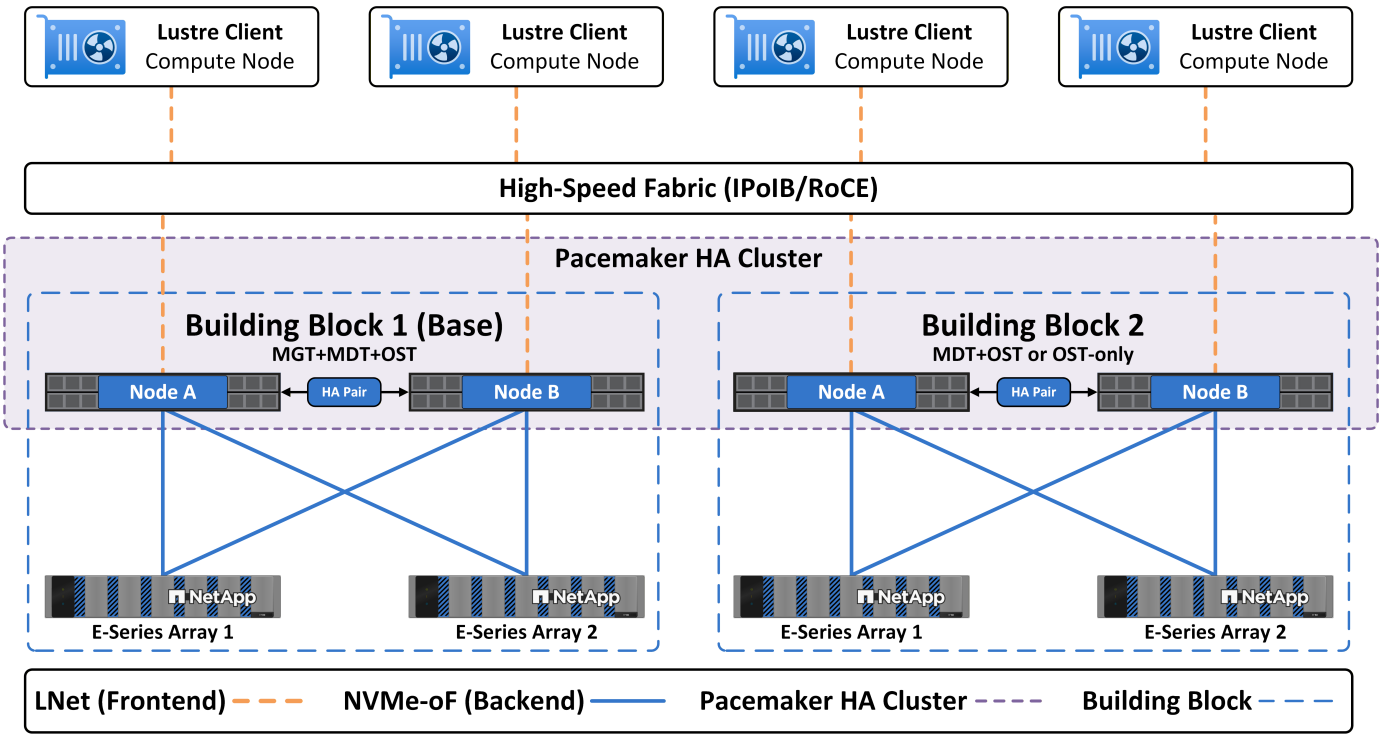
Présentation de la solution

Lustre avec le stockage NetApp E-Series combine le système de fichiers parallèle open source Lustre avec le stockage bloc NetApp E-Series pour offrir une infrastructure évolutive et haute performance adaptée aux charges de travail intensives en données. Chaque élément de base associe deux nœuds de serveur OSS/MDS configurés pour la haute disponibilité (HA) à deux baies E-Series 100 % flash. La connectivité NVMe-oF du backend transporte les E/S bloc vers les baies et un réseau LNet distinct connecte les clients et les nœuds de serveur OSS/MDS pour l'accès parallèle au système de fichiers.

Un élément de base héberge le serveur de gestion (MGS), les cibles de métadonnées initiales (MDT) et les cibles de stockage d'objets (OST). Des éléments de base supplémentaires s'enregistrent auprès du MGS de base pour augmenter la capacité de métadonnées, la capacité de données et le débit agrégé à mesure que les besoins de la charge de travail augmentent.

La figure suivante illustre un exemple de déploiement avec plusieurs éléments de base et des clients Lustre connectés via LNet.

Présentation de la solution de stockage Lustre avec NetApp E-Series



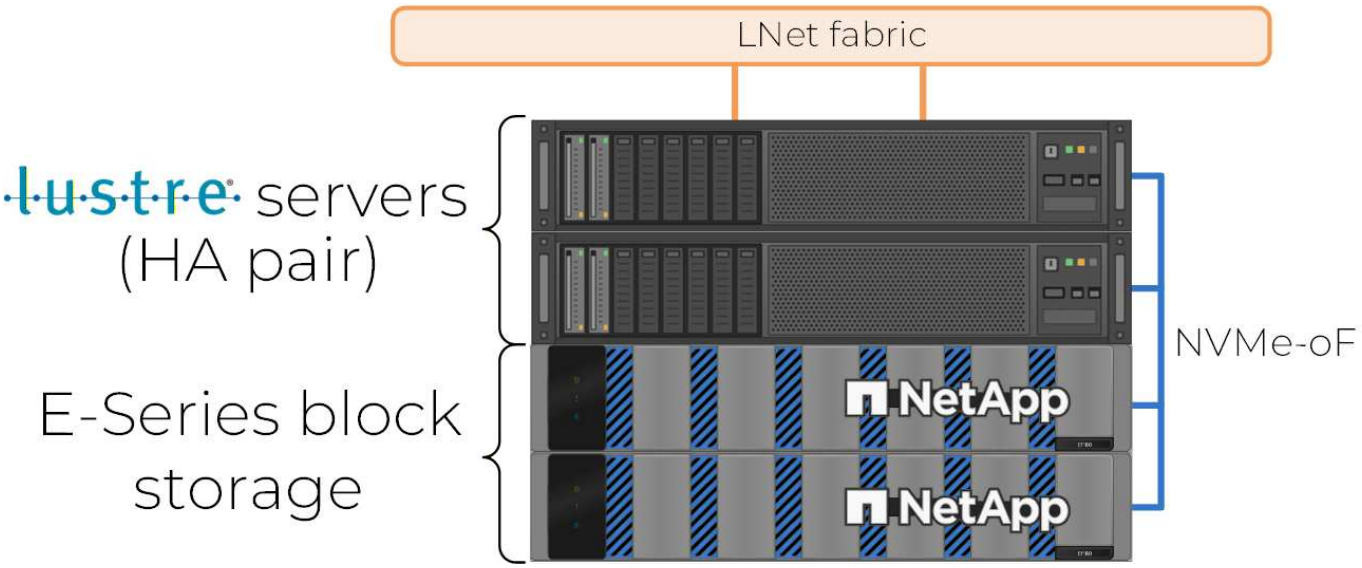
Lustre avec NetApp E-Series Storage - Architecture de la solution

Découvrez comment Lustre avec NetApp système de stockage E-Series utilise des éléments de base, la haute disponibilité, la topologie réseau et les clients afin que vous puissiez planifier la capacité, les performances et le basculement.

Conception d'éléments de base

Un élément de base est l'unité évolutive fondamentale de la solution NetApp E-Series Lustre. Chaque élément de base se compose de deux nœuds serveur configurés en tant que paire à haute disponibilité (HA) fournissant les fonctions de serveur de stockage d'objets Lustre (OSS) et de serveur de métadonnées (MDS), de deux baies NetApp E-Series 100 % flash, d'une connectivité backend dédiée NVMe over Fabrics (NVMe-oF) entre les serveurs et les baies, ainsi que d'une connectivité réseau frontend Lustre (LNet) dédiée pour les clients et la communication inter-nœuds. Un cluster HA Pacemaker et Corosync peut contenir la paire de serveurs d'un ou de plusieurs éléments de base. La collection NetApp `ansible-lustre` utilise l'automatisation Ansible pour déployer et configurer les services Lustre.

élément de base Lustre



Les éléments de base s'adaptent à la croissance requise du système de fichiers. Chaque système de fichiers nécessite un élément de base. L'élément de base héberge le serveur de gestion (MGS) sur une cible de gestion (MGT) et fournit la capacité initiale de la cible de métadonnées (MDT) et de la cible de stockage d'objets (OST). Les éléments de base supplémentaires étendent le système de fichiers et enregistrent leurs cibles MDT et OST auprès du MGS de l'élément de base. Cette approche modulaire permet à la capacité et aux performances de s'adapter indépendamment selon les besoins de la charge de travail.

NetApp recommande les profils de configuration d'élément de base suivants pour la plupart des déploiements. Les nombres de MGS, MDT et OST indiqués sont des valeurs par défaut recommandées, optimisées pour maximiser les performances et la capacité des systèmes de stockage E-Series. Ajustez le nombre de cibles, la taille des volumes et l'allocation des disques E-Series dans l'inventaire Ansible afin de répondre aux besoins de la charge de travail. Par exemple, les déploiements nécessitant une plus grande capacité de stockage de métadonnées peuvent prévoir davantage de MDT et moins d'OST au sein du même matériel d'élément de base.

Le tableau suivant récapitule les types d'éléments de base et les nombres cibles recommandés.

Type	MGS	MDTs	OST	Utiliser
Base	1	8	32	Premier élément de base d'un système de fichiers. Héberge le MGS ainsi que la capacité initiale du MDT et de l'OST.
MDT+OST	0	8	32	Ajoute des métadonnées et de la capacité de données. S'enregistre auprès de l'élément de base MGS.
OST uniquement	0	0	32	Augmente uniquement la capacité de données. Enregistrez-vous auprès de l'élément de base MGS.

- **Type de disque et configuration du pool :** E-Series prend en charge les groupes de volumes RAID traditionnels et les Dynamic Disk Pools (DDP). Pour les disques NVMe TLC, utilisez des groupes de volumes RAID 6 pour le stockage OST et des groupes de volumes RAID 1 pour le stockage MGS et MDT. Pour les disques NVMe QLC, utilisez DDP pour le pool de disques partagé.
- **Répartition des cibles :** Les cibles Lustre de chaque élément de base sont réparties de manière équilibrée entre les nœuds de serveur OSS/MDS et les deux baies E-Series. La configuration répartit les E/S entre les processeurs des serveurs et les contrôleurs de baies afin d'améliorer les performances du chemin NVMe-oF et de maintenir la redondance. Voir "distribution cible et connectivité NVMe-oF" dans "Composants matériels".

Les systèmes d'exploitation pris en charge, les versions de Lustre, le firmware de la baie et les composants élément de base associés sont répertoriés dans "Outil de matrice d'interopérabilité (IMT) NetApp" sous **E-Series Lustre**. Pour des informations détaillées sur le nombre de volumes, la disposition des disques et les recommandations de dimensionnement, consultez "Composants matériels". Pour les composants logiciels, consultez "Composants logiciels".

Recommandations de mise à l'échelle

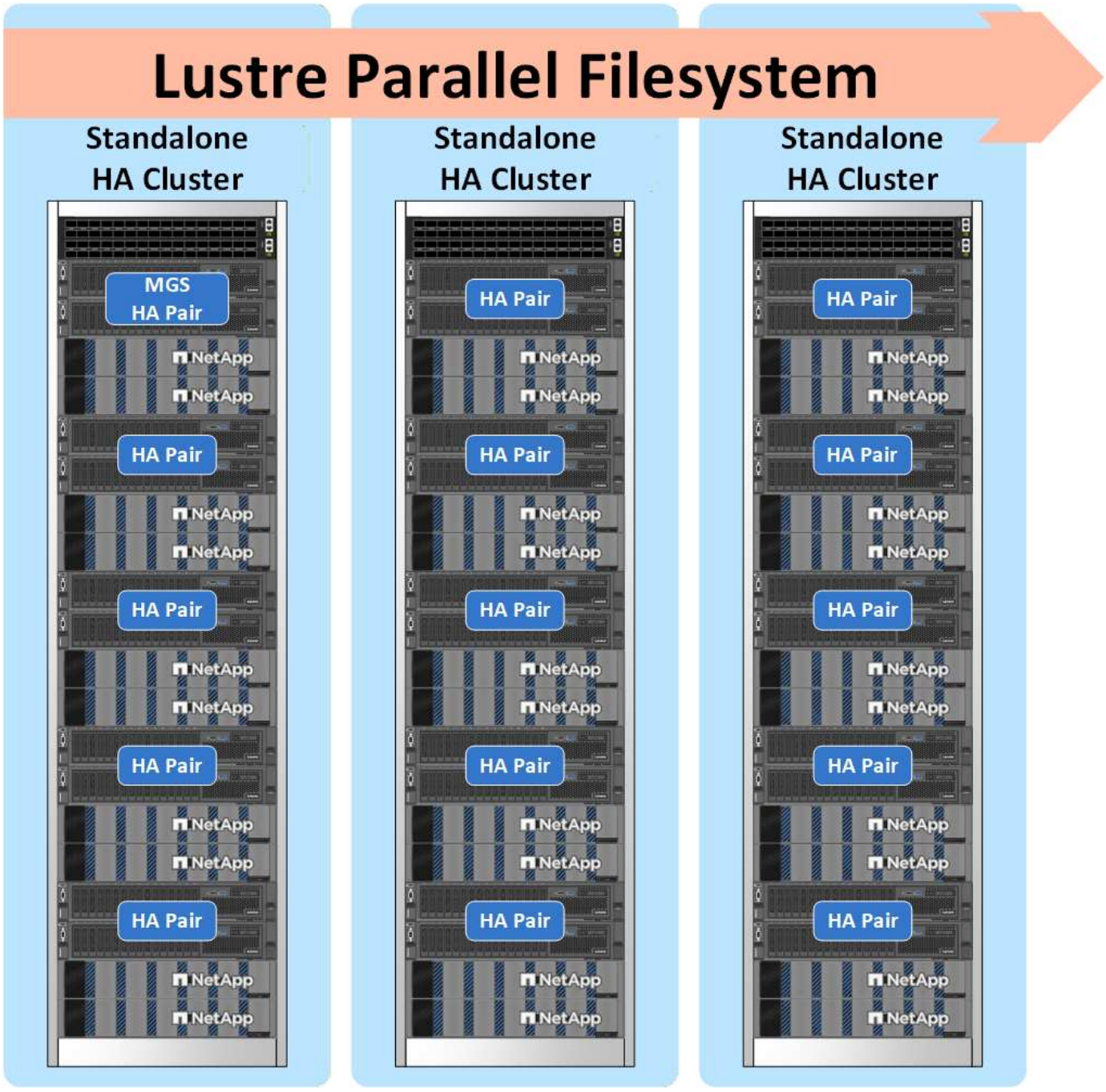
Le système de fichiers Lustre s'adapte en ajoutant des éléments de base lorsque la capacité des métadonnées, des données, ou les deux, devient limitante. Ajustez les MDT en fonction du nombre de fichiers et des IOPS des métadonnées ; ajustez les OST en fonction des besoins en capacité et en débit agrégé.

1. **Un seul élément de base par système de fichiers :** Déployez exactement un élément de base ; il héberge le MGS et les cibles initiales MDT et OST.
2. **Profil d'élément de base supplémentaire :** Ajoutez un élément de base OST uniquement lorsque la capacité MDT existante est suffisante et que la capacité ou le débit de données doit être augmenté. Ajoutez un élément de base MDT+OST lorsque les performances des métadonnées ou la capacité de l'espace de noms deviennent le facteur limitant.

3. **Limite de cluster HA :** Limitez chaque cluster HA Pacemaker et Corosync à cinq éléments de base (dix nœuds de serveur Lustre). Répartissez les déploiements importants de manière égale en plusieurs clusters HA afin d'éviter les problèmes de ressources dans les grandes configurations.

La figure suivante illustre jusqu'à cinq éléments de base empilés dans un rack 42U (un cluster Pacemaker HA par rack). Plusieurs racks peuvent participer à un même système de fichiers Lustre ; seul l'élément de base héberge le MGS.

Mise à l'échelle des racks du système de fichiers parallèle Lustre



Pour des exemples de dimensionnement, voir "Guide de dimensionnement".

Architecture HA

Lustre avec le système de stockage NetApp E-Series utilise une architecture haute disponibilité (HA) à disque partagé intégrée au système de stockage NetApp E-Series. Pacemaker gère les ressources HA, et Corosync assure l'appartenance au cluster et la messagerie. Ensemble, ils gèrent le basculement de la cible Lustre entre les deux nœuds de serveur OSS/MDS de chaque élément de base. La technologie NVMe-oF multipath connecte les deux nœuds de serveur OSS/MDS aux mêmes volumes E-Series, permettant ainsi à chaque nœud de prendre le relais des services cibles en cas de besoin.

Chaque MGS, MDT et OST est configuré comme une ressource HA avec ses dépendances dans un groupe de ressources Pacemaker. Pacemaker garantit que les ressources démarrent et s'arrêtent dans le bon ordre, restent colocalisées sur le même nœud et ne s'exécutent que sur un seul nœud à la fois. Un accès simultané à la même cible depuis les deux nœuds pourrait corrompre le système de fichiers sous-jacent.

- **Basculement de cible :** Les deux nœuds du serveur OSS/MDS sont pairs au sein du cluster, mais chaque cible Lustre n'est active que sur un seul nœud à la fois. Une ressource de surveillance Pacemaker contrôle l'état de chaque cible et de ses dépendances et déclenche un basculement lorsqu'une cible devient indisponible sur son nœud actuel. En cas de panne, Pacemaker redémarre le groupe de ressources concerné sur le nœud partenaire dans l'ordre approprié, et les clients se reconnectent automatiquement à la cible à son nouvel emplacement une fois les services rétablis. Chaque cible étant active sur un seul nœud, le système de fichiers n'est jamais exposé à des écritures simultanées provenant des deux nœuds.
- **Accès au stockage partagé :** Les chemins NVMe-oF multipath connectent chaque nœud serveur OSS/MDS à tous les contrôleurs E-Series dans l'élément de base, de sorte que chaque volume cible est accessible depuis n'importe quel nœud. Si un nœud serveur tombe en panne, le nœud partenaire dispose déjà de chemins actifs vers les mêmes volumes et peut prendre en charge les cibles sans reconfigurer la connectivité du stockage. Si un contrôleur de baie ou un chemin tombe en panne, le multipath redirige les E/S vers un contrôleur restant. Cette double redondance permet aux cibles Lustre de basculer entre les nœuds serveurs OSS/MDS ou les contrôleurs de baie sans perdre l'accès aux volumes sous-jacents.
- **Isolation STONITH :** En cas de panne, Pacemaker peut parfois ne pas pouvoir communiquer avec le nœud défaillant pour confirmer que ses cibles sont arrêtées. Avant de redémarrer ces cibles ailleurs, Pacemaker isole le nœud défaillant, idéalement en coupant son alimentation, afin de garantir qu'il est hors service. L'isolation empêche un scénario de « split-brain » dans lequel les deux nœuds accèdent simultanément à la même cible et corrompent le système de fichiers sous-jacent, préservant ainsi l'intégrité des données lors d'un basculement. NetApp recommande `fence_redfish` pour les serveurs dotés de contrôleurs de gestion de carte mère (BMC) compatibles Redfish. D'autres agents d'isolation, tels que `fence_apc`, sont également pris en charge.

Pour l'administration du cluster et la configuration du fencing, consultez le "Guide d'utilisation de HA" dans le "collection ansible-lustre".

Architecture réseau

La solution utilise des chemins réseau distincts pour le trafic du backend de stockage et le trafic du frontend LNet.

- **Backend (NVMe-oF) :** Les nœuds serveur OSS/MDS se connectent aux baies E-Series via NVMe-oF en utilisant NVMe/InfiniBand ou NVMe/RoCE. Le chemin NVMe-oF transporte les E/S par blocs entre les services cibles Lustre et les volumes E-Series. Pour le câblage backend EF80 à six HCA, voir "Matériel Lustre pour rack et câbles". Pour le placement optimal des cibles sur les nœuds et les baies, voir "distribution cible" dans "Composants matériels".
- **Frontend (LNet) :** Les clients Lustre et les nœuds serveurs OSS/MDS communiquent via LNet en utilisant le type de réseau `@o2ib` avec la pile NVIDIA OFED. InfiniBand et RoCE sont tous deux des transports frontend validés. La collection ansible-lustre configure toutes les interfaces frontend pour LNet multi-rails,

qui agrège plusieurs ports afin d'augmenter la bande passante et d'assurer la redondance des chemins pour le trafic client et inter-nœuds.

- **Gestion et cluster :** Un réseau de gestion hors bande assure la gestion des serveurs, l'accès au BMC et la gestion des baies. Les agents de clôture STONITH, tels que `fence_redfish`, utilisent ce réseau pour accéder aux BMC des serveurs et couper l'alimentation d'un nœud défaillant. Corosync peut également exécuter la communication du cluster via ce réseau en tant qu'anneau supplémentaire, parallèlement à la structure frontale. Configurer Corosync avec plusieurs anneaux ajoute de la redondance pour la messagerie du cluster et réduit le risque qu'une seule panne réseau perturbe le quorum.

Clients Lustre

Un client Lustre charge le module noyau client, établit une connectivité LNet avec les nœuds serveurs OSS/MDS et monte le système de fichiers afin de présenter aux applications un espace de noms unique, cohérent et conforme à POSIX. Les E/S sont distribuées en parallèle entre les OST actifs. Les nœuds clients sont externes à un élément de base et utilisent le système de fichiers via le tissu LNet frontal. Les systèmes d'exploitation clients ne sont pas listés dans IMT pour cette solution ; utilisez la "Matrice de support Lustre" pour connaître les distributions clientes et les versions de noyau testées et prises en charge par la version de Lustre déployée (par exemple, Red Hat Enterprise Linux 9, SUSE Linux Enterprise Server 15 et Ubuntu 24.04).

Les nœuds clients nécessitent une connectivité au même fabric LNet et au même type de réseau que les nœuds serveurs OSS/MDS. Si nécessaire, un routeur LNet peut relier des sous-réseaux ou des fabrics LNet distincts.

NetApp fournit des paquets RPM serveur Lustre précompilés pour Rocky Linux 9.8 et Red Hat Enterprise Linux 9.8. NetApp ne fournit pas de paquets clients précompilés. Compilez les paquets clients pour le système d'exploitation et le noyau du client à partir du code source de Lustre dans le "dépôt netapp-lustre".

Une fois les paquets clients installés, le rôle optionnel `lustre_client` dans le "collection ansible-lustre" configure les interfaces réseau client et LNet, active LNet multi-rails lorsqu'il est sélectionné, valide la connectivité et gère une unité de montage systemd persistante. Le rôle ne compile pas Lustre. Il installe les paquets clients uniquement lorsque `lustre_client_packages` est renseigné. Si vous le souhaitez, les clients peuvent être configurés manuellement. Pour des procédures étape par étape, consultez "Déployez la solution" et le "documentation du rôle lustre_client".

Lustre avec le stockage NetApp E-Series - Composants matériels

Utilisez ces exigences matérielles pour les serveurs, les baies NetApp E-Series, l'agencement du stockage et la mise en réseau lorsque vous dimensionnez et commandez Lustre avec le stockage NetApp E-Series. Pour les composants logiciels, consultez "Composants logiciels".

Configuration requise pour le serveur

La solution utilise un modèle BYOS (Bring Your Own Server). Chaque élément de base nécessite deux nœuds de serveur OSS/MDS qui respectent ou dépassent les spécifications suivantes.

Spécifications des nœuds

Le tableau suivant répertorie les exigences minimales du serveur par nœud OSS/MDS.

Composant	Exigence
Quantité	2 nœuds par élément de base
Processeur	AMD EPYC ou Intel Xeon, 32 cœurs ou plus
Mémoire	256 Go DDR5 (minimum)
HCAs réseau	6 HCA double port 200Gb (voir Connectivité HCA)
emplacements PCIe	2× PCIe Gen5 x16 et 4× PCIe Gen5 x8 (voir Exigences relatives aux emplacements PCIe)
Disques de démarrage	2× disques en RAID 1 (RAID logiciel ou matériel recommandé)

NetApp a validé la solution à l'aide de serveurs Lenovo ThinkSystem SR665 V3. Les serveurs équivalents d'autres fournisseurs sont pris en charge s'ils répondent à ces exigences.

Connectivité HCA

Le tableau suivant répertorie les exigences HCA, port frontal LNet et chemin NVMe-oF par nœud de serveur OSS/MDS.

HCA par nœud	Ports LNet par nœud Lustre	Chemins NVMe-oF par nœud Lustre	Exemple HCA
6 (12 ports)	4	8 au total (4 par baie)	MCX755106AS-HEAT (carte PCIe gen5 double port 200Gb)

NetApp recommande six HCA par nœud pour maximiser la bande passante des baies de stockage EF80.

Exigences relatives aux emplacements PCIe

Chaque nœud serveur OSS/MDS d'un élément de base EF80 comprend six cartes HCA double port : deux pour LNet et quatre pour NVMe-oF. La largeur de l'emplacement détermine la bande passante disponible pour chaque HCA, donc veuillez vérifier la largeur électrique de chaque emplacement et de chaque riser, et non celle de la carte seule. Une carte Gen5 x16 installée dans un emplacement Gen5 x8 fonctionne en x8.

- **Cartes LNet HCA :** Installez-les dans les emplacements PCIe Gen5 x16 pour atteindre le débit frontal maximal.
- **Cartes HCA NVMe-oF :** Installez-les dans des emplacements PCIe Gen5 x8 ou PCIe Gen5 x16.
- **Équilibrage NUMA :** Répartissez les HCAs de manière égale entre les deux zones NUMA afin que chaque zone héberge deux interfaces LNet et quatre interfaces NVMe-oF.

Le tableau suivant répertorie les emplacements minimums pour chaque nœud serveur OSS/MDS dans un élément de base EF80.

Type de trafic	HCA par nœud	Largeur minimale de la fente	Emplacements par zone NUMA
LNet	2	PCIe Gen5 x16	1
NVMe-oF	4	PCIe Gen5 x8	2

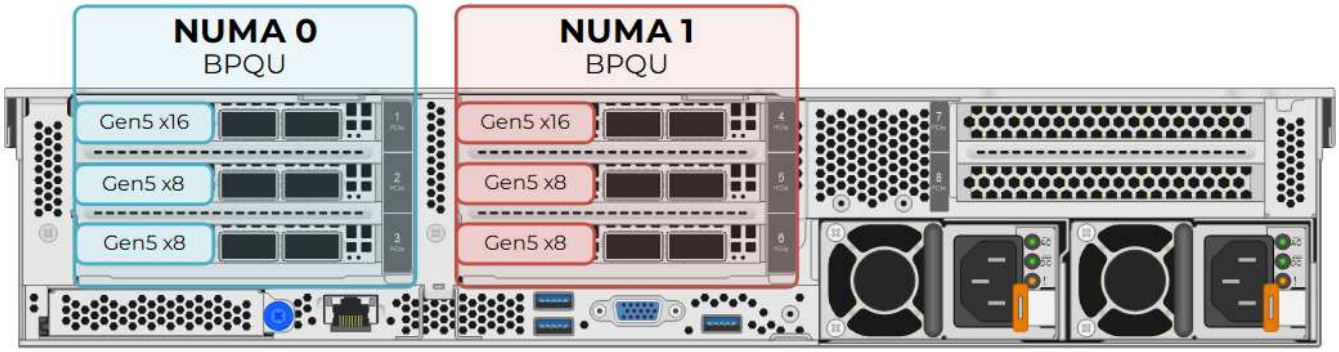
Vous pouvez, si vous le souhaitez, répartir les ports d'une carte HCA double port afin qu'un port prenne en charge le trafic LNet et l'autre le trafic NVMe-oF. Installez ces quatre cartes HCA dans des emplacements PCIe Gen5 x16 pour que chaque port LNet atteigne son débit maximal, puis ajoutez deux cartes HCA supplémentaires dans des emplacements Gen5 x8 ou Gen5 x16 pour les ports NVMe-oF restants.

▼ **Exemples de configurations de riser pour Lenovo ThinkSystem SR665 V3**

Les deux combinaisons de risers suivantes répondent aux exigences d'emplacement pour un élément de base EF80.

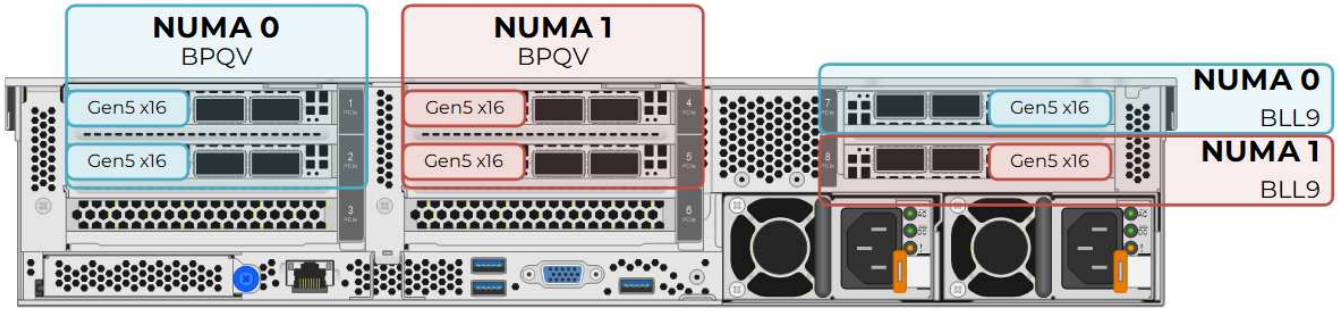
Largeurs minimales des fentes

Installez un élévateur BPQU dans les positions d'élévateur 1 et 2. Chaque élévateur BPQU fournit un emplacement PCIe Gen5 x16 et deux emplacements PCIe Gen5 x8. Ensemble, les élévateurs offrent deux emplacements Gen5 x16 pour LNet et quatre emplacements Gen5 x8 pour NVMe-oF, répartis équitablement entre les zones NUMA.



Tous les emplacements Gen5 x16

Installez un élévateur BPQV aux positions d'élévateur 1 et 2, et un élévateur BLL9 à la position d'élévateur 3. Chaque élévateur BPQV fournit deux emplacements PCIe Gen5 x16. L'élévateur BLL9 fournit l'emplacement 7 sur la zone NUMA 0 et l'emplacement 8 sur la zone NUMA 1, tous deux PCIe Gen5 x16. Cette combinaison offre six emplacements Gen5 x16, trois par zone NUMA, et prend en charge la répartition du trafic LNet et NVMe-oF sur les ports du même HCA.



élément de base EF80 standard

La plateforme EF80 est la plateforme standard validée pour cette version de la solution.

Spécifications du tableau

Le tableau suivant répertorie les spécifications des matrices EF80 pour un élément de base (deux matrices).

Composant	Spécification
Modèle	NetApp EF80
facteur de forme	Châssis de base 2U, 24 emplacements SSD NVMe internes
Contrôleurs	Deux contrôleurs (A et B)
Disques	24× SSD NVMe par baie
connectivité E/S	8 ports hôtes NVMe/IB ou NVMe/RoCE de 200 Gb par baie dans la conception Lustre validée

La plateforme EF80 prend en charge jusqu'à douze ports hôtes par baie lorsque trois modules d'E/S hôte à deux ports sont installés dans chaque contrôleur. La conception Lustre validée utilise des modules d'E/S hôte dans les emplacements 1 et 2 pour huit ports hôtes par baie.

Chaque baie EF80 utilise également le module d'E/S dédié de l'emplacement 4 pour la mise en miroir inter-contrôleurs. Connectez le port 4a du contrôleur A au port 4a du contrôleur B, et le port 4b du contrôleur A au port 4b du contrôleur B. Ces connexions assurent la mise en miroir du cache et le transfert des E/S, et ne sont pas utilisées pour le trafic NVMe-oF de l'hôte. Voir "Câblez les connexions de mise en miroir intercontrôleurs EF50 et EF80".

Pour connaître l'intégralité des spécifications de la série EF pour tous les modèles, consultez le "NetApp Fiche technique de la baie 100 % flash de la série EF".

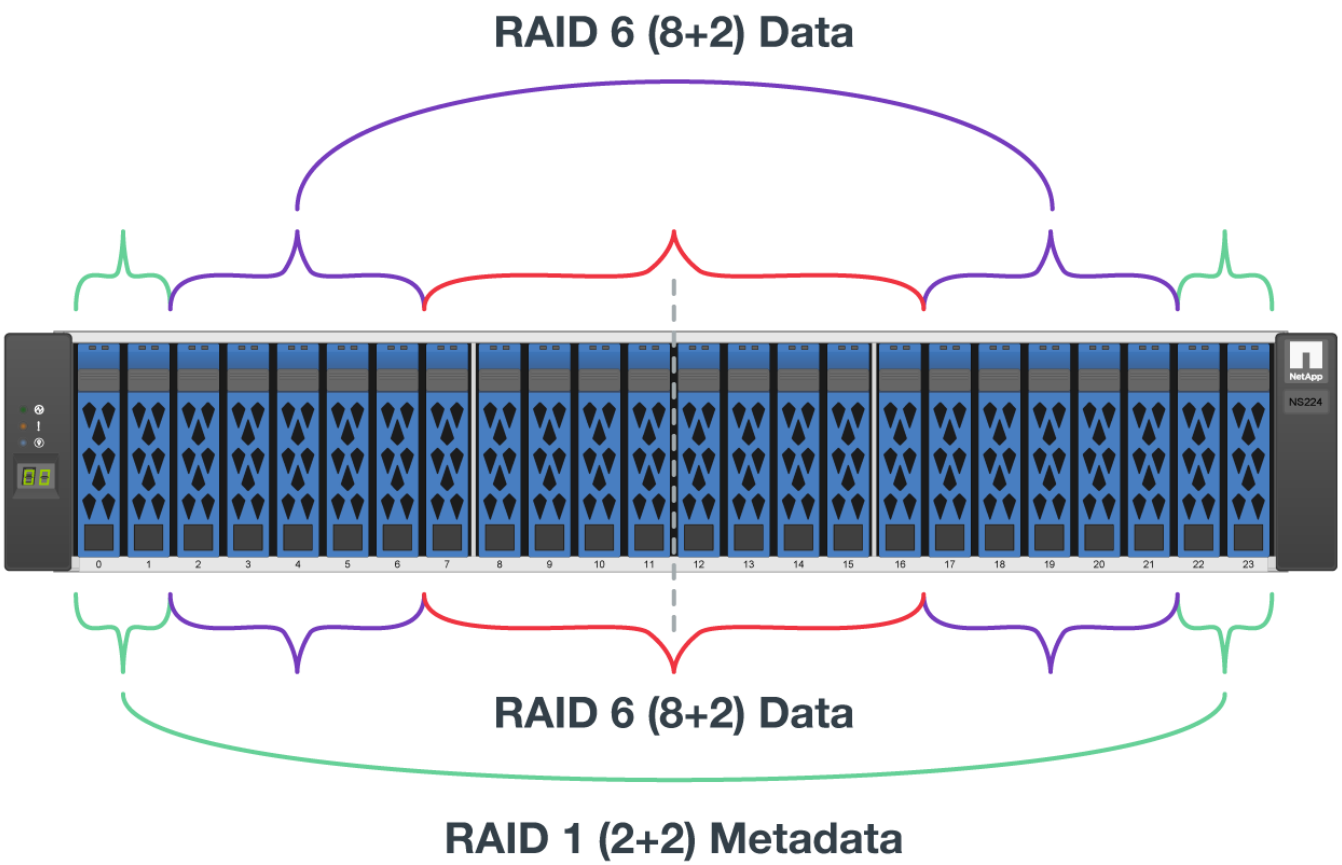
Disposition des disques (24 disques par baie)

Chaque baie EF80 d'un élément de base utilise la totalité des vingt-quatre emplacements pour disques NVMe :

- 4× disques en RAID 1 pour le stockage MGS/MDT (groupe de volumes partagés ou allocation DDP)
- 10× disques en RAID 6 pour le stockage OST (premier pool OST)
- 10 disques en RAID 6 pour le stockage OST (deuxième pool OST)

Cette configuration s'applique lorsque vous utilisez des disques de 3,84 To, 7,68 To ou 15,3 To avec des groupes de volumes traditionnels. Lorsque vous utilisez des disques Capacity Flash (QLC) de 30,7 To ou 61,4 To, provisionnez un seul pool de disques dynamiques pour l'ensemble des vingt-quatre disques et créez des volumes RAID 1 au sein du pool pour les volumes MGS et MDT, tout en utilisant des volumes RAID 6 par défaut pour les OST.

Les disques durs de 1,92 To ne sont actuellement pas recommandés pour cette solution. Utilisez l'une des capacités de disque validées listées ci-dessus.



Volume et nombre de cibles (élément de base)

Le tableau suivant répertorie les nombres MGS, MDT et OST pour un élément de base.

Type de cible	Nombre par BB	RAID / pool	Remarques
MGS	1	RAID 1	élément de base uniquement; tableau 1
MDT	8	RAID 1	4 par tableau
OST	32	RAID 6 (TLC) ou DDP (QLC)	16 par tableau

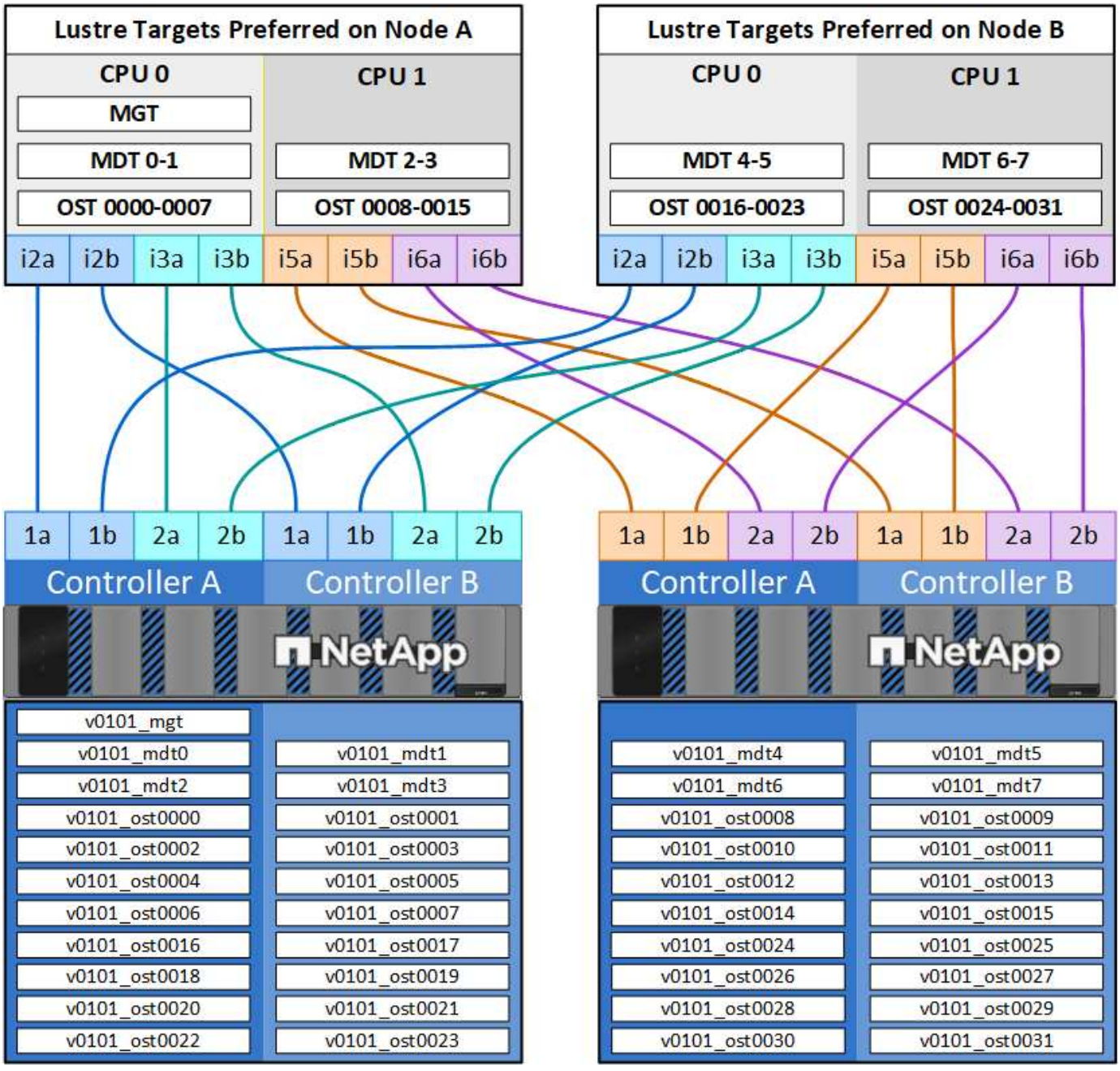
Utilisez les directives de dimensionnement de volume suivantes comme point de départ dans l'inventaire Ansible.

Volume	Taille recommandée	Remarques
MGS	5–10 Gio	Données de configuration uniquement
MDT	Capacité RAID 1 ÷ Nombre de MDT	Environ 1 à 2 TiB chacun (valeur typique)
OST	Capacité RAID 6 ou DDP ÷ Nombre d'OST par pool	Évolue avec la capacité du lecteur ; voir "Guide de dimensionnement"

Répartition du volume de l'élément de base principal

La figure suivante illustre l'emplacement recommandé des cibles Lustre et la connectivité NVMe-oF entre les nœuds de serveur OSS/MDS et les baies E-Series dans un élément de base EF80. Le schéma désigne la cible de gestion par **MGT** (cible de gestion) ; une cible MGT héberge le service MGS (serveur de gestion) mentionné ailleurs dans ce document.

Distribution cible de l'élément de base (EF80)



Le tableau suivant indique comment les volumes sont répartis entre les deux baies et quel serveur chaque cible préfère.

Type de cible	Baie 1	Baie 2	Serveur 1	Serveur 2	Remarques
MGS	1	0	1	0	élément de base uniquement

Type de cible	Baie 1	Baie 2	Serveur 1	Serveur 2	Remarques
MDT	4	4	4	4	Chaque serveur privilégie 2 MDT de chaque array
OST	16	16	16	16	8 par pool RAID 6 × 2 pools par baie, ou allocation DDP équivalente
Volumes totaux	**21**	**20**	**21**	**20**	

Sélection du pool de stockage : TLC vs QLC

Choisissez le type de pool de la série E en fonction de la capacité des disques NVMe dans les baies :

- **Disques de 3,84 To, 7,68 To et 15,3 To :** Utilisez des **groupes de volumes RAID 6** pour les volumes OST et des **groupes de volumes RAID 1** pour les volumes MGS et MDT, ou utilisez DDP pour le pool de disques partagé.
- **Disques Flash (QLC) de 30,7 To et 61,4 To :** utilisez uniquement les **Dynamic Disk Pools (DDP)**. Créez des volumes RAID 1 au sein du DDP pour le stockage MGS et MDT. Créez des volumes RAID 6 pour le stockage OST à partir du DDP.

Pour obtenir des estimations de la capacité utilisable par élément de base en fonction de la taille et de la disposition des disques, voir "Guide de dimensionnement".

Exigences réseau

Backend (NVMe-oF)

- NVMe/InfiniBand ou NVMe/RoCE entre chaque nœud OSS/MDS et chaque baie E-Series
- MTU 9000 sur les interfaces backend NVMe/RoCE (typique)
- Huit chemins relient chaque nœud Lustre aux baies de stockage (quatre à chaque baie)
- EF80 utilise six HCA par nœud (i2, i3, i5 et i6 pour NVMe-oF ; i1 et i4 pour LNet). Connectez le nœud A aux ports du contrôleur dont les étiquettes se terminent par `a`, comme `1a` et `2a`. Connectez le nœud B aux ports du contrôleur dont les étiquettes se terminent par `b`, comme `1b` et `2b`. Voir "Câblage backend EF80 six-HCA".

Frontend (LNet)

- IPoIB ou RoCE pour le type de réseau LNet (`@o2ib`)
- MTU 9000 sur les interfaces frontales RoCE (typique)
- LNet multirail recommandé lorsque quatre ports frontaux sont disponibles (EF80 : i1 et i4 HCAs, les deux ports)
- Infrastructure de commutation dédiée InfiniBand ou RoCE reliant les nœuds serveurs OSS/MDS et les clients Lustre
- RoCE sans perte (PFC) recommandé sur les fabrics RoCE

Gestion

- Réseau de gestion hors bande pour les BMC des serveurs et les ports de gestion des baies
- Réseau Corosync dédié ou infrastructure frontale partagée (selon le site)

Lustre avec le système de stockage E-Series NetApp - Composants logiciels

Examinez la pile logicielle et l'automatisation Ansible qui déploient Lustre avec le système de stockage NetApp E-Series. Pour les serveurs, le stockage et les réseaux, voir "Composants matériels".

Logiciel élément de base

Le tableau suivant répertorie les composants logiciels utilisés par les baies E-Series et les nœuds serveur OSS/MDS. Utilisez la "Outil de matrice d'interopérabilité (IMT) NetApp" comme source officielle pour connaître les versions et combinaisons de composants prises en charge. Recherchez **E-Series Lustre** et confirmez les versions du système d'exploitation, de Lustre, de SANtricity OS, de DOCA-OFED, du firmware HCA et du package HA avant le déploiement.

Composant	Fonction
OS SANtricity	Fonctionne sur chaque système de stockage E-Series et fournit SANtricity System Manager, une haute disponibilité à double contrôleur actif avec basculement automatique des chemins d'E/S, un équilibrage de charge automatique et une surveillance des chemins, des Dynamic Disk Pools et un RAID traditionnel, des reconstructions automatiques de disques, une protection de cache en miroir, Data Assurance, une surveillance proactive de l'état des disques, ainsi que des modifications en ligne des logiciels, du firmware et de la configuration.
Lustre	Ce système offre aux clients un espace de noms unique conforme à la norme POSIX tout en répartissant les métadonnées et les données de fichiers sur plusieurs cibles pour un accès parallèle. Le serveur de gestion (MGS) stocke la configuration du système de fichiers, les cibles de métadonnées (MDT) gèrent l'espace de noms et les cibles de stockage d'objets (OST) stockent les données de fichiers sur des volumes E-Series. Cette séparation permet de faire évoluer les performances et la capacité par l'ajout d'éléments de base.
Red Hat Enterprise Linux (RHEL) ou Rocky Linux	Fournit le noyau, les services système et l'infrastructure de paquets pour les services cibles Lustre, le clustering haute disponibilité, ainsi que la connectivité de stockage et réseau. Les dépôts des systèmes d'exploitation pris en charge fournissent également Pacemaker, Corosync, les agents de fence, NVMe-oF et les paquets multipath, testés ensemble en tant que pile de solutions.

Composant	Fonction
Services de cluster à haute disponibilité (Pacemaker, Corosync et fence agents)	Ensemble, ces services assurent une haute disponibilité coordonnée pour les cibles Lustre sur les nœuds de serveur OSS/MDS. Ils maintiennent l'appartenance au cluster et le quorum, surveillent les ressources, orchestrent le basculement ordonné et isolent les nœuds non réactifs avant d'activer les cibles partagées ailleurs. Cette coordination empêche un accès en mode split-brain aux volumes E-Series et protège l'intégrité du système de fichiers en cas de défaillance.
NVIDIA DOCA-OFED	Fournit des pilotes HCA, des bibliothèques RDMA et des utilitaires de gestion pour InfiniBand et les architectures RoCE. La même pile logicielle prend en charge l'accès NVMe-oF à faible latence au système de stockage E-Series en backend et la communication LNet à haut débit en frontend avec les clients Lustre.

NetApp fournit des RPM préconfigurés pour serveur Lustre pour Rocky Linux 9.8 et Red Hat Enterprise Linux 9.8.

Automatisation Ansible

Lustre avec le NetApp système de stockage E-Series est livré et déployé à l'aide de l'automatisation Ansible depuis un nœud de contrôle disposant d'un accès de gestion aux baies E-Series et aux nœuds serveurs OSS/MDS. Un inventaire Ansible modélise la solution souhaitée, incluant les baies, les serveurs, les affectations de stockage, les interfaces réseau et les ressources du cluster HA.

La collection Ansible Lustre NetApp E-Series orchestre le déploiement de bout en bout en utilisant les collections SANtricity et Host. Ensemble, ces collections provisionnent et mappent le système de stockage E-Series, préparent le système d'exploitation du serveur, configurent la connectivité NVMe-oF et LNet, déploient les cibles Lustre et créent les ressources Pacemaker. Cette automatisation rend la configuration reproductible et simplifie l'ajout d'éléments de base à un système de fichiers existant.

Le tableau suivant récapitule les principaux logiciels utilisés sur le nœud de contrôle Ansible. Consultez la "Collection Ansible NetApp E-Series Lustre" pour connaître les dépendances logicielles actuelles.

Package	Dépendance ou finalité
Python	Fournit l'environnement d'exécution pour Ansible et les modules Python requis.
ansible-core	Nécessite Python et exécute les collections Ansible NetApp E-Series.
Packages Python supplémentaires	`cryptography`, `ipaddr`, `netaddr`, `passlib`, `resolvelib`, `jinja2` et `pyyaml`
Collection Ansible NetApp E-Series Lustre	Configure Lustre, LNet, la connectivité hôte NVMe-oF et les ressources Pacemaker.
NetApp Collection Ansible SANtricity E-Series	Configure les baies E-Series, les pools de stockage, les volumes et les mappages d'hôtes via l'API Web Services SANtricity.
NetApp Collection Ansible de l'hôte E-Series	Prépare les nœuds serveur OSS/MDS pour la connectivité LNet et NVMe-oF aux systèmes de stockage E-Series.

Logiciel client Lustre

Le logiciel client Lustre permet à un système d'accéder au système de fichiers parallèle via le réseau frontal LNet. Chaque client nécessite des modules noyau Lustre et des utilitaires client Lustre compatibles à la fois avec son noyau en cours d'exécution et avec la version de Lustre déployée sur les serveurs.

Utilisez "Matrice de support Lustre" pour trouver un système d'exploitation client et un noyau compatibles avec votre version de Lustre. NetApp ne fournit pas de paquets clients Lustre précompilés. Compilez les paquets clients pour le système d'exploitation client et le noyau à partir du code source Lustre fourni par NetApp dans le "dépôt netapp-lustre". Une fois les paquets clients Lustre installés, le rôle `lustre_client` peut configurer les interfaces réseau, LNet et créer une unité de montage systemd persistante pour le système de fichiers Lustre.

Pour connaître les étapes de déploiement du client, consultez "Déployez la solution".

Lustre avec le stockage NetApp E-Series - Guide de dimensionnement

Dimensionnez Lustre avec les éléments de base de stockage NetApp E-Series selon vos besoins en capacité et en métadonnées, décidez quand ajouter de la capacité et examinez les facteurs qui affectent les performances.

Dimensionnement de la capacité

Un élément de base composé de deux baies EF80 offre la configuration cible Lustre suivante. Pour connaître l'emplacement optimal des cibles et les chemins NVMe-oF, consultez "Répartition du volume de l'élément de base principal" dans "Composants matériels".

Composant	Nombre	Taille de volume typique (par cible)
MGS	1	5-10 Gio
MDT	8	La taille varie en fonction de la capacité du disque et de la configuration RAID
OST	32	La taille varie en fonction de la capacité du disque et de la configuration RAID

Chaque baie EF80 utilise vingt-quatre disques NVMe. Les capacités prises en charge sont de 3,84 To, 7,68 To et 15,3 To avec les groupes de volumes traditionnels (RAID 1 pour MGS/MDT et RAID 6 pour OST) ou DDP, et de 30,7 To ou 61,4 To avec les disques Capacity Flash (QLC) en mode DDP uniquement. Les disques de 1,92 To ne sont actuellement pas recommandés pour cette solution. Pour la disposition des emplacements de disques et la sélection des pools, voir "Composants matériels".

Le tableau suivant indique la capacité utilisable approximative d'un élément de base EF80 (deux baies, 48 disques). La capacité utilisable de la baie n'est pas la même que la capacité utilisable finale du système de fichiers Lustre. L'allocation des inodes MDT, les journaux, le surprovisionnement et les pourcentages de volume d'inventaire réduisent la capacité disponible pour les applications.

Capacité du disque	Disposition (par tableau)	Utilisable approximatif (élément de base)
3,84 To	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	138,08 TB
3,84 To	DDP (24 lecteurs, 2 réservés)	133,63 To
7,68 To	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	276,35 TB
7,68 To	DDP (24)	267,47 To
15,3 To	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	552,88 To
15,3 To	DDP (24)	535,15 To
30,7 To	DDP uniquement	1 070,35 TB
61,4 To	DDP uniquement	2 162,70 TB

Dimensionnez séparément les métadonnées et les données, puis définissez les tailles de volume dans l'inventaire Ansible. Les modèles de déploiement formatent les cibles avec ldiskfs ; MDT utilise le ratio d'inodes par défaut de Lustre et les modèles OST définissent `-i 4096`.

- **Métadonnées (MDT) :** Dimensionnez les MDT en fonction du nombre de fichiers prévu. Prévoyez environ le double du nombre de fichiers attendu pour la croissance. Une MDT sous-dimensionnée peut empêcher la création de nouveaux fichiers même lorsque la capacité de l'OST reste suffisante.
- **Données (OST) :** Dimensionnez les OST en fonction de la capacité utilisable et des besoins de débit.
- **MGS :** Utilisez une petite cible de gestion (5 à 10 Gio) uniquement pour la configuration du système de fichiers.

Ajustez la taille des volumes MDT et OST dans l'inventaire pour la capacité de disque sélectionnée. Modifiez `format_options.mkfsoptions` dans `eseries_lustre_filesystem_mdt` ou `eseries_lustre_filesystem_ost` uniquement lorsqu'un site nécessite un ratio d'inodes différent. Pour plus d'informations sur les ratios d'inodes, consultez la "Lustre Wiki : Réglage Lustre". Pour savoir quand ajouter des éléments de base, consultez Mise à l'échelle.

Mise à l'échelle

Ajoutez des éléments de base lorsque la capacité ou la charge des métadonnées l'exigent :

- **OST uniquement :** Le nombre de fichiers et la charge des métadonnées sont dans la capacité MDT existante, et vous avez besoin de plus de capacité de données ou de débit agrégé. Ajoute 32 OST et deux nœuds de serveur OSS/MDS.
- **MDT+OST :** Le nombre de fichiers ou le débit de métadonnées approche la capacité du MDT, ou vous avez besoin à la fois d'un débit de métadonnées et d'une capacité de données accrus. Ajoute 8 MDT et 32 OST.

Utilisez `mdt.*.md_stats` sur les MDT existants pour confirmer si les métadonnées sont saturées. Limitez chaque cluster Pacemaker/Corosync à cinq éléments de base (dix nœuds de serveur OSS/MDS). Pour les déploiements plus importants, répartissez les éléments de base de manière égale entre plusieurs clusters HA. Voir "Architecture de la solution" pour les règles de mise à l'échelle.

L'exemple suivant illustre un système de fichiers avec un élément de base et un élément de base OST uniquement dans un cluster HA.

Élément de base	Type	Serveurs	Tableaux	MGS	MDT	OST
BB1	Base	2	2	1	8	32
BB2	OST uniquement	2	2	0	0	32
Total		4	4	1	8	64

BB2 enregistre ses cibles OST auprès du MGS dans BB1 en utilisant `mgsnode=`. Les indices OST pour BB2 sont attribués globalement (par exemple, OST0032 à OST0063) afin qu'aucun indice n'entre en conflit avec BB1.

Performances

La validation formelle utilise IOR, mdtest et fio des clients Lustre pour caractériser le débit relatif, les IOPS et le comportement des métadonnées, et pour valider le basculement haute disponibilité. Ces tests ne publient pas de chiffres de performance garantis. Les benchmarks synthétiques mesurent un comportement optimal et peuvent ne pas refléter la performance de l'application.

Les performances augmentent approximativement avec le nombre d'éléments de base. Les résultats obtenus dépendent du type de disque et de la configuration du pool, du nombre de chemins NVMe-oF, du tissu LNet et du nombre de clients, ainsi que de la répartition des E/S entre données et métadonnées.

Contactez votre équipe commerciale NetApp pour obtenir des recommandations de dimensionnement et de performance adaptées à votre charge de travail.

Pour connaître la configuration des disques et le nombre de volumes, consultez "Composants matériels". Pour connaître les étapes de déploiement, consultez "Déployez la solution". Pour obtenir des conseils sur l'inventaire au niveau du terrain, consultez "Personnalisez l'inventaire Ansible de Lustre".

Déployez la solution

Déployez Lustre avec le stockage NetApp E-Series

Déployez Lustre avec le stockage NetApp E-Series en effectuant dans l'ordre les tâches de préparation du matériel, de l'environnement, d'inventaire, de déploiement avec Ansible et de vérification.

Prérequis

Avant de commencer le déploiement, assurez-vous des points suivants :

- Vous avez suivi les instructions de dimensionnement et d'élément de base dans "Guide de dimensionnement" et sélectionné le matériel qui correspond à "Composants matériels".
- Vous avez confirmé que la combinaison serveur, HCA, E-Series, système d'exploitation, Lustre, SANtricity OS et DOCA-OFED est répertoriée pour **E-Series Lustre** dans le "Outil NetApp Interoperability Matrix Tool".
- Tous les nœuds de serveur OSS/MDS et les baies E-Series pour les éléments de base prévus sont sur site et prêts à être installés en rack.
- Vous disposez des identifiants et de l'accès réseau pour les BMC des serveurs, les ports de gestion des

baies et l'hôte qui servira de nœud de contrôle Ansible.

> NetApp prend en charge le déploiement du cluster Lustre HA et des clients Lustre avec le "Collection Ansible NetApp E-Series Lustre". Ne configurez pas manuellement les cibles Lustre, LNet, la connectivité hôte NVMe-oF ou les ressources Pacemaker.

Flux de travail de déploiement

Veuillez effectuer les tâches suivantes dans l'ordre :

1. "Matériel Lustre pour rack et câbles".

   Installez les serveurs Lustre et les baies EF80 dans le rack, puis câblez les fabrics NVMe-oF du backend et LNet du frontend.

2. "Préparez l'environnement de déploiement de Lustre".

   Configurez les baies et les serveurs, puis installez Ansible et ses dépendances sur le nœud de contrôle.

3. "Personnalisez l'inventaire Ansible de Lustre".

   Sélectionnez le modèle pour l'architecture du site et renseignez les serveurs, les baies, les réseaux, les éléments de base et les identifiants.

4. "Déployez le cluster Lustre HA et les clients".

   Préparez le système d'exploitation du serveur, installez NVIDIA DOCA-OFED, exécutez le playbook de déploiement principal et configurez les clients Lustre.

5. "Vérifiez et étendez le déploiement de Lustre".

   Vérifiez l'état du cluster, les points de montage et la capacité avant la mise en production. Étendez l'inventaire existant lorsque le système de fichiers a besoin d'un autre élément de base.

Informations connexes

- "Collection Ansible NetApp E-Series Lustre"
- "NetApp Dépôt source Lustre"
- "Guide de l'utilisateur d'administration HA"
- "Guide d'optimisation des performances"
- "Modèles de déploiement du client Lustre"

Matériel de rack et de câblage Lustre

Installez en rack les serveurs Lustre et les baies NetApp EF80, puis câblez les réseaux NVMe-oF du backend et LNet du frontend pour chaque élément de base.

Avant de commencer

Consultez le nombre de HCA, les exigences relatives aux emplacements PCIe, les exigences relatives aux chemins et les recommandations concernant la MTU dans "Composants matériels".

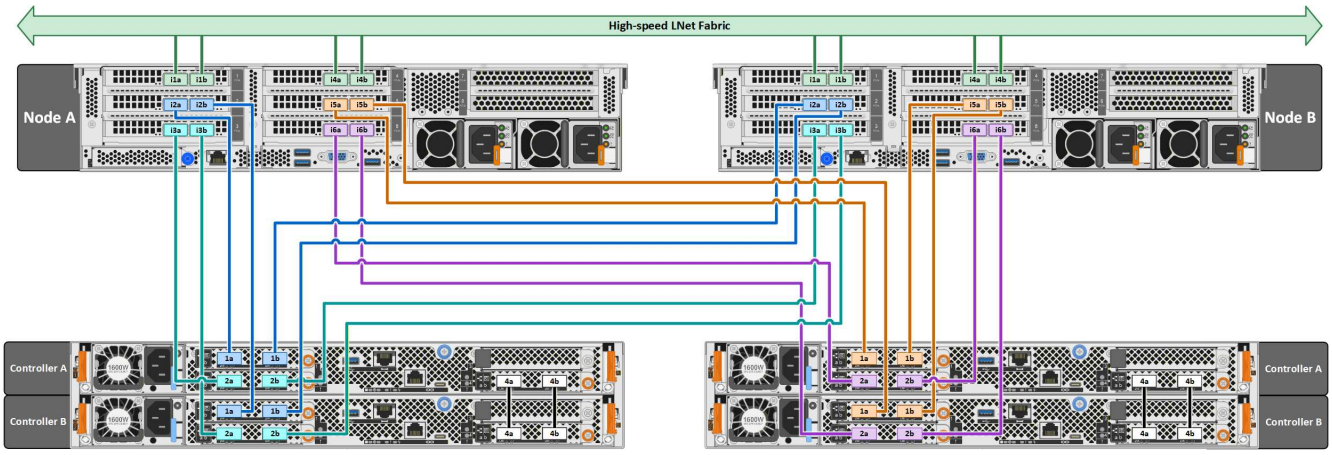
Étapes

1. Installez et câblez tous les serveurs Lustre et les systèmes de stockage E-Series conformément à la topologie de l'élément de base.

2. Pour chaque baie EF80, connectez le port 4a du contrôleur A au port 4a du contrôleur B, et le port 4b du contrôleur A au port 4b du contrôleur B. Ces connexions dédiées assurent la duplication entre les contrôleurs et ne sont pas utilisées pour le trafic hôte. Voir "Câblez les connexions de mise en miroir intercontrôleurs EF50 et EF80".

3. Câblez le réseau NVMe-oF dorsal entre les serveurs et les baies selon la topologie EF80 à six HCA.

Câblage arrière EF80 six-HCA (RoCE ou InfiniBand)



4. Connectez les quatre interfaces LNet frontales (les HCA i1 et i4) sur chaque serveur Lustre et les clients Lustre à la matrice de commutation InfiniBand ou RoCE dédiée.

5. Lors de l'utilisation de RoCE, configurez le réseau pour un fonctionnement sans perte avec le contrôle de flux prioritaire (PFC). Les playbooks Ansible configurent RoCE sans perte sur toutes les interfaces de l'élément de base.

Après avoir terminé

"Préparez les serveurs, les baies et le nœud de contrôle Ansible".

Préparez l'environnement de déploiement de Lustre

Configurez l'accès à la gestion et les paramètres de base sur chaque baie EF80 et serveur Lustre, puis préparez un nœud de contrôle Ansible pour déployer la solution.

Préparez les baies E-Series

Étapes

1. Configurez l'interface de gestion sur chaque contrôleur et vérifiez que SANtricity System Manager est accessible.
2. Définissez le nom de la baie et les informations d'identification administrateur que l'inventaire Ansible utilisera.
3. Vérifiez que la baie est dans un état optimal et que tous les disques sont présents et signalés correctement.

Pour obtenir des instructions sur la configuration de la connexion de gestion, consultez le "Documentation de la série E".

Préparez les serveurs Lustre

Étapes

1. Configurez le contrôleur de gestion de la carte mère (BMC) sur chaque serveur pour un accès hors bande et définissez les paramètres système UEFI pour des performances maximales.
2. Installez un système d'exploitation compatible (RHEL ou Rocky Linux), puis configurez le nom de l'hôte et le port réseau de gestion.
3. Préparez tous les packages de micrologiciels et de logiciels pour répondre aux exigences de la solution **E-Series Lustre**, comme décrit dans "Composants logiciels".

Préparez le nœud de contrôle Ansible

Désignez un système Linux virtuel ou physique disposant d'un accès réseau de gestion à tous les serveurs Lustre et aux baies E-Series comme nœud de contrôle Ansible.

Les étapes suivantes utilisent Ubuntu 24.04. Pour obtenir des instructions concernant une autre distribution Linux, consultez le "Documentation d'installation d'Ansible".

Étapes

1. Installez Python, pip et le paquet d'environnement virtuel Python :

```
sudo apt-get install python3 python3-pip python3-setuptools python3-venv
```

2. Créez et activez un environnement virtuel Python :

```
python3 -m venv ~/pyenv
source ~/pyenv/bin/activate
```

3. Installez Ansible et les packages Python requis dans l'environnement virtuel activé :

```
pip install "ansible-core>=2.19" cryptography jinja2 ipaddr netaddr \
  passlib pyyaml resolvelib
```

4. Installez la collection Ansible Lustre NetApp E-Series. L'installation de la collection installe également ses collections dépendantes, y compris les collections SANtricity et Host :

```
ansible-galaxy collection install netapp_eseries.lustre
```

5. Vérifiez les versions installées d'Ansible, de Python et de la collection :

```
ansible --version
python3 --version
ansible-galaxy collection list netapp_eseries.lustre
```

6. Configurez une connexion SSH sans mot de passe entre le nœud de contrôle et chaque serveur Lustre. Remplacez `<ip_or_hostname>` par l'adresse IP de gestion ou le nom de l'hôte d'un serveur :

```
ssh-keygen
ssh-copy-id <ip_or_hostname>
```

Répétez `ssh-copy-id` pour chaque serveur Lustre.

Ne configurez pas l'accès SSH sans mot de passe aux baies de la série E. Ansible gère les baies via l'API des services Web SANtricity en utilisant les identifiants définis dans l'inventaire.

Après avoir terminé

"Personnalisez l'inventaire Ansible".

Personnalisez l'inventaire Ansible de Lustre

Créez une copie de travail d'un modèle de déploiement EF80 et personnalisez son inventaire pour vos serveurs, baies, réseaux et éléments de base Lustre.

Conservez l'inventaire sélectionné dans son répertoire de structure et de plateforme. Conservez le répertoire partagé `playbooks` et le fichier partagé `passwords.yml` à la racine de `building_blocks`.

Sélectionnez un modèle de déploiement

Commencez par utiliser le modèle qui correspond à l'architecture réseau du site :

- **InfiniBand:** "IB_H2_IB_DAC_A2_EF80/lustre_building_block_cluster_1"
- **RoCE :** "ROCE_H2_ROCE_DAC_A2_EF80/lustre_building_block_cluster_1"

Configurer les paramètres spécifiques au site

Définissez les valeurs spécifiques au site dans l'inventaire, ainsi que les paramètres du référentiel local et des règles udev. Les références de ligne dans le tableau pointent vers les fichiers de modèle RoCE EF80 dans la balise `ansible-lustre release/1.0.0`. Le modèle InfiniBand utilise les mêmes noms de variables et la même structure, à l'exception de la variable d'interface LNet, qui est indiquée dans le tableau.

Champ	Variable Ansible	Fichier modèle (`release/1.0.0`)	Portée	Remarques
Nom du système de fichiers	`eseries_lustre_filesystem_mgt.format_options.fsname`, `eseries_lustre_filesystem_mdt.format_options.fsname`, `eseries_lustre_filesystem_ost.format_options.fsname`	"ha_cluster.yml#L76", "#L90", "#L110"	À l'échelle du cluster	À placer sous `format_options` pour les cartes MGT, MDT et OST. Maximum 8 caractères. Utilisez la même valeur dans les trois emplacements. Une clé `fsname` de premier niveau est sans effet.
Identifiants MGS NIDs pour l'enregistrement des cibles	`eseries_lustre_filesystem_mdt.format_options.mgsnode`, `eseries_lustre_filesystem_ost.format_options.mgsnode`	"ha_cluster.yml#L98", "#L118"	À l'échelle du cluster	Sous MDT et OST `format_options`. Incluez les NID LNet frontaux des deux nœuds de la paire haute disponibilité. Listez d'abord les NID du nœud configuré comme propriétaire MGS préféré, suivis de ceux du nœud partenaire.
Nom du cluster Pacemaker	`eseries_lustre_pacemaker_cib_properties_overrides.cluster_properties.cluster-name`	"ha_cluster.yml#L228"	À l'échelle du cluster	Imbriqué sous les remplacements de propriétés CIB de Pacemaker. Nom unique pour le cluster HA.
Adresses de clôture BMC	`eseries_lustre_pacemaker_fencing_agents.fence_redfish.ip`	"ha_cluster.yml#L211", "#L214", "#L217", "#L220"	Répétez pour chaque serveur Lustre	Adresse IP BMC Redfish pour chaque nœud serveur.
Destinataires des courriels d'alerte	`eseries_lustre_alert_email_list`	"ha_cluster.yml#L233"	À l'échelle du cluster	Destinataires des notifications de ressources et de défaillance HA.
Domaine de messagerie	`eseries_lustre_mail_service_overrides.conf.mydomain`	"ha_cluster.yml#L240"	À l'échelle du cluster	Domaine Postfix utilisé pour envoyer des e-mails d'alerte.

Champ	Variable Ansible	Fichier modèle (`release/1.0.0`)	Portée	Remarques
Sources de temps NTP	`eseries_lustre_time_sync_overrides.server_pools`	"ha_cluster.yml#L250"	À l'échelle du cluster	Sources de temps pour les nœuds du cluster. Spécifiez un pool NTP ou un ou plusieurs serveurs NTP individuels. Incluez la directive requise `pool` ou `server` ainsi que des options telles que `iburst`.
Table de correspondance PCI-nom udev	`eseries_ip_udev_rules`	"ha_cluster.yml#L172"	Répétez selon la disposition des emplacements HCA	Associez chaque emplacement PCI à un nom d'interface persistant (valeurs à L182). Collectez avec `lspci -nnvmm`.
Dépôt de paquets local (facultatif)	`eseries_common_yum_repos`	"ha_cluster.yml#L256", "#L261"	À l'échelle du cluster	Pour les installations hors réseau, décommentez le bloc et définissez le dépôt `url` (L264).
Adresse IP de gestion du serveur	`ansible_host`	"host_vars/lustre_01.yml#L10"	Répétez pour chaque serveur Lustre	Un `host_vars/lustre_NN.yml` fichier par serveur.
Interfaces de backend NVMe-oF	`eseries_nvme_roce_interfaces` (RoCE) / `eseries_nvme_ib_interfaces` (InfiniBand)	"RoCE host_vars/lustre_01.yml#L21", "InfiniBand host_vars/lustre_01.yml#L17"	Répétez pour chaque serveur Lustre	Adresses d'interface backend utilisées pour le trafic de stockage NVMe-oF vers les baies E-Series.
Interfaces de cluster LNet	`eseries_roce_interfaces` (RoCE) / `eseries_ipoib_interfaces` (InfiniBand)	"RoCE host_vars/lustre_01.yml#L47", "InfiniBand host_vars/lustre_01.yml#L38"	Répétez pour chaque serveur Lustre	Adresses d'interface LNet frontales (valeurs RoCE à L56 ; InfiniBand valeurs à L47).
Adresses IP des nœuds Corosync	`eseries_lustre_corosync_node_ips`	"host_vars/lustre_01.yml#L58"	Répétez pour chaque serveur Lustre	Répertoriez toutes les adresses IP des nœuds du cluster HA (valeurs à L62).
NIDs du service cible Lustre	`lustre_target_service_nodes`	"host_vars/lustre_01.yml#L64"	Répétez pour chaque serveur Lustre	Incluez les NID LNet frontaux (`@o2ib`) des deux nœuds de la paire haute disponibilité afin que les services cibles restent accessibles après le basculement (valeurs à L67).

Champ	Variable Ansible	Fichier modèle (`release/1.0.0`)	Portée	Remarques
IP de gestion de baie	`ansible_host`	"host_vars/netapp_01.yml#L13"	Répétez pour chaque baie E-Series	Adresse IP de gestion d'un contrôleur ; l'URL de l'API en découle.
Appartenance au groupe d'éléments de base	`mgs` / `mds_NN` / `oss_NN`	"lustre_inventory.yml"	Par élément de base	Chaque nom d'hôte d'inventaire doit correspondre au nom de base de son fichier `host_vars` correspondant. Par exemple, l'hôte `lustre_01` utilise `host_vars/lustre_01.yml`, et `netapp_01` utilise `host_vars/netapp_01.yml`. Incluez le groupe MGS uniquement dans le premier élément de base. Incluez les groupes MDS dans l'élément de base et dans chaque élément de base d'expansion MDT+OST ; omettez les groupes MDS uniquement dans les éléments de base d'expansion OST uniquement. Enregistrez les cibles d'expansion auprès du MGS existant avec `mgsnode`.

Répétez la configuration pour chaque hôte et élément de base. Créez un `host_vars/lustre_NN.yml` fichier par serveur Lustre et un `host_vars/netapp_NN.yml` fichier par baie, puis répétez les groupes d'inventaire pour chaque élément de base supplémentaire.

Les playbooks de déploiement chargent les identifiants de la baie, de Pacemaker et du BMC depuis le fichier partagé `building_blocks/passwords.yml`. Renseignez ce fichier et chiffrez-le avec Ansible Vault avant le déploiement, ou remplacez les valeurs de manière sécurisée lors de l'exécution avec `--extra-vars`. Ne stockez pas les identifiants en clair dans le système de contrôle de version.

> Les adresses IP, les noms d'hôtes, les noms d'interfaces et le dimensionnement des volumes dans les modèles de déploiement sont des exemples. Modifiez tous les fichiers d'inventaire de l'environnement avant d'exécuter un playbook.

Après avoir terminé

"Déployez le cluster Lustre HA et les clients".

Déployez le cluster Lustre HA et les clients

Préparez le système d'exploitation du serveur, installez NVIDIA DOCA-OFED, déployez

le cluster Lustre HA et configurez les clients Lustre à l'aide des playbooks Ansible NetApp.

Avant de commencer, complétez et sécurisez l'inventaire décrit dans "Personnalisez l'inventaire Ansible de Lustre".

Déployez le cluster Lustre HA

Étapes

1. À partir du `building_blocks/playbooks` répertoire de votre copie de travail, préparez le système d'exploitation sur les serveurs Lustre. Remplacez `<inventory_path>` par le chemin d'accès à votre répertoire d'inventaire personnalisé :

   ```
   ansible-playbook -i <inventory_path>/lustre_inventory.yml
   deploy_lustre_os/deploy_lustre_os_playbook.yml
   ```

2. Installez NVIDIA DOCA-OFED sur chaque serveur Lustre. Compilez les modules du noyau avec le noyau installé à l'étape précédente. Suivez la procédure correspondant à votre noyau dans "Installation et mise à niveau de NVIDIA DOCA-Host". L'exemple suivant illustre une installation de DOCA-OFED sur RHEL ou Rocky Linux.

   a. Installez le paquet du dépôt hôte DOCA et actualisez le cache des paquets. Remplacez `<doca_host_package>` par le paquet hôte DOCA correspondant à votre système d'exploitation et à votre architecture :

      ```
      dnf install -y wget tar
      wget https://www.mellanox.com/downloads/DOCA/<doca_host_package>.rpm
      rpm -i <doca_host_package>.rpm
      dnf clean all
      dnf makecache
      ```

   b. Compilez les modules noyau DOCA pour le noyau en cours d'exécution. Le script affiche le chemin du paquet qu'il crée :

      ```
      dnf install -y doca-extra
      /opt/mellanox/doca/tools/doca-kernel-support
      ```

   c. Installez le paquet du dépôt de modules du noyau créé par le script, puis actualisez le cache des paquets. Remplacez `<doca_kernel_repo>` par le chemin d'accès de l'étape précédente :

      ```
      rpm -Uvh <doca_kernel_repo>.rpm
      dnf makecache
      ```

   d. Installez les paquets DOCA-OFED pour l'espace utilisateur, le noyau et NVMe. Remplacez `<kernel_version>` par la version du noyau en cours d'exécution :

```
dnf install -y doca-ofed-userspace
dnf install -y --disablerepo=doca doca-kernel-<kernel_version>
dnf install -y kmod-mlnx-nvme
```

3. Exécutez le playbook de déploiement principal :

```
ansible-playbook -i <inventory_path>/lustre_inventory.yml
lustre_playbook.yml
```

> Ajustez `--forks`la valeur en fonction de la taille du déploiement pour contrôler le nombre d'hôtes qu'Ansible configure en parallèle. Par exemple, ajoutez `--forks 20`pour un cluster plus important.

Le playbook provisionne les groupes de volumes E-Series ou les pools et volumes DDP, configure la connectivité hôte NVMe-oF, formate et monte les cibles Lustre, configure LNet, applique un réglage des performances et configure les ressources Pacemaker HA.

> Un déploiement composé d'un seul élément de base forme un cluster Pacemaker à deux nœuds. Le nœud A dispose d'un vote supplémentaire en cas de basculement. Pour garantir le quorum dans un cluster à deux nœuds, envisagez de configurer un qdevice. Lorsqu'un déploiement comprend plusieurs éléments de base, chaque élément de base contribue une paire de serveurs à deux nœuds au cluster Pacemaker, jusqu'à la limite documentée du cluster. Vérifiez la configuration du fencing et du quorum dans le "Guide de l'utilisateur d'administration HA" afin que le cluster puisse récupérer les cibles en toute sécurité lors d'une défaillance de nœud.

Déployer des clients Lustre

Déployez des clients Lustre sur n'importe quel système nécessitant un accès au système de fichiers en personnalisant le modèle de déploiement du client et en exécutant le playbook du client.

Étapes

1. Sélectionnez un système d'exploitation client et un noyau compatibles à partir du "Matrice de support Lustre". Compilez et installez les paquets clients Lustre pour ce noyau à partir de "netapp-lustre" s'ils ne sont pas déjà installés.
2. Créez une copie de travail du `clients`répertoire du modèle, y compris son fichier `passwords.yml`partagé, et sélectionnez le modèle qui correspond à votre fabric :
   - **InfiniBand :** "IB_lustre_cluster_clients"
   - **RoCE :** "ROCE_lustre_cluster_clients"
3. Personnalisez `lustre_client_inventory.yml`, `host_vars` et `group_vars` pour vos clients. Les références de ligne dans le tableau pointent vers les fichiers de modèle client RoCE dans la balise `ansible-lustre release/1.0.0` ; le modèle client InfiniBand utilise les mêmes variables, sauf indication contraire.

Champ	Variable Ansible	Fichier modèle (`release/1.0.0`)	Portée	Remarques
Hôtes clients	`lustre_clients`	"lustre_client_inventory.yml#L4"	Par client	Répertoriez chaque nom de l'hôte client.
Utilisateur SSH client et mot de passe become	`ssh_client_user` / `ssh_client_become_pass`	"passwords.yml#L2"	À l'échelle du cluster	Définissez le mot de passe become uniquement si l'utilisateur SSH n'est pas `root`.
Interfaces LNet	`eseries_lustre_lnet`	"group_vars/all.yml#L7"	À l'échelle du cluster	Noms des interfaces LNet côté client (valeurs à L13).
Monter les NIDs MGS	`lustre_client_mounts.mgsnode`	"group_vars/all.yml#L17"	À l'échelle du cluster	NID MGS utilisés pour monter le système de fichiers (valeurs à L20).
Nom du système de fichiers	`lustre_client_mounts.fsname`	"group_vars/all.yml#L21"	À l'échelle du cluster	Doit correspondre au cluster `fsname`.
Point de montage	`lustre_client_mounts.mount_point`	"group_vars/all.yml#L22"	À l'échelle du cluster	Répertoire de montage du client.
IP de gestion du client	`ansible_host`	"host_vars/client_01.yml#L4"	Répéter par client	Un `host_vars/client_NN.yml` fichier par client.
Interfaces de structure de stockage	`eseries_roce_interfaces`	"host_vars/client_01.yml#L10"	Répéter par client	Adresses de l'interface frontale (valeurs à la ligne 15). Le modèle InfiniBand les utilise `eseries_ipoib_interfaces` ici.

Répétez la configuration pour chaque client. Créez un fichier `host_vars/client_NN.yml` pour chaque client et ajoutez l'hôte correspondant à `lustre_client_inventory.yml`. Renseignez le fichier partagé `clients/passwords.yml` et chiffrez-le avec Ansible Vault avant d'exécuter le playbook client.

4. Exécutez le playbook client à partir du répertoire du modèle client :

```
ansible-playbook -i lustre_client_inventory.yml
lustre_client_playbook.yml
```

Le playbook client configure les interfaces réseau du client et LNet, et peut créer une unité de montage systemd persistante pour le système de fichiers Lustre.

Après avoir terminé

"Vérifiez le déploiement".

Vérifiez et étendez le déploiement de Lustre

Confirmez l'état du cluster, les montages Lustre et la capacité du système de fichiers avant la mise en production, puis utilisez le même inventaire pour ajouter des éléments de base en cas de besoin.

Vérifiez le déploiement

Étapes

1. Depuis un serveur Lustre, vérifiez que toutes les ressources Pacemaker sont démarrées sur les nœuds attendus :

```
pcs status
```

2. Depuis un client Lustre, vérifiez que le système de fichiers Lustre est monté :

```
mount -t lustre
```

3. Depuis un client Lustre, vérifiez la capacité du système de fichiers et assurez-vous que les cibles MDT et OST sont visibles :

```
lfs df -h
```

Pour la validation des performances avec IOR, mdtest et fio, voir "Guide de dimensionnement".

Pour la migration contrôlée de la cible ou les tests de basculement de la paire haute disponibilité, suivez les procédures dans le "Guide de l'utilisateur d'administration HA".

Étendre le système de fichiers

Pour augmenter la capacité ou améliorer les performances des métadonnées, étendez votre inventaire existant avec un élément de base supplémentaire. Ne créez pas d'inventaire distinct pour le nouvel élément de base. Le playbook attribue automatiquement des indices MDT et OST uniques et enregistre les nouvelles cibles auprès du MGS existant.

Étapes

1. Ajoutez les nouveaux éléments de base hôtes, baies et groupes de ressources (MDT+OST ou OST uniquement) au même inventaire et aux fichiers `host_vars` et `group_vars` que vous avez utilisés pour le déploiement d'origine.
2. Réexécutez le playbook de déploiement principal sur l'inventaire mis à jour.

Consultez "Architecture de la solution" les règles de mise à l'échelle et "Guide de dimensionnement" les conseils sur le moment d'ajouter chaque type d'élément de base.

Lustre avec le stockage NetApp E-Series - Ressources associées

Utilisez ces liens pour la documentation, les outils et les logiciels associés lors de la planification ou de l'expansion d'un déploiement Lustre avec stockage NetApp E-Series. Pour les étapes de dimensionnement et de déploiement, consultez "Guide de dimensionnement" et "Déployez la solution".

Ressources NetApp

- "Outil NetApp Interoperability Matrix Tool". Recherchez **E-Series Lustre**.
- "NetApp Fiche technique de la baie 100 % flash de la série EF"
- "NetApp Documentation de SANtricity System Manager"
- "collection NetApp ansible-lustre". Pour obtenir des conseils sur le champ d'inventaire, consultez "Personnalisez l'inventaire Ansible de Lustre".
- "NetApp Dépôt source Lustre". NetApp fournit des RPM serveur Lustre précompilés pour Rocky Linux 9.8 et Red Hat Enterprise Linux 9.8. Créez des packages clients Lustre pour le système d'exploitation client et le noyau à partir du code source NetApp.

Communauté Lustre

- "système de fichiers Lustre"
- "Matrice de support Lustre" pour les systèmes d'exploitation clients et les noyaux
- "Lustre Wiki : Réglage Lustre"

Solutions d'infrastructure d'IA NetApp connexes

- "BeeGFS sur NetApp avec stockage série E"
- "Déploiement d'IBM Spectrum Scale avec le stockage NetApp E-Series"
- "NVIDIA DGX SuperPOD avec NetApp EF-Series"