Від стійки до POD:

NVIDIA GB300 NVL72 і платформа Vera Rubin

Потреби штучного інтелекту в обчислювальній інфраструктурі невблаганно зростають. NVIDIA має відповідь на ці виклики.

Індустрія розроблення штучного інтелекту потребує величезних обчислювальних ресурсів для інференсу і ще більших — для навчання передових мовних моделей. Попри те, що нові високопродуктивні рішення виходять регулярно, готових систем з інтерконектом, здатним ефективно об'єднати кілька десятків прискорювачів, на ринку майже немає. NVIDIA закриває цю потребу стійковими комплексами, які працюють як єдина система: **GB300 NVL72** сьогодні і **Vera Rubin** — у найближчій перспективі.

Це готові рішення «під ключ», укомплектовані GPU, процесорами ARM, комутаторами, DPU і SmartNIC. Розберімо, що вони собою являють, з чого складаються, наскільки продуктивні — і, головне, чого вимагають від інженерної інфраструктури будівлі, у якій їх доведеться розміщувати.

Що таке GB300 NVL72

GB300 NVL72 — це серверна шафа (**рис. 1**) з 36 центральних процесорів NVIDIA Grace і 72 графічних процесорів B300 на архітектурі Blackwell Ultra. Комплекс призначений для навчання та інференсу сучасних великих мовних моделей із трильйонами параметрів і для інтеграції в інфраструктуру великих ЦОД та сектору ШІ.

Побудовано систему на фірмовій модульній архітектурі MGX, яка забезпечує гнучкість конфігурування та широкі можливості масштабування. Сукупний обсяг швидкої пам'яті HBM3e становить близько 20 ТБ — це на 50 % більше, ніж у попередника, і дозволяє розгортати найвимогливіші моделі без потреби в агресивній квантизації. З урахуванням пам'яті LPDDR5X процесорів Grace обсяг швидкої пам'яті сягає 37 ТБ, а сумарна кількість ядер Arm Neoverse V2 — 2592.

Така кількість високопродуктивних компонентів робить систему надзвичайно енергоємною: споживання шафи сягає 142 кВт. Жодні серверні вентилятори для охолодження такого пристрою не підходять, тому застосовується пряме рідинне охолодження. Крім того, NVIDIA інтегрувала Transformer Engine другого покоління з підтримкою форматів FP4 і FP8: вони дозволяють зберегти достатню точність при істотній економії обчислювальних ресурсів, не вдаючись до INT8.

Рис. 1. Компонування стійки GB300 NVL72 на прикладі реалізації Supermicro

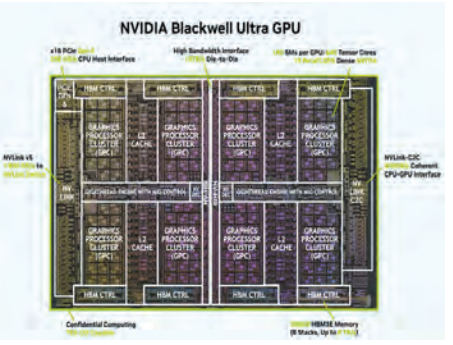

Рис. 2. Структура графічного процесора NVIDIA Blackwell Ultra (B300)

Ключова відмінність Blackwell Ultra (**рис. 2**) від звичайного Blackwell — не в кількості, а в профілі. NVIDIA заявляє у півтора раза більше щільних FP4-обчислень і приблизно вдвічі вищу продуктивність механізму уваги. Саме останній стає вузьким місцем на моделях із міркуванням і довгим контекстом, тож приріст спрямований точно в поточний профіль навантаження.

Обчислювальний лоток: базовий будівельний блок

Рис. 3. Обчислювальний лоток NVIDIA GB300 NVL72

Якщо в попередньому поколінні одиницею комплектації був суперчип із двох GPU й одного CPU, то

в GB300 NVL72 практичною одиницею став обчислювальний лоток. Один лоток (**рис. 3**) містить чотири GPU Blackwell Ultra і два процесори Grace, зв'язані інтерконектом NVLink-C2C, який забезпечує мінімальну затримку при передачі даних між графічними й центральними процесорами.

У кожному лотку GB300:
- **4 GPU B300** — сукупно 1152 ГБ HBM3e (по 288 ГБ на процесор); кожен процесор дає 15 PFLOPS щільних обчислень NVFP4, звідки й береться показник 1080 PFLOPS на шафу;
- **2 CPU Grace** — близько 1 ТБ LPDDR5X сукупної системної пам'яті;
- **дві мезонінні мережеві плати** (**рис. 4**) з двома кристалами ConnectX-8 на кожній — разом чотири адаптери по 800 Гбіт/с, тобто один SuperNIC на один GPU;
- **один DPU BlueField-3 B3240** з двома портами по 400 Гбіт/с;
- **накопичувач M.2 NVMe** під операційну систему і диски E1.S NVMe як швидкий локальний кеш.

Вісімнадцять таких лотків формують шафу. Розподіл ролей між мережевими адаптерами тут принциповий: ConnectX-8 обслуговує трафік «схід–захід» між графічними процесорами, BlueField-3 — трафік «північ–південь» до сховища і зовнішніх систем. Спроба звести обидва потоки на одну фабрику заради економії ламає розрахункову модель усієї архітектури.

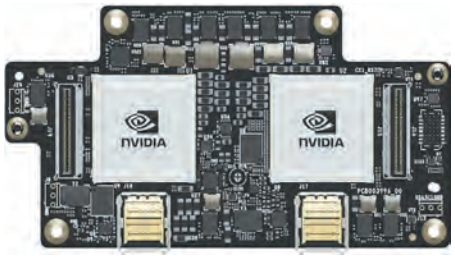

Рис. 4. Мезонінна мережева плата NVIDIA з двома кристалами ConnectX-8

Інфраструктура NVLink

NVLink п'ятого покоління

NVLink 5 — основна сполучна ланка всього комплексу. Технологія об'єднує графічні й центральні процесори в один великий кластер із мінімальними затримками. Інтерфейс має пропускну здатність 1,8 ТБ/с на GPU — це сумарна двонаправлена цифра, тобто по 900 ГБ/с у кожному напрямку. Саме вона дозволяє створювати міжз'єднання на 72 графічних процесори. Для реалізації такої розгалуженої мережі NVIDIA впровадила у стійку нетривіальні компоненти — NVLink Switch і мідну об'єднавчу панель, завдяки яким NVLink масштабується значно ширше, ніж у традиційних системах DGX і HGX.

NVLink Switch

Спеціалізовані комутаційні лотки NVLink встановлюються в центрі шафи між масивами обчислювальних вузлів у кількості дев'яти штук, по два кристали NVSwitch у кожному. Кожен GPU має вісімнадцять лінків п'ятого покоління — по одному на кожен внутрішньостійковий NVSwitch. Так формується повнозв'язний домен першого рівня на 72 GPU із сукупною неблокуючою пропускною здатністю 130 ТБ/с, за рахунок чого швидкість передачі залишається стабільною навіть на пікових навантаженнях.

Арифметика тут зійшлася точно. Кожен комутаційний лоток несе 144 порти (по 72 на кристал), дев'ять лотків дають 1296 портів — рівно стільки, скільки потребують 72 графічних процесори по вісімнадцять лінків кожен. Жодного порту в надлишку, жодного в нестачі.

Саме завдяки роботі NVLink Switch вдається комутувати таку велику кількість графічних процесорів у єдину мережу — без цих вузлів технологія NVLink могла б об'єднати лише кілька GPU. Це і зробило можливою концепцію об'єднання всіх графічних процесорів в один великий CUDA-процесор зі спільною пам'яттю.

Важлива для експлуатації деталь: NVLink Switch — це керований комутатор під управлінням мережевої ОС NVOS, а не пасивний бекплейн. Він є повноцінним об'єктом моніторингу. Керування розділами домену (NVLink Partition) виконується через API глобального менеджера фабрики GFM: розділи ізолюють доступ до пам'яті GPU між завданнями й захищають одне завдання від збоїв у сусідньому. Для корпоративного мультикористувацького розгортання цей механізм треба закладати в проєкт експлуатації одразу.

NVLink Spine: мідна об'єднавча панель

Фізично домен NVLink реалізовано не оптикою, а мідною об'єднавчою панеллю — так званим NVLink Spine. Це касети з майже п'ятьма тисячами пасивних мідних провідників сумарною довжиною понад 3,2 км, змонтовані в задній частині шафи. Обчислювальні й комутаційні лотки під'єднуються до них наосліп (blind-mate): лоток, який засувають у стійку, автоматично стикується у панель через високоточні плаваючі роз'єми, без ручної комутації.

Це свідомий інженерний вибір, і Дженсен Хуанг свого часу назвав його ціну прямо: якби ті самі з'єднання виконали оптикою, лише трансивери й ретаймери для живлення NVLink Spine з'їли б близько 20 кВт на шафу. Мідь віддає ці 20 кіловат обчисленням і водночас прибирає з переліку відмов найчисленніший тип компонентів. Оптика починається там, де закінчується шафа.

У GB300 конструкцію збережено, але довжину провідників збільшено приблизно наполовину — під вищі вимоги до смуги. А от у поколінні Rubin NVIDIA від касет із кабелями відмовляється на користь друкованої об'єднавчої плати на всю ширину шафи: менше ручної праці на складанні й менше механічних відмов на роз'ємах.

Мережева інфраструктура: InfiniBand та Ethernet

NVIDIA грамотно розставила пріоритети, приділивши мережевій інфраструктурі не менше уваги, ніж графічним рішенням. Щоб забезпечити ефективний інтерконект для різних екосистем, пропонується вибір із двох платформ — **Quantum-X800 InfiniBand** і **Spectrum-X Ethernet**.

Рішення на базі InfiniBand по праву вважаються найбільш просунутими в галузі кластеризації, оскільки цей стандарт із самого початку розроблявся для побудови таких систем. Він забезпечує високу пропускну здатність і низькі затримки, що ідеально підходить для сучасних обчислювальних навантажень.

Однак інфраструктура InfiniBand не завжди легко впроваджується в ЦОД, які раніше використовували класичний Ethernet. Саме для таких замовників NVIDIA розробила платформу Spectrum-X, яка дозволяє отримати більшість переваг InfiniBand без повної перебудови мережевої інфраструктури.

Quantum-X800 InfiniBand

Платформа Quantum-X800 InfiniBand (**рис. 5**) побудована на новій версії стандарту InfiniBand XDR і призначена для інтеграції в найвимогливіші інфраструктури HPC та ШІ. Комутатори мають до 144 портів OSFP 800G, а також виділений порт для Unified Fabric Manager. Сімейство працює у зв'язці з мережевими адаптерами ConnectX-8 SuperNIC, вбудованими в кожен обчислювальний лоток.

Рис. 5. Мережевий комутатор Quantum InfiniBand у компонуванні 1U

Spectrum-X Ethernet

Платформа Spectrum-X Ethernet (**рис. 6**) на базі комутаторів SN5600 і SN5610 (64 порти по 800 Гбіт/с або 128 портів по 400 Гбіт/с) застосовується для розгортання інфраструктур найбільших хмарних систем генеративного ШІ. Пропускна здатність рішення сягає 51,2 Тбіт/с. Передбачається спільна робота з DPU BlueField-3, хоча замість DPU можна використати й адаптери ConnectX.

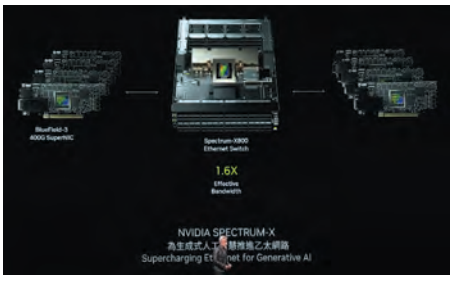

Рис. 6. Платформа NVIDIA Spectrum-X

Саме Spectrum-X лежить в основі еталонної архітектури NVIDIA для GB300 NVL72 — про її топологію йтиметься нижче.

NVIDIA BlueField-3

BlueField-3 (**рис. 7**) — процесор обробки даних, ще одна обчислювальна ланка стійки GB300 NVL72. Його основне завдання — розвантажити центральний процесор Grace, виконуючи мережеві й сервісні функції на власних потужностях DPU. BlueField-3 можна перепрограмувати під особливості мережевого трафіку та для посилення безпеки мережевого оточення комплексу.

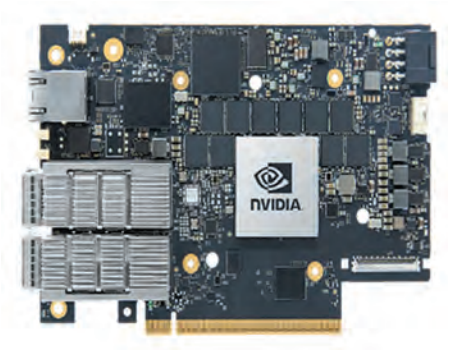

Рис. 7. Процесор обробки даних NVIDIA BlueField-3 B3240 — саме ця модель встановлена в кожному обчислювальному лотку GB300 NVL72

В архітектурі GB300 NVL72 BlueField-3 закриває три задачі:

- **позасмугове забезпечення ресурсів** — DPU працює як оптимізована платформа для площини керування ЦОД, даючи автоматизоване виділення ресурсів і еластичність під змінні навантаження;
- **прискорення роботи зі сховищем** — технологія BlueField SNAP дозволяє віддаленим накопичувачам працювати як локальні;
- **безпечна інфраструктура** — DPU працює в режимі нульової довіри незалежно від хоста і забезпечує міжмережевий екран та мікросегментацію.

Чи вдалий це вибір?

Сукупна пропускна здатність плати B3240 — близько 480 Гбіт/с, і саме вона визначає стелю дискової смуги на вузол: до 40 ГБ/с. Тут варто зупинитися, бо це одне з небагатьох місць архітектури, яке заслуговує критичного погляду.

Передусім зазначимо: вибору тут немає. У специфікаціях OEM-виробників слоти PCIe обчислювального лотка позначені як неконфігуровані — один BlueField-3 і два модулі ConnectX-8, замінити не можна. У першому промисловому розгортанні GB300 у CoreWeave шафа містить рівно вісімнадцять DPU, по одному на лоток.

Логіка NVIDIA цілком виправдана. BlueField-3 — не просто швидкий мережевий адаптер, а самостійна обчислювальна платформа: шістнадцять ядер Arm Cortex-A78, до 32 ГБ DDR5 і апаратні прискорювачі криптографії, дедуплікації та протоколів зберігання. Він виконує роботу, на яку мережева карта не розрахована в принципі — ізоляцію за моделлю нульової довіри незалежно від хоста, SNAP, забезпечення площини керування. Швидший NIC його не замінив би, а симетрія смуги між «схід–захід» і «північ–південь» архітектурі й не потрібна.

Слабке місце, однак, є, і воно конкретне. В одному лотку сусідять два покоління мережевого кремнію: ConnectX-8 на 800 Гбіт/с і BlueField-3, ядро якого — контролер класу ConnectX-7 на 400 Гбіт/с. Причому двопортовість оманлива: плата віддає близько 480 Гбіт/с сумарно на обидва порти, а не 800. NVIDIA рекомендує два порти заради резервування, а не заради смуги. Підсумкова асиметрія між фабриками — приблизно 6,7 до 1.

Де це стає відчутним: при дезагрегованому інференсі з вивантаженням KV-кешу на мережеве сховище. Показово, що NVIDIA визнає це сама — в еталонній архітектурі для платформи HGX B300 прямо рекомендовано обирати модель B3240 замість звичної B3220 саме з розрахунку на майбутні навантаження на кшталт вивантаження KV-кешу на швидке мережеве сховище. Остаточна ж відповідь NVIDIA — BlueField-4 на 800 Гбіт/с у поколінні Vera Rubin.

Практичний висновок для проєктувальника: стеля 40 ГБ/с на вузол — це даність, а не параметр налаштування. Архітектуру сховища і стратегію роботи з KV-кешем треба рахувати від цієї цифри, а не від смуги обчислювальної фабрики.

Мережеві фабрики та топологія

Еталонна архітектура NVIDIA будує для GB300 NVL72 три фізичні мережі, і їх не можна зливати в одну.

GPU Compute (схід–захід) — RDMA-фабрика на комутаторах Spectrum-X у неблокуючій топології fat tree, оптимізованій за рейками: кожен leaf-комутатор обслуговує ту саму позицію GPU в усіх лотках. Це дає найкоротший шлях для колективних операцій NCCL. Сукупна пропускна здатність на лоток — 3200 Гбіт/с.

CPU Converged (північ–південь) — конвергентна мережа для сховища, внутрішньосмугового керування, користувацького доступу та підтримувальних серверів. Кожен лоток підключається двома портами 400 Гбіт/с до двох різних комутаторів. Поділ підмереж — через VLAN поверх спільної фізичної фабрики.

Out-of-Band Management — окрема гігабітна мережа на комутаторах SN2201: порти BMC серверів, BMC самих DPU, керуючі порти комутаторів. Три підключення на лоток, 54 порти на шафу. Ізоляція тут фізична — це вимога безпеки.

Орієнтири щодо смуги, які варто закладати в технічне завдання: не менше 25 Гбіт/с на GPU в напрямку користувацької мережі та не менше 12,5 Гбіт/с на GPU до сховища.

Подвійна площина: навіщо ділити 800G на два по 400G

Найцікавіше проєктне рішення еталонної архітектури — схема 2–4–5–800 із подвійною площиною. Кожен порт ConnectX-8 на 800 Гбіт/с розбивається на два інтерфейси по 400 Гбіт/с, і кожен із них іде на leaf-комутатор іншої, повністю незалежної фабрики.

Мотив прямий: одиничний лінк від адаптера до одного комутатора є єдиною точкою відмови для цілого GPU. У подвійній площині кожна площина несе половину трафіку; за відмови однієї інша приймає навантаження, а деградація лінійно відповідає втраченій смузі. Балансування й обробку відмов виконує сам ConnectX-8 на апаратному рівні.

Чому не звичайний LAG: агрегація каналів має лише локальне уявлення про трафік і розподіляє потоки статичним хешем. Для ШІ-навантажень із низькою ентропією потоків такий розподіл систематично незбалансований. Це один із випадків, коли класична мережева практика ЦОД дає гірший результат, ніж спеціалізоване рішення.

Варто знати й те, що еталонна архітектура описує також одноплощинний варіант топології. Він дешевший, але повертає в схему ту саму єдину точку відмови, заради усунення якої подвійну площину й вигадали. Вибір між ними — це вибір між вартістю фабрики та доступністю кластера, і робити його треба свідомо, а не за замовчуванням.

Масштабована одиниця

Архітектура оперує поняттям ScalableUnit — одна шафа GB300 NVL72, тобто 18 обчислювальних вузлів і 72 GPU. Протестоване масштабування — до восьми таких одиниць; приблизно від 1024 вузлів у топологію вводиться рівень super-spine (**табл. 1**).

Таблиця 1. Параметри масштабування GB300 NVL72

Вузлів	GPU	Шаф	Leaf	Spine	Трансиверів вузол→leaf	Трансиверів між комутаторами
36	144	2	8	4	144	288
72	288	4	16	8	288	576
144	576	8	32	12	576	1152

Два зауваження до цих цифр. По-перше, NVIDIA свідомо недозаповнює leaf-комутатори: кожен здатний обслужити три шафи, але конфігурується на дві — це вимога програмної моделі GB300 NVL72, а не помилка проєктування. По-друге, кількість трансиверів зростає швидше за кількість GPU: на восьми шафах лише фабрика «схід–захід» потребує понад 1700 оптичних модулів. Це і стаття витрат, і домінантний внесок у частоту відмов усієї інсталяції.

Дванадцять керуючих вузлів площини керування з увімкненою високою доступністю — саме таку кількість NVIDIA вважає достатньою для великого корпоративного розгортання. Вузли можуть бути на x86 (два сокети, 32 ядра, 512 ГБ DDR5) або на ARM — у варіанті з двома процесорами Grace і 480 ГБ LPDDR5X. Кожен несе чотири однопортові адаптери ConnectX-7 на 200G, два накопичувачі NVMe по 7,68 ТБ під кеш даних і дзеркальну пару під операційну систему.

До дванадцяти підтримувальних серверів, під'єднаних чотирма портами по 200 Гбіт/с. Вони забезпечують керування, розгортання, моніторинг і контроль для решти кластера.

Програмне забезпечення сховища. Окрема примітка в документі NVIDIA каже прямо: керуюче ПЗ для систем зберігання постачають партнери, і до складу програмного стека Enterprise RA воно не входить. Саме сховище — також. Це класична діра в кошторисі: замовник купив «готове рішення», а через місяць з'ясовується, що дані зберігати ніде й нема чим керувати.

Разом — близько двох десятків серверів плюс окремо закуплена система зберігання. Навіть на двошафовій інсталяції це помітна частка бюджету, і планувати її треба на етапі концепції, а не після підписання договору.

Продуктивність GB300 NVL72

Продуктивності стійки вистачить для розроблення навіть найвимогливіших моделей із трильйонами параметрів. NVIDIA заявляє до 50-кратного зростання виходу ШІ-фабрики порівняно з платформами покоління Hopper і 30-кратне прискорення інференсу трильйонних

моделей у реальному часі за рахунок п'ятого покоління NVLink. Офіційна специфікація комплексу має такий вигляд (**табл. 2**).

Таблиця 2. Специфікація комплексу GB300 NVL72

Параметр	NVIDIA GB300 NVL72
Конфігурація	72 GPU Blackwell Ultra, 36 CPU Grace
Пропускна здатність NVLink	130 ТБ/с
Швидка пам'ять	37 ТБ
Пам'ять GPU | смуга	20 ТБ | до 576 ТБ/с
Пам'ять CPU | смуга	17 ТБ LPDDR5X | 14 ТБ/с
Кількість ядер CPU	2592 ядра Arm Neoverse V2
FP4 Tensor Core	1440 | 1080 PFLOPS
FP8 / FP6 Tensor Core	720 PFLOPS
INT8 Tensor Core	24 POPS
FP16 / BF16 Tensor Core	360 PFLOPS
TF32 Tensor Core	180 PFLOPS
FP32	6 PFLOPS
FP64 / FP64 Tensor Core	100 TFLOPS

Перше значення у рядку FP4 наведене з урахуванням розрідженості, друге — для щільних обчислень. Два рядки заслуговують окремого коментаря, бо саме вони визначають, для яких задач машина придатна, а для яких — ні.

INT8 — 24 POPS. У попередньому поколінні GB200 NVL72 цей показник становив 720 POPS. Падіння тридцятикратне. NVIDIA свідомо прибрала більшу частину відповідного кремнію, звільнивши площу під формати FP4 і FP8, на яких сьогодні працює інференс сучасних моделей. Для замовника з парком навчених під INT8 моделей це означає необхідність переходу на FP8 — інакше GB300 працюватиме гірше за попередника.

FP64 — 100 TFLOPS. Тут падіння ще драматичніше: GB200 NVL72 дає близько 2880 TFLOPS подвійної точності. Висновок для проєктувальника однозначний: GB300 NVL72 — це машина для ШІ, а не для класичного HPC. Розрахункові задачі, які спираються на подвійну точність — обчислювальна гідродинаміка, скінченні елементи, наукове моделювання, — на цій платформі виконуватимуться в рази повільніше, ніж на спеціалізованих рішеннях попередніх поколінь. Якщо замовник планує змішане навантаження, розділяти ШІ-контур і HPC-контур треба на етапі концепції, а не після закупівлі.

Vera Rubin

Якщо GB300 був заміною кремнію в наявній конструкції, то Vera Rubin змінює саму одиницю проєктування. Платформу оголошено на GTC у березні 2026 року і виведено в повносерійне виробництво до GTC Taipei. Вона складається із семи нових чипів: CPU Vera, GPU Rubin, комутатора NVLink 6, SuperNIC ConnectX-9, DPU BlueField-4, Ethernet-комутатора Spectrum-6 і прискорювача інференсу Groq 3 LPU.

Ці чипи зібрані у п'ять типів шаф, які працюють як єдиний POD — цілісний ШІ-суперкомп'ютер:

- **Vera Rubin NVL72** — обчислювальні GPU-шафи;
- **Vera CPU** — шафи процесорних вузлів для пісочниць і CPU-фаз агентних робочих процесів;
- **Groq 3 LPX** — шафи прискорювачів інференсу;
- **BlueField-4 STX** — шафи сховища для контекстної пам'яті;
- **Spectrum-6 SPX** — шафи мережевої комутації.

Заявлений приріст — десятикратна пропускна здатність агентних сценаріїв порівняно з платформою Grace Blackwell, навчання на чверті кількості GPU і десятикратне зниження вартості мільйона токенів на інференсі.

Що змінилося в компонентах

Vera CPU відходить від ліцензованих ядер Neoverse: 88 власних ядер NVIDIA Olympus, сумісних з ARM. Rubin GPU переходить на пам'ять HBM4 і несе Transformer Engine продуктивністю близько 50 PFLOPS у форматі NVFP4.

NVLink 6 подвоює смугу до 3,6 ТБ/с на GPU, а сукупна пропускна здатність шафи зростає зі 130 до 260 ТБ/с. Комутаційний лоток додає внутрішньомережеві обчислення для прискорення колективних операцій маршрутизації MoE-моделей.

ConnectX-9 подвоює зовнішню смугу до 1,6 Тбіт/с на GPU, по вісім адаптерів на лоток. BlueField-4 несе програмно-визначену мережу на 800 Гбіт/с, вбудовану ізоляцію орендарів і архітектуру довіреного ресурсу ASTRA.

Розділення prefill і decode

Найнесподіваніший елемент платформи — прискорювач Groq 3 LPU, що з'явився внаслідок придбання компанії Groq наприкінці 2025 року й замінив раніше анонсований Rubin CPX.

Ідея інженерно чиста. Інференс великої моделі складається з двох фаз із протилежними вимогами: prefill (обробка вхідного контексту) впирається в обчислення, decode (генерація токенів) — у затримку доступу до пам'яті. NVIDIA розводить ці фази по різному кремнію: GPU Rubin виконують prefill і механізм уваги, LPU виконують decode, а програмний шар Dynamo маршрутизує роботу між ними потоково. Змін у коді CUDA при цьому не потрібно — LPU працює як прискорювач до наявного стека.

Шафа LPX містить 256 прискорювачів і робить ставку на великий обсяг SRAM на кристалі замість HBM: сукупно 128 ГБ пам'яті з пропускною здатністю близько 40 ПБ/с. Для послідовного decode така пара «менша ємність — надзвичайна смуга» працює краще за HBM. Заявлений ефект від зв'язки Vera Rubin NVL72 з шафами LPX — до 35-кратного зростання пропускної здатності інференсу на мегават для трильйонних моделей.

Для системного інтегратора звідси випливає цілком конкретна річ: у проєкті ШІ-фабрики покоління Rubin з'являється окремий тип шафи з власним профілем живлення й охолодження, а розрахунок продуктивності перестає бути лінійною функцією кількості GPU.

Оптика переїжджає всередину комутатора

Vera Rubin виводить у виробництво Spectrum-X Ethernet Photonics — перші комутатори зі спільно упакованою оптикою (CPO) на SerDes 200 Гбіт/с із комутаційною здатністю 102 ТБ/с. Заявлені показники варто читати саме як експлуатаційні: уп'ятеро краща енергоефективність, уп'ятеро довший час безвідмовної роботи ШІ-інсталяції й на 30% швидше введення в експлуатацію порівняно з мережами на знімних трансиверах.

Поверніться до таблиці масштабованих одиниць: понад 1700 оптичних модулів на восьми шафах — це і споживання, і найчисленніша популяція компонентів, що відмовляють, і левова частка часу монтажу та продзвонювання. CPO прибирає більшу частину цієї популяції. Для інтегратора це, ймовірно, найважливіша зміна покоління після самої щільності.

Безпека на рівні шафи

Vera Rubin NVL72 спроєктовано з наскрізними **конфіденційними обчисленнями**: дані шифруються на високошвидкісних з'єднаннях, а апаратна атестація підтверджує цілісність системи. Поверх BlueField-4 працює програмна платформа DOCA: ізоляція мультитенантних мереж, політики нульової довіри, виявлення загроз під час виконання та наскрізне шифрування на швидкості 800 Гбіт/с без навантаження на CPU хоста.

Для сценаріїв, де в контурі є регульовані дані, це вперше дає інтегратору технічну відповідь замість організаційної.

Інженерія будівлі: де закінчується типовий ЦОД

Саме тут криється головний ризик проєкту, і саме про нього найчастіше згадують в останню чергу.

Для калібрування: за даними Uptime Institute, середня корпоративна шафа споживає близько 9 кВт. GB300 NVL72 у номіналі бере 132–142 кВт із піками близько 155 кВт, важить понад півтори тонни у спорядженому стані й відводить у рідину близько 90 % тепла. Vera Rubin піднімає планку майже вдвічі — до 190–230 кВт. А заявлена на 2027 рік архітектура Kyber розрахована на близько 600 кВт на шафу й вимагає переходу на розподіл живлення 800 В постійного струму з винесеними в окремі шафи модулями живлення та охолодження.

Що з цього випливає практично.

Живлення. У GB300 — вісім полиць по 33 кВт (шість блоків по 5,5 кВт у кожній), розподіл через шину постійного струму. Внутрішньостійкові комутатори SN2201 живляться від тієї ж шини. Резервування закладається в межах шафи; якщо будівля цього не забезпечує, компонування доводиться переглядати.

Охолодження. Повітряне охолодження не розглядається як опція взагалі. Пряме рідинне охолодження означає CDU, вторинний контур, вимоги до температури зворотної води та якості теплоносія. У GB300 інтегровано виявлення протікань на рівні лотка й на рівні шафи — це штатна підсистема, яку треба заводити в моніторинг нарівні з електрикою.

Механіка й будівля. Півтори тонни на шафу — це навантаження на перекриття, маршрути завезення й підйомне обладнання. Стійка займає ту саму площу, що й звичайна 42U, тож щільність навантаження на квадратний метр зростає непропорційно.

Програмний стек. Його не варто планувати за залишковим принципом: Mission Control для операцій, AI Enterprise для платформи, Dynamo для дезагрегованого інференсу, Run:ai або Kubernetes і Slurm для оркестрації, NetQ для валідації фабрики. Ліцензування AI Enterprise — підписка з розрахунку на GPU, і це помітна частина сукупної вартості володіння.

* * *

GB300 NVL72 — найдосконаліше на сьогодні стійкове рішення NVIDIA для сектору ШІ та HPC. Компанія не помиляється, позиціонуючи його як один великий GPU: усі компоненти шафи працюють як єдине ціле завдяки NVLink п'ятого покоління. Приріст відносно попередника отримано всередині незмінної інженерної оболонки, і саме це робить перехід реалістичним для замовників, які вже мають відповідний машинний зал.

Vera Rubin цю оболонку змінює: подвоєна внутрішня смуга, оптика всередині комутаторів, окремий тип шафи під фазу decode, конфіденційні обчислення на рівні стійки і споживання, що наближається до чверті мегавата. Одиницею проєктування стає не шафа, а POD із п'яти різнотипних шаф.

За два покоління NVIDIA послідовно піднімає рівень абстракції продажу: від прискорювача до шафи, від шафи до POD. Для системного інтегратора висновок неприємний, але чіткий: компетенція зміщується з IT у бік інженерії будівлі. Той, хто рахує кіловати, гідравліку й кабельний журнал так само впевнено, як топології Clos, отримує проєкти. Той, хто продовжує мислити категоріями серверної кімнати, — ні.

Андрій РОГОВ,
директор ПП «Конус»
andrey.rogov@gmail.com