EN

# GTC 2026 — Article B4

## Vera Rubin & Feynman

GTC 2026 Series · BOTUM Analysis · March 2026

March 2026

## Table of Contents

## Introduction — When hardware becomes strategy

At GTC 2026, Jensen Huang did not just present software. He laid the physical foundations of the agentic era. Vera Rubin, Vera Rubin Ultra, NVLink Fusion, Feynman: four announcements that redraw the GPU infrastructure map for the coming years.

For enterprise solution architects, these announcements are not speculation: they determine the infrastructure choices to make right now to be ready for 2026-2028.

---

## 1. Vera Rubin — Next-gen rack-scale GPU platform

Vera Rubin is the GPU architecture succeeding Blackwell. It is not just a performance upgrade: it is a paradigm shift in how AI infrastructure is designed.

### Key specifications

| Spec | Vera Rubin (GB200 NVL72) |
|---|---|
| GPUs per rack | 576 GPUs |
| FP4 compute | 1.5 exaflops per rack |
| HBM memory | 30 TB per rack (HBM3e) |
| NVLink | 5th gen, 1.8 Tb/s bidirectional |
| Cooling | Liquid required (rack-native architecture) |
| Availability | H2 2026 (hyperscalers already in production) |

### What changes for enterprise

- 1.5 exaflops per rack = an entire frontier model can be trained in hours, not weeks
- Rack-native architecture: individual GPU servers are gone - the rack is the base unit
- Unified memory: 30 TB HBM3e per rack enables multi-million token contexts
- Liquid cooling required: every infrastructure decision must factor in liquid cooling now

i BOTUM signal: companies planning a data center in 2025-2026 must plan liquid cooling. Retrofitting later will cost 3-5x more.

---

## 2. Vera Rubin Ultra — The quantum leap of June 2026

Vera Rubin Ultra is the maximalist version of Vera Rubin. Available in June 2026, it doubles the capabilities of standard Vera Rubin with two GB300 GPUs connected via NVLink.

- 2x GB300: doubles memory and bandwidth vs GB200
- Inference of models > 1T parameters without sharding: one rack can run GPT-4 class models

- 40% lower latency on multi-turn agentic inference workloads
- Priority for hyperscalers and strategic clients: waitlist already significant

i For enterprise: if you are planning an AI project for H2 2026 or 2027, Vera Rubin Ultra is the target infrastructure. Start discussions with your cloud provider now.

---

## 3. NVLink Fusion — The strategic opening

NVLink Fusion is the most strategically significant announcement for the enterprise ecosystem. NVIDIA opens NVLink to third-party CPUs: AMD EPYC, Intel Xeon, and Arm processors.

### Why this is a major shift

Until now, NVLink was exclusive to the NVIDIA ecosystem (GPU + Grace CPU). NVLink Fusion means:

- AMD-first enterprises can now pair NVIDIA GPUs without changing their CPU
- Intel Xeon + NVIDIA GPU hybrid systems benefit from NVLink bandwidth (vs PCIe 5x slower)
- Arm architectures (AWS Graviton, NVIDIA Grace) become legitimate targets for AI workloads
- Cost reduction: no need to replace the entire CPU fleet to migrate to NVIDIA GPU

| Scenario | Before NVLink Fusion | After NVLink Fusion |
|---|---|---|
| AMD EPYC + NVIDIA GPU | PCIe only, high latency | Native NVLink, 5x faster |
| Intel Xeon + NVIDIA GPU | PCIe Gen5 bottleneck | NVLink Fusion, HPC bandwidth |
| Arm + NVIDIA GPU | Closed ecosystem (Grace only) | Open: AWS Graviton, Ampere... |
| HPC -> AI migration | Full fleet replacement | GPU overlay on existing fleet |

---

## 4. Feynman 2028 — NVIDIAs long-term vision

Jensen Huang announced Feynman at GTC 2026 - the Blackwell successor planned for 2028. This is not yet a product spec, but a declaration of intent.

### What we know

- Code name: Feynman - named after physicist Richard Feynman
- Generation: post-Vera Rubin, 2 generations after Blackwell
- 2028 target: NVIDIA maintains its annual architecture announcement cadence
- Presumed focus: quantum-classical hybrid inference, based on market signals

### Enterprise planning implications

The 2-year Feynman announcement confirms NVIDIAs cadence: one major architecture every 12-18 months. For CIOs and enterprise architects, this means:

- Infrastructure as a service > on-premise to stay agile against obsolescence cycles
- Flexible cloud contracts: avoid 5-year lock-in on specific GPU hardware
- GPU FinOps: model total cost of ownership including rapid obsolescence

i The 3-year rule: a GPU purchased today is last-gen in 18 months. Calculate your ROI on a maximum of 24 months for on-premise hardware.

---

## 5. Architect decision guide — Cloud or on-prem?

Faced with these announcements, every CIO asks the same question: "Do we invest in on-prem Vera Rubin, or stay on cloud?" Here is the BOTUM decision framework:

| Criteria | Cloud (recommended if...) | On-prem (recommended if...) |
|---|---|---|
| Workload volume | < 10,000 GPU req/day | > 50,000 GPU req/day |
| Data compliance | No strong constraint | GDPR, healthcare, defense |
| GPU Ops team | Team < 3 people | Dedicated GPU Ops team |
| ROI horizon | < 24 months | > 36 months with intensive use |
| Cooling | Existing infrastructure | Liquid available or planned |
| Capex budget | Opex preference | Capex budget > $2M |

### BOTUM recommendation for 2026

- Pilot phase (0-12 months): cloud systematically - AWS, Azure, GCP already have Vera Rubin access
- Scale phase (12-36 months): hybrid - cloud for peaks, GPU colocation for baseline
- Mature phase (36+ months): on-prem if volume justifies + trained GPU Ops team

Need help evaluating your GPU infrastructure?
BOTUM teams support CIOs in their AI infrastructure strategy. From cloud vs on-prem evaluation to Vera Rubin deployment, we cover the full spectrum. Contact:
www.botum.ca/contact

-> Online version: blog.botum.ca/gtc2026-b4-vera-rubin-feynman-gpu-infrastructure/