FR

GTC 2026 — Billet B4

Vera Rubin & Feynman

Serie GTC 2026 · Analyse BOTUM · Mars 2026

Mars 2026

Sommaire

Introduction — Quand le materiel devient strategie

Au GTC 2026, Jensen Huang n'a pas presente que des logiciels. Il a pose les fondations physiques de l'ere agentique. Vera Rubin, Vera Rubin Ultra, NVLink Fusion, Feynman : quatre annonces qui redessinentla carte de l'infrastructure GPU pour les prochaines annees.

Pour les architectes solutions enterprise, ces annonces ne sont pas de la speculation : elles determinent les choix d'infrastructure a faire maintenant pour etre pret en 2026-2028.

---

1. Vera Rubin — La plateforme rack-scale de nouvelle generation

Vera Rubin est l'architecture GPU qui succed au Blackwell. Elle n'est pas qu'un upgrade de performance : c'est un changement de paradigme dans la facon de concevoir l'infrastructure IA.

Specifications cles

Spec	Vera Rubin (GB200 NVL72)
GPU par rack	576 GPU
Puissance FP4	1,5 exaflops par rack
Memoire HBM	30 TB par rack (HBM3e)
NVLink	5e generation, 1,8 Tb/s bidirectionnel
Refroidissement	Liquide obligatoire (architecture rack-native)
Disponibilite	H2 2026 (deja en production chez hyperscalers)

Ce qui change pour l'enterprise

- 1,5 exaflops par rack = un modele frontier entier peut etre entraine en heures, pas en semaines
- Architecture rack-native : fini les serveurs GPU individuels — l'unite de base est le rack
- Memoire unifiee : 30 TB HBM3e par rack permet des contextes de plusieurs millions de tokens
- Refroidissement liquide obligatoire : toute decision d'infrastructure doit integrer le liquide des maintenant

i Signal BOTUM : les entreprises qui planifient un data center en 2025-2026 doivent prevoir le refroidissement liquide. Retrofitter plus tard coutera 3-5x plus cher.

---

2. Vera Rubin Ultra — Le saut quantique de juin 2026

Vera Rubin Ultra est la version maximaliste de Vera Rubin. Disponible en juin 2026, elle double les capacites du Vera Rubin standard avec deux GPU GB300 connectes via NVLink.

- 2x GB300 : double la memoire et la bande passante par rapport au GB200
- Inference de modeles > 1T parametres sans sharding : un seul rack peut faire tourner GPT-4 class
- Latence reduite de 40% sur les workloads d'inference agentique multi-tour
- Priorite aux hyperscalers et clients strategiques : liste d'attente deja importante

i Pour les entreprises : si vous planifiez un projet IA H2 2026 ou 2027, Vera Rubin Ultra est l'infrastructure cible. Commencez les discussions avec votre fournisseur cloud maintenant.

3. NVLink Fusion — L'ouverture strategique

NVLink Fusion est l'annonce la plus strategique pour l'ecosysteme enterprise. NVIDIA ouvre NVLink aux CPU tiers : AMD EPYC, Intel Xeon, et processeurs Arm.

Pourquoi c'est un changement majeur

Jusqu'a present, NVLink etait exclusif a l'ecosysteme NVIDIA (GPU + CPU Grace). NVLink Fusion signifie :

- Les entreprises AMD-first peuvent maintenant coupler des GPU NVIDIA sans changer leur CPU
- Les systemes hybrides CPU Intel + GPU NVIDIA beneficient de la bande passante NVLink (vs PCIe 5x plus lent)
- Les architectures Arm (AWS Graviton, NVIDIA Grace) deviennent des cibles legitimes pour les workloads IA
- Reduction des couts : pas besoin de remplacer l'ensemble du parc CPU pour migrer vers GPU NVIDIA

Scenario	Avant NVLink Fusion	Apres NVLink Fusion
AMD EPYC + GPU NVIDIA	PCIe seulement, latence elevee	NVLink natif, 5x plus rapide
Intel Xeon + GPU NVIDIA	PCIe Gen5, goulot d'etranglement	NVLink Fusion, bande passante HPC
Arm + GPU NVIDIA	Ecosysteme ferme (Grace only)	Ouvert : AWS Graviton, Ampere...
Migration HPC -> IA	Remplacement complet du parc	GPU overlay sur parc existant

4. Feynman 2028 — La vision long terme

Jensen Huang a annonce Feynman au GTC 2026 — le successeur de Blackwell prevu pour 2028. Ce n'est pas encore une spec produit, mais une declaration d'intention.

Ce qu'on sait

- Nom de code : Feynman — nomme apres Richard Feynman, physicien quantique

- Generation : post-Vera Rubin, soit 2 generations apres Blackwell
- Annonce 2028 : NVIDIA maintient son rythme annuel d'annonces d'architectures
- Focus presume : inference quantique-classique hybride, selon les signaux du marche

Implication pour la planification enterprise

L'annonce de Feynman a 2 ans confirme la cadence de NVIDIA : une nouvelle architecture majeure chaque 12-18 mois. Pour les DSI et architectes enterprise, cela signifie :

- Infrastructure as a service > on-premise pour rester agile face aux cycles d'obsolescence
- Contrats cloud flexibles : eviter les lock-in a 5 ans sur du materiel GPU specifique
- FinOps GPU : modeliser le cout total de possession en incluant l'obsolescence rapide

i La regle des 3 ans : un GPU achete aujourd'hui est genere-precedente dans 18 mois. Calculez votre ROI sur 24 mois maximum pour le materiel on-premise.

---

5. Guide de decision architecte — Cloud ou on-prem ?

Face a ces annonces, la question que posent tous les DSI est la meme : "On investit dans du Vera Rubin on-premise, ou on reste sur le cloud ?". Voici le framework de decision BOTUM :

Critere	Cloud (recommande si...)	On-premise (recommande si...)
Volume workload	< 10 000 req GPU/jour	> 50 000 req GPU/jour
Conformite donnees	Pas de contrainte forte	RGPD, secteur sante, defense
Equipe GPU Ops	Equipe < 3 personnes	Equipe GPU Ops dediee
Horizon ROI	< 24 mois	> 36 mois avec usage intensif
Refroidissement	Infrastructure existante	Liquide disponible ou planifie
Budget capex	Preference opex	Budget capex > 2M$

La recommandation BOTUM pour 2026

- Phase pilote (0-12 mois) : cloud systematiquement — AWS, Azure, GCP ont deja acces a Vera Rubin
- Phase scale (12-36 mois) : hybrid — cloud pour les pics, colocation GPU pour la base
- Phase mature (36+ mois) : on-prem si volume justifie + equipe GPU Ops formee

Besoin d'aide pour evaluer votre infrastructure GPU ?
Les equipes BOTUM accompagnent les DSI dans leur strategie infrastructure IA. De l'evaluation cloud vs on-prem au deploiement Vera Rubin, nous couvrons tout le spectre.
Contact : www.botum.ca/contact

-> Version en ligne : blog.botum.ca/gtc2026-b4-vera-rubin-feynman-infrastructure-gpu/