CISCO Connect

AI-Ready Network

Construire et opérer le réseau Datacenter de demain

Mathieu Moriceau
Account Executive Datacenter

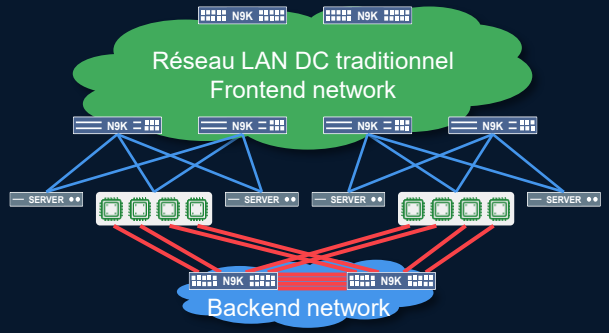
François Couderc
Principal Solutions Engineer

21 mai 2026

Agenda

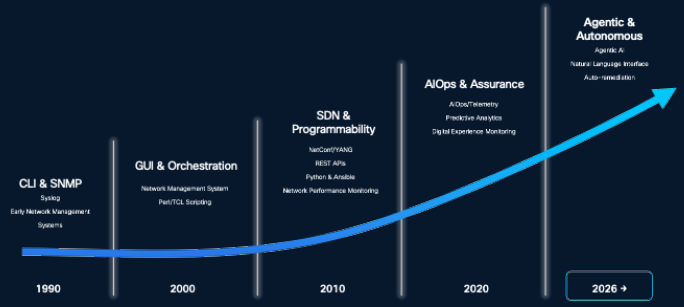
1. Le réseau Datacenter au service de l'IA

- Appréhender les impacts de l'IA sur les réseaux Datacenter
- Connaître les composants clés et leur interactions (switches, GPUs, NICs)
- Exemples de designs et bonnes pratiques



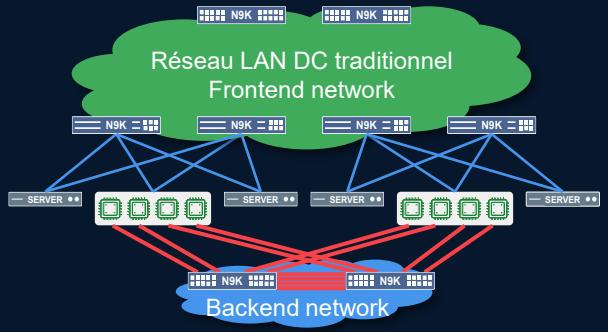
2. L'IA au service du réseau Datacenter

- Passage d'un réseau programmable et ses contrôleurs logiciels à l'ère de l'AgenticOps
- Interactions LLM avec l'infrastructure Datacenter
- Démo

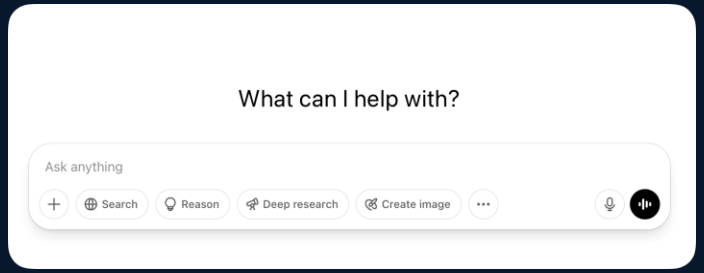


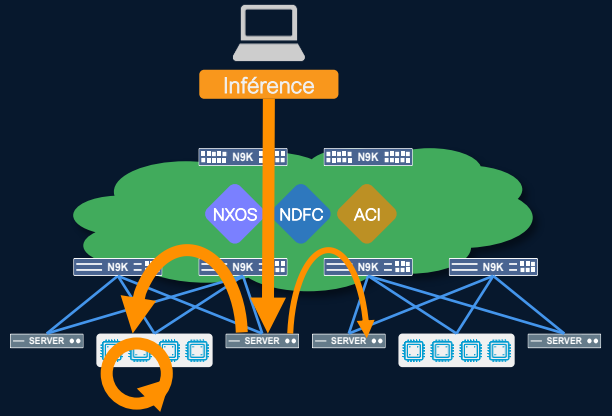
© 2026 Cisco and/or its affiliates. All rights reserved.

Le réseau Datacenter au service de l'IA



Inférence



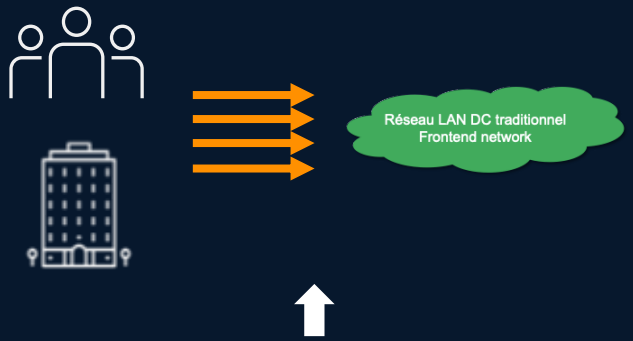


- Une requête utilisateur est traitée localement sur le serveur (intra/inter GPUs / intra node)
- Modèle réseau client / serveur traditionnel. Peu d'impact sur l'infrastructure réseau LAN Datacenter
- La rapidité de la transaction est le principal KPI
- Ne s'applique pas en cas d'inférence distribuée (voir slides suivants)
  - Contexte de très grande taille (137KB VRAM kvcache par token par serveur) =>1To de VRAM (modèle + contexte + cache)
  - Découplage des phases prefill / decode (transformers LLM, VLM, GPUs spécialisées Rubin CPX, ...)
  - Optimisation des ressources GPU via un service désagrégé (ex: Dynamo)

© 2026 Cisco and/or its affiliates. All rights reserved.

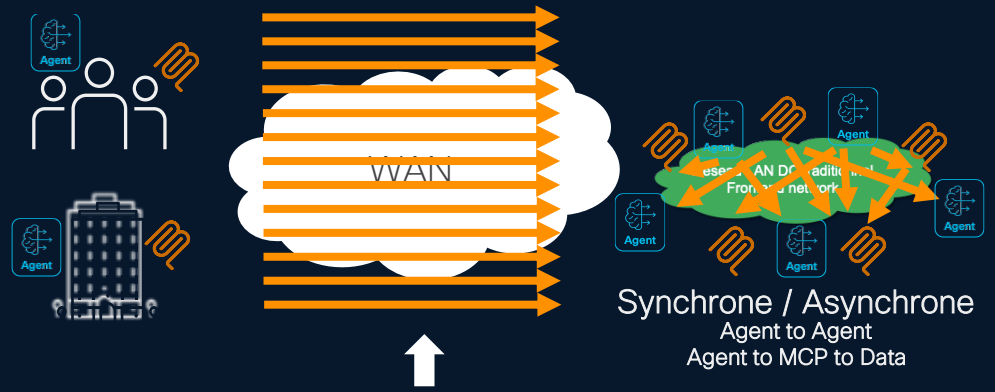
Inférence & impact Agentic AI / serveurs MCP

Trafic réseau traditionnel



Requêtes à “vitesse humaine”

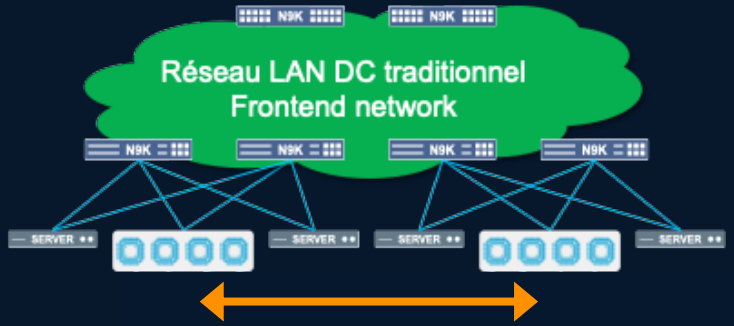
Trafic réseau à l’aire Agentic

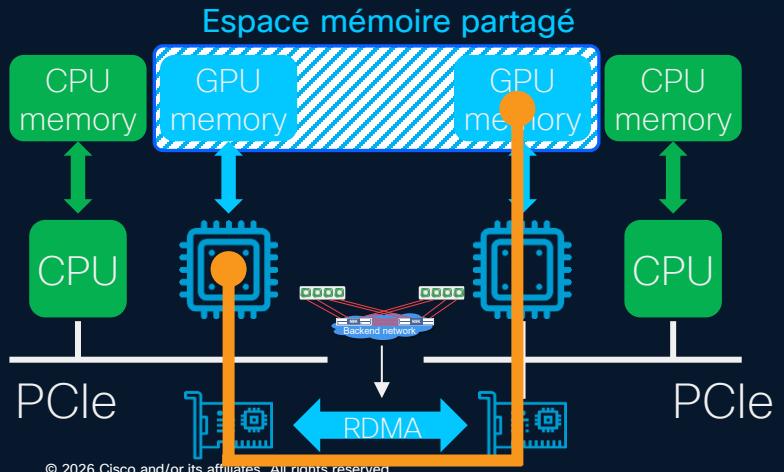


Requêtes à la “vitesse de l’agent”, 24h/24 et 7j/7

- Forte pression sur l’infrastructure WAN & DC Frontend
  - First token latency (FTL) : délai jusqu’à la génération du premier token
  - Time to Last Token (TTL) : délai jusqu’à la génération du dernier token
  - Throughput : nombre de tokens/s ou de requêtes/s traités
- Augmentation des serveurs 100G à l’accès
- Généralisation des Spines 400G
- Augmentation des surfaces d’attaque

Training, Fine Tuning, Inference distribuée





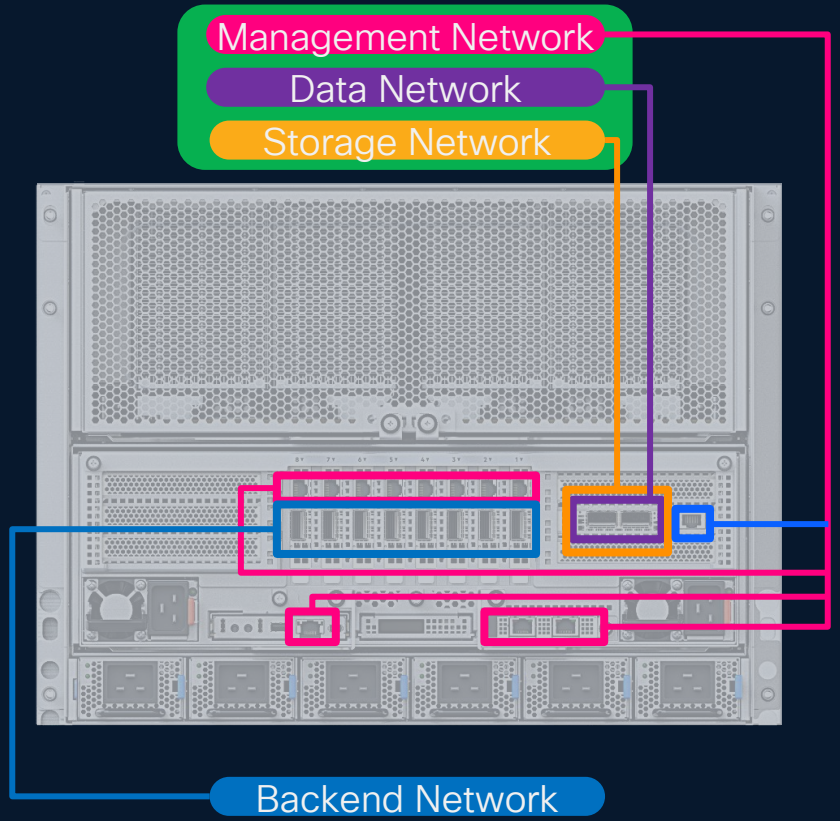
- Echanges mémoire entre les GPUs (gradients, paramètres du modèle)
- Bypass de la CPU pour accès direct avec faible latence

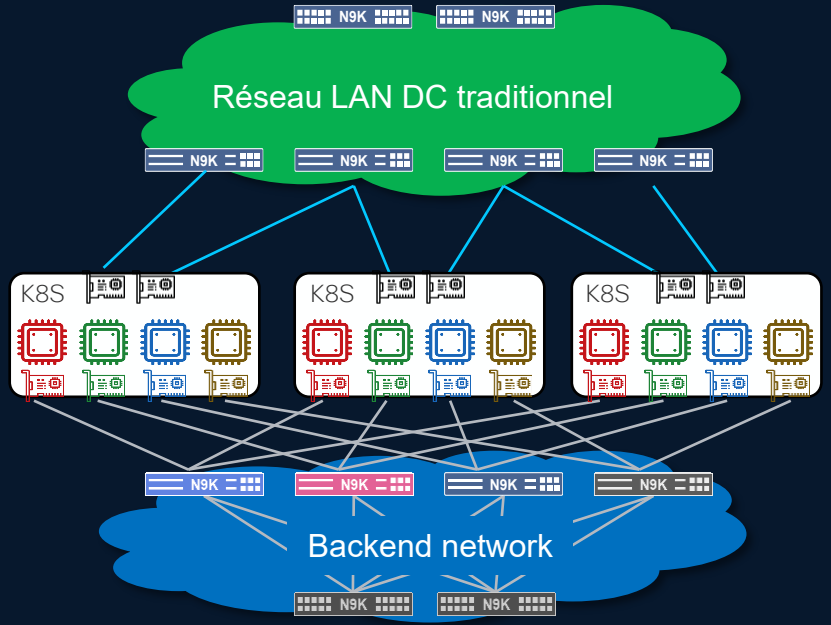
- Les GPUs absorbent aisément 200Gbps / 400Gbps / 800Gbps
- Carte NIC dédiée pour l'inter-GPUs / inter-nodes

- Débits très importants, ultra bursty (On / Off mode)
- Très peu de flows, problème d'entropie pour le partage de charge

© 2026 Cisco and/or its affiliates. All rights reserved.

Réseau Backend Inter-GPUs





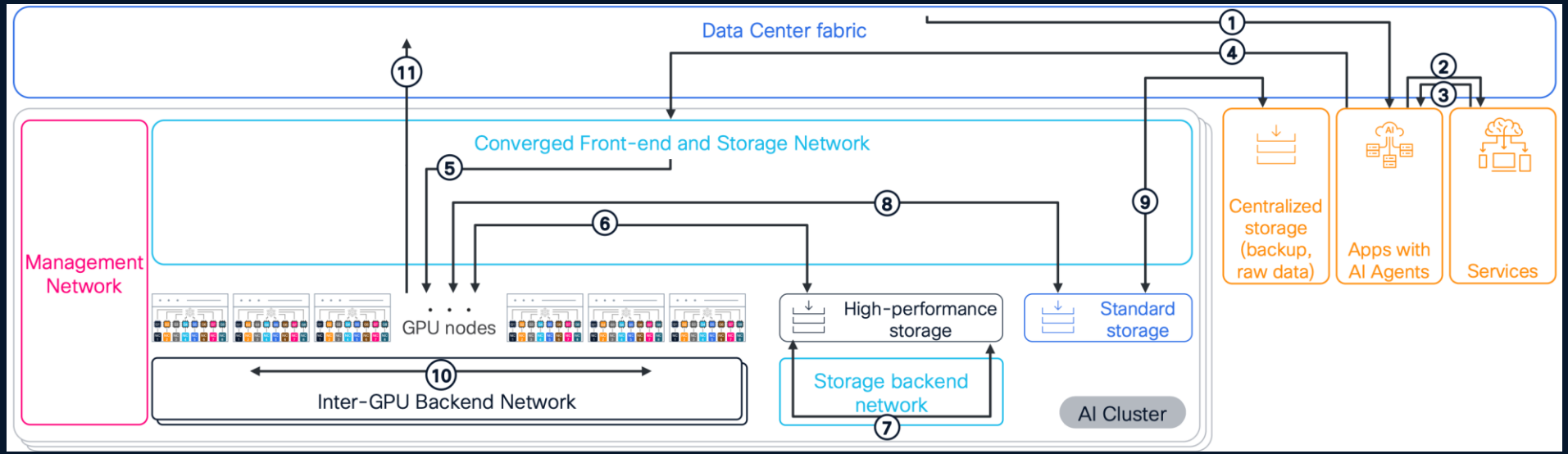
Techno pure L3 ou VXLAN
Taux d'oversubscription 1:1
Pas de redondance réseau

© 2026 Cisco and/or its affiliates. All rights reserved.

De la qualité du réseau dépend le taux d'utilisation des GPUs

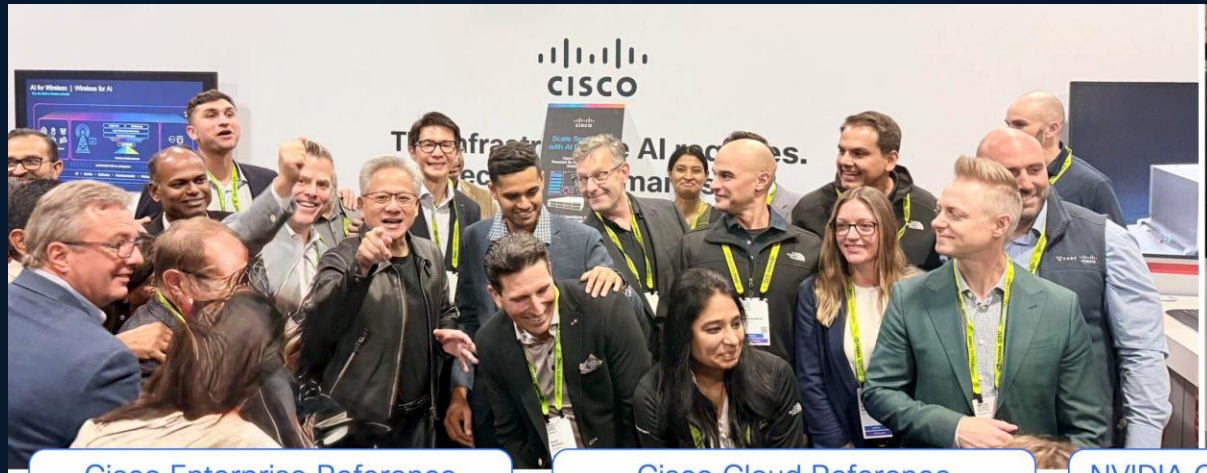
cisco

Exemple d’une architecture à + de 1000 GPUs



© 2026 Cisco and/or its affiliates. All rights reserved.

Architectures de référence Cisco avec NVIDIA



Cisco Enterprise Reference Architecture (ERA)
Up to 1024 GPUs in a cluster

With Cisco Silicon One and Cloud Scale switches and NVIDIA Spectrum-X Ethernet technology

Available since March 2025 (NVIDIA GTC)

Cisco Cloud Reference Architecture (CRA)
1K+ GPUs in a cluster

With Cisco Silicon One switches and NVIDIA Spectrum-X Ethernet technology

Available since Oct 2025 (NVIDIA GTC DC)

NVIDIA Cloud Partner Reference Architecture (NCP RA)
1K+ GPUs in a cluster

With Cisco N9100 switches (powered by NVIDIA Ethernet silicon) and NVIDIA Spectrum-X Ethernet technology

Available since Oct 2025 (NVIDIA GTC DC)

Operational simplicity of Cisco NX-OS, Nexus Dashboard, SONiC, and Nexus Hyperfabric

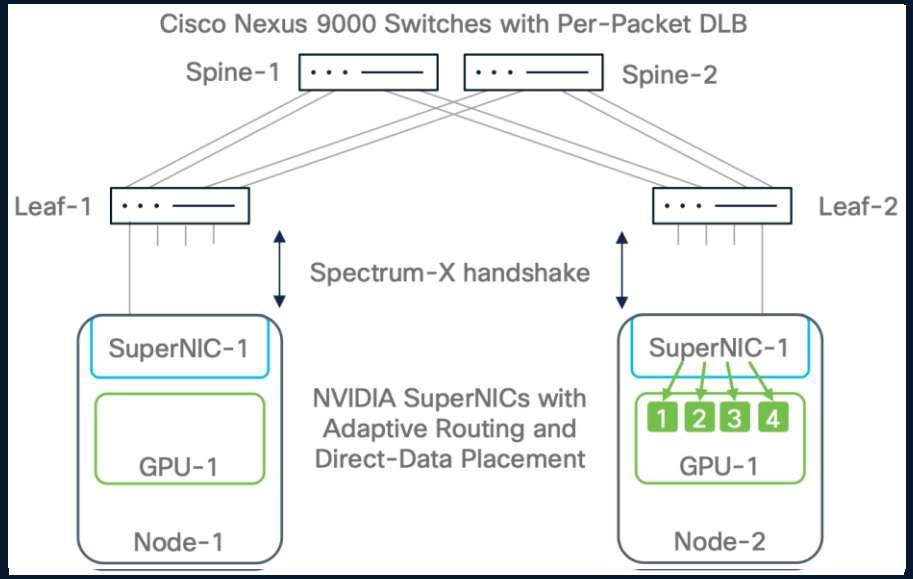
Hyperscalers, Neoclouds, Sovereign clouds, and Enterprises

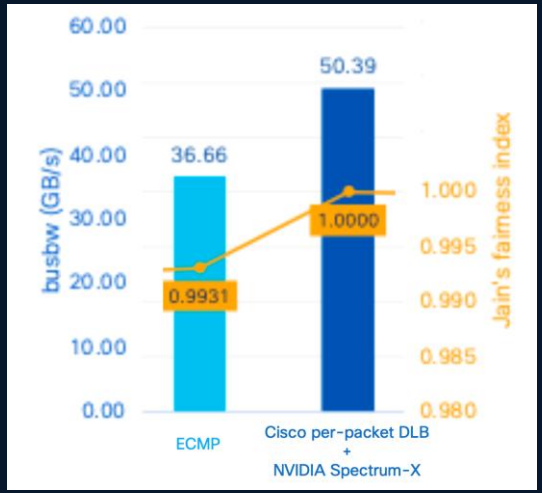
© 2026 Cisco and/or its affiliates. All rights reserved.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Per-Packet DLB + Spectrum-X Adaptive routing

Etape 1





- 37% de gain de performance constaté
- 64 NVIDIA H100 GPUs connectées à un réseau backend Nexus 9000

© 2026 Cisco and/or its affiliates. All rights reserved.

Portfolio Nexus 9000 400G et 800G

Plates-formes Nexus 9300 compatibles NVIDIA Spectrum-X en 10.6(1)F

400G

Nexus 9316D-GX
16p 400G

Nexus 9500
Carte 16p 400G

Nexus 9332D-GX2B
32p 400G

Nexus 9348D-GX2A
48p 400G

800G

Nexus 9400
64p 400G
128p 200G
256p 100G

Nexus 9364D-GX2A
64p 400G

1.6T

Nexus 9808
288p 400G
800G ready

Nexus 9164E

NVIDIA Spectrum-X Ethernet silicon

Nexus 9364E
64p 800G
128p 400G

Dual duplex LC 400G Transceiver

2H CY2026

Nexus 9364F
64p 1.6T
128p 400G

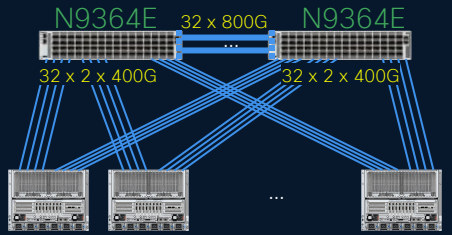
© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

Exemple de Backend avec cisco + NVIDIA

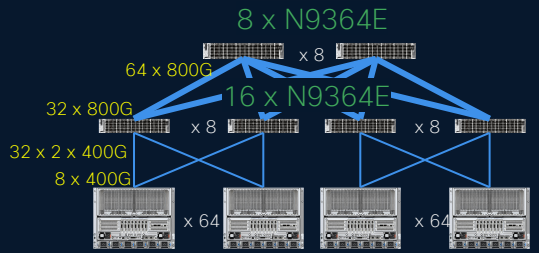
Alignement de nos Reference Architecture sur ceux de NVIDIA pour les besoins Entreprise

Design 128 GPUs



16 serveurs – 8 GPUs – 8 NICs 400G

Design 1024 GPUs



128 serveurs – 8 GPUs – 8 NICs 400G

Nexus Dashboard

Provisioning & Operations dédiés AI Networking

Intégration dans nos switches des fonctionnalités NVIDIA Spectrum-X à valeur ajoutée (ex: Adaptive Routing)

Continuer à utiliser la Gamme Nexus 9000 tout en étant dans une architecture de référence qualifiée de bout en bout à la fois par NVIDIA et Cisco

© 2026 Cisco and/or its affiliates. All rights reserved.

cisco

AIOps et Agentic AI:
L'IA au service du réseau Datacenter

AIOps - Intelligence Artificielle pour les Opérations IT

Piliers fonctionnels

Détection d'anomalies (Anomaly Detection)

- Le système apprend le comportement "normal" (baseline dynamique) et alerte uniquement sur les vraies déviances – pas juste sur des seuils statiques.

Root Cause Analysis (RCA) automatique

- Corrélation d'événements multi-couches pour identifier la cause réelle, pas juste les symptômes.

Analyse prédictive

- Anticiper les pannes avant qu'elles surviennent (saturation de liens, instabilité BGP, épuisement de ressources TCAM).

Recommandations actionnables

- Pas seulement "il y a un problème" mais "voici quoi faire, et voici le risque si tu ne fais rien".

Automatisation de la remédiation

- Déclenchement automatique ou semi-automatique de workflows correctifs.

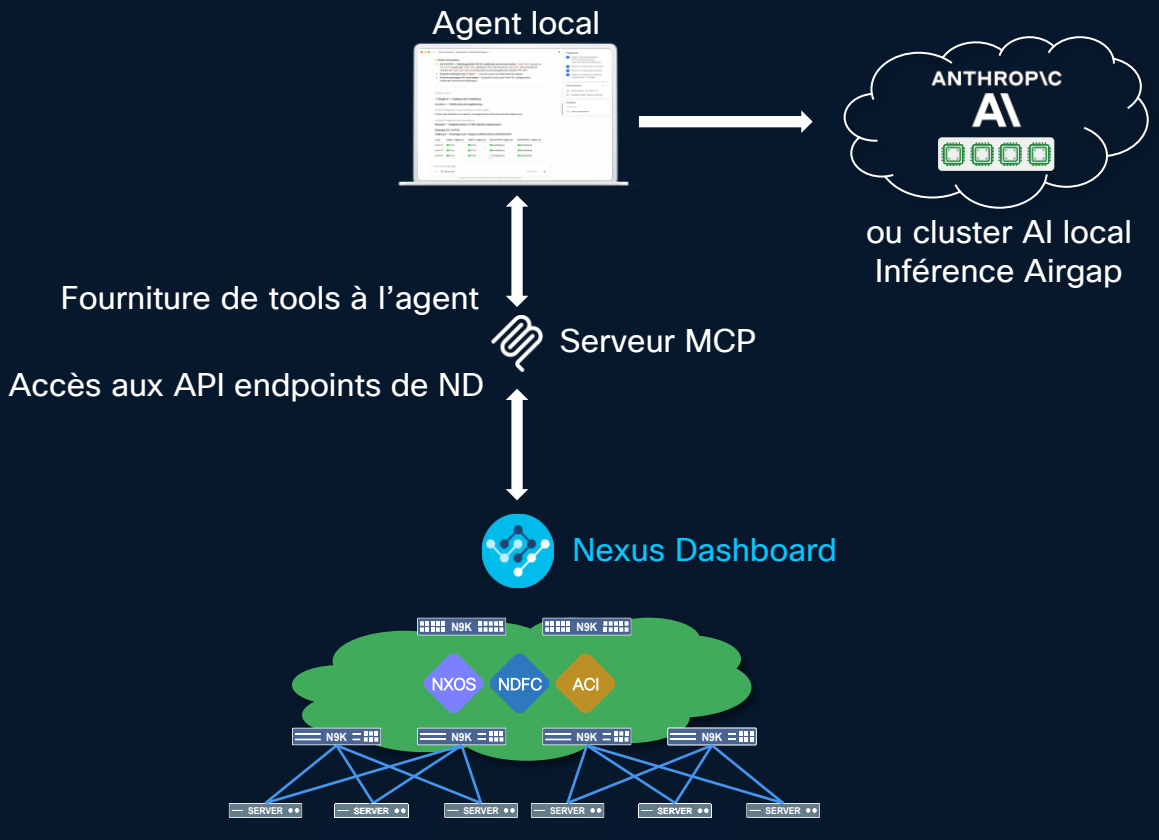
Attentes & Mesure des Bénéfices

ROI concret

Métrique	Sans AIOps	Avec AIOps	Gain
MTTD (Mean Time to Detect)	15-60 min	1-5 min	~90%
MTTR (Mean Time to Resolve)	2-8h	20-45 min	~75%
Faux positifs alertes	60-80% du volume	< 15%	Réduction du bruit énorme
Pannes évitées (prédictif)	0	30-50% des incidents	ROI direct
Temps ingénieur par incident P1	4-6h	30-60 min	Libère ~5h/incident
Audit de conformité	Manuel, jours	Continu, automatique	Conformité permanente

© 2026 Cisco and/or its affiliates. All rights reserved.

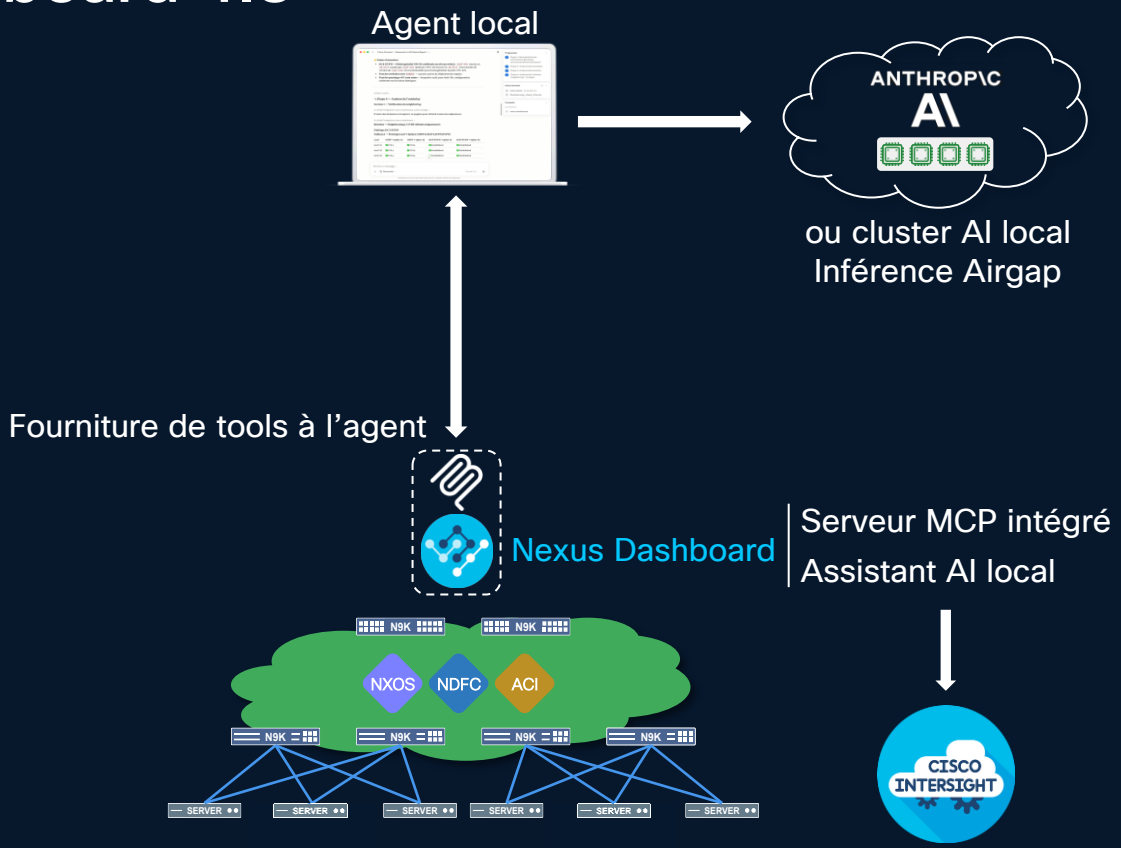
Démo



© 2026 Cisco and/or its affiliates. All rights reserved.

Nexus Dashboard 4.3

Roadmap – juillet 2026

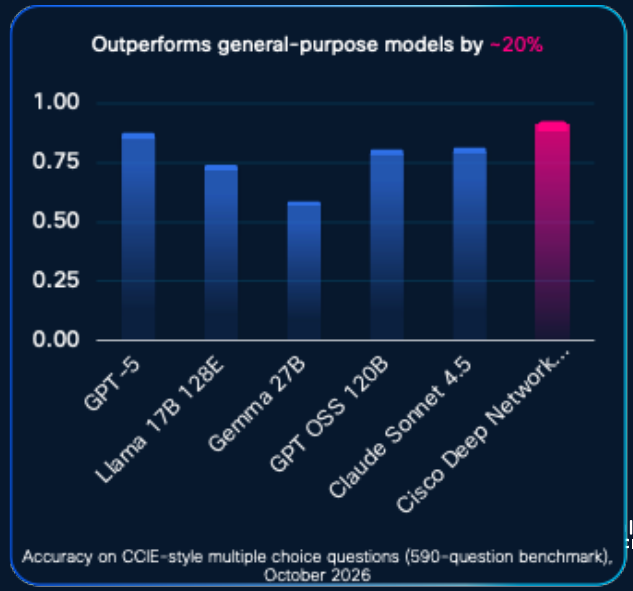


© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco AI Canvas donne vie à l'AgenticOps

Lancement été 2026 en modèle Cloud

1. Multi-technos / Multi-architectures
2. Interaction Homme-machine native
3. Cisco Deep Network Model



© 2026 Cisco and/or its affiliates. All rights reserved.

Conclusion

Conclusion

Evaluer la nécessité d'avoir ou non un Backend GPUs sur son infra LAN Datacenter

- Le backend GPUs devient un système où le software, les NICs et les switches forment une entité unifiée.
- Cisco dispose des atouts pour une mise en œuvre rapide et performante :
  - Des switches 400G/800G hautes performances et faible latence
  - Un partenariat unique avec NVIDIA
  - Un modèle opérationnel éprouvé (Ethernet/IP, NX-OS, stack de management)

Les solutions AIOps permettent de fluidifier les opérations et d'accélérer le troubleshooting

- Importance de Nexus Dashboard qui consolide la télémétrie de l'infrastructure
- Permettre aux experts de se concentrer sur des tâches à valeur ajoutée

© 2026 Cisco and/or its affiliates. All rights reserved.


CISCO