E-BOOK

INFRAESTRUTURA DE REFRIGERAÇÃO LÍQUIDA PARA NVIDIA VERA RUBIN NVL72

Objetivos, aplicações, arquitetura DLC, CDU, TCS, FWS, dimensionamento e boas práticas para AI Factories e supercomputação.

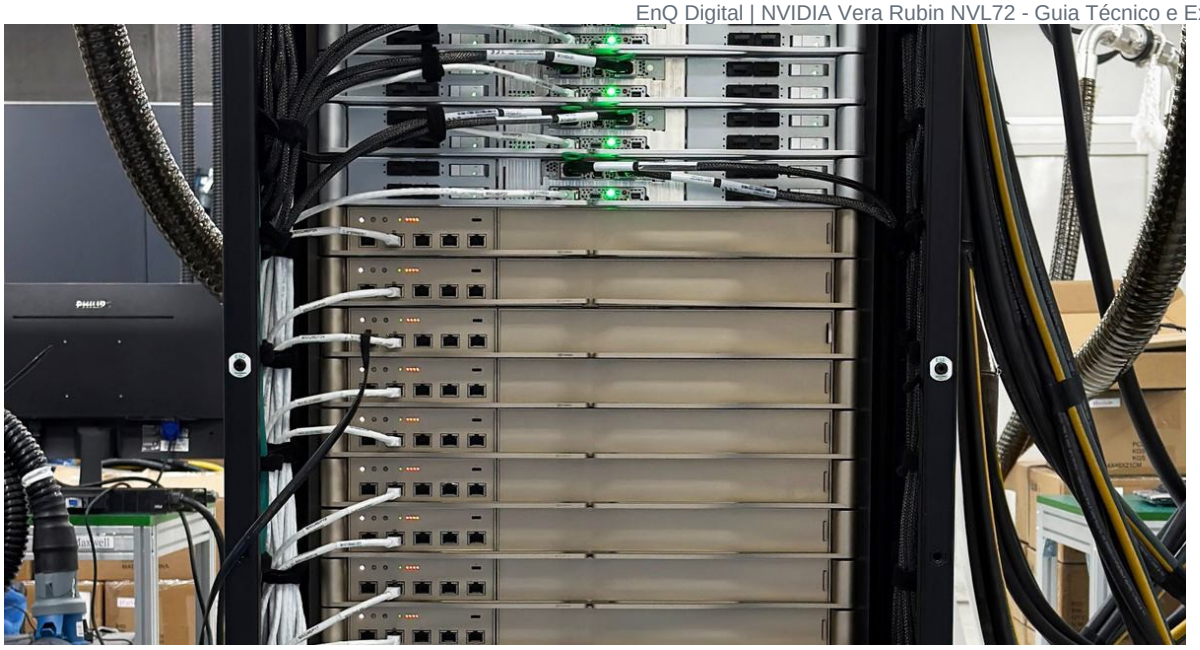
GUIA TÉCNICO E EXECUTIVO

Galeria - plataforma Vera Rubin em imagens reais

Registro visual de racks e módulos associados à plataforma Vera Rubin. As fotografias foram incorporadas a partir do e-book técnico da EnQ Digital e ajudam a contextualizar a densidade física do NVL72, o rack de potência e os módulos de interconexão.

Vera Rubin NVL72 em exposição ao lado do rack de potência 800 VDC.

Vista de produto do rack Vera Rubin NVL72, destacando a densidade dos módulos.



Detalhe real de módulos, cabos e indicadores em ambiente de validação.

Fonte: Galeria do e-book EnQ Digital; créditos originais: NVIDIA e fotografias editoriais indicadas na seção de referências.

1. Por que construir um supercomputador desta classe?

Um Vera Rubin NVL72 não deve ser tratado apenas como um rack de GPUs de altíssima densidade. Ele é o núcleo de uma AI Factory: uma plataforma destinada a transformar grandes volumes de dados e energia elétrica em treinamento de modelos, inferência, raciocínio de IA, simulação científica e resultados de negócio em escala.

> **OBJETIVO CENTRAL**
>
> Reduzir drasticamente o tempo para treinar, ajustar e executar modelos e simulações que seriam impraticáveis em infraestrutura convencional - preservando baixa latência entre GPUs, grande memória agregada e alta eficiência por unidade de trabalho.

1.1 O valor não está apenas nos FLOPS

- Time-to-solution: reduzir semanas ou meses de treinamento, simulação e processamento para dias ou horas.
- Modelos maiores: treinar, ajustar e servir LLMs, modelos multimodais e Mixture-of-Experts com centenas de bilhões ou trilhões de parâmetros.
- Raciocínio em escala: sustentar inferência com long context, agentes e cadeias de raciocínio que consomem muito mais compute que um chatbot tradicional.
- Soberania e privacidade: manter dados, modelos e propriedade intelectual dentro de uma infraestrutura controlada pela organização ou pelo país.
- Convergência AI + HPC: usar a mesma fábrica para IA, análise massiva de dados, simulação numérica e workloads científicos.
- Economia por token e por resultado: em workloads grandes e contínuos, desempenho por watt, utilização e velocidade de comunicação passam a determinar o TCO.
- Monetização: provedores de cloud, data centers e AI factories podem transformar a infraestrutura em GPU-as-a-Service, training-as-a-service e inference-as-a-service.

1.2 Quando faz sentido - e quando não faz

Um cluster NVL72 se justifica quando existe demanda sustentada por computação paralela, grandes datasets, forte comunicação GPU-GPU, necessidade de baixa latência, soberania de dados ou volume de inferência suficiente para amortizar a infraestrutura. Para projetos pequenos, experimentais ou com utilização esporádica, cloud pública ou clusters menores podem apresentar melhor relação econômica.

Indicador	Sinal de que uma AI Factory pode fazer sentido
Treinamento	Modelos proprietários grandes, treinamento recorrente ou fine-tuning contínuo.
Inferência	Bilhões de tokens/dia, baixa latência e agentes executando fluxos complexos.
Dados	Petabytes de dados ou datasets estratégicos que não devem sair do domínio controlado.
HPC	Simulações altamente paralelas, CFD, FEA, clima, molecular dynamics ou seismic.
Negócio	Capacidade será usada continuamente ou revendida como serviço de GPU/IA.
Soberania	Requisitos de residência de dados, segurança, compliance ou propriedade intelectual.

2. Para quais áreas um supercomputador Vera Rubin é útil?

A mesma infraestrutura pode atender IA generativa e HPC. O ganho aparece principalmente quando o problema pode ser paralelizado e exige grande largura de banda de memória e comunicação entre aceleradores.

Área	Workloads típicos	Objetivo
AI Labs / LLM	Treinamento de foundation models, MoE, multimodal, reasoning, fine-tuning e RL.	Criar modelos proprietários e reduzir tempo de treinamento.
Cloud / Data Center	GPUaaS, bare metal GPU, inference endpoints, clusters dedicados.	Monetizar capacidade e atender clientes de IA/HPC.
Bancos e Finanças	Fraude, AML, risco, pricing, modelos tabulares, agentes, análise de documentos e atendimento.	Processar grandes bases e executar IA privada com baixa latência.
Saúde e Life Sciences	Drug discovery, dinâmica molecular, genômica, protein design, imagem médica e modelos biomédicos.	Reduzir ciclos de pesquisa e testar mais hipóteses computacionalmente.
Indústria / Engenharia	CFD, FEA, digital twins, generative engineering, robótica e visão industrial.	Diminuir prototipagem física e acelerar otimização de produtos/processos.
Energia / Óleo e Gás	Seismic imaging, reservoir simulation, otimização de rede, previsão de demanda e manutenção preditiva.	Aumentar resolução de modelos e reduzir tempo de decisão operacional.
Clima e Meio Ambiente	Weather forecasting, climate modeling, assimilação de dados e modelos Earth-2-like.	Executar previsões mais rápidas e de maior resolução.
Telecom	Planejamento de RF, otimização de rede, IA para NOC, agentes e modelos de tráfego.	Automatizar operação e prever congestionamentos/falhas.
Governo / Sovereign AI	LLMs nacionais, pesquisa, defesa cibernética, ciência e serviços públicos.	Preservar soberania de dados e capacidade estratégica de IA.
Cybersecurity	Detecção em larga escala, análise de malware, modelos de comportamento e SOC com agentes.	Correlacionar grandes volumes de telemetria e acelerar resposta.
Mídia / Conteúdo	Geração de vídeo, imagem, áudio, tradução e renderização neural.	Produzir conteúdo multimodal em grande escala.
Pesquisa científica	Física, química, materiais, astronomia, bioinformática e métodos híbridos AI+HPC.	Explorar espaços de solução antes inviáveis por custo computacional.

EXEMPLO PRÁTICO

Uma instituição pode usar o mesmo cluster durante o dia para inferência e agentes corporativos e, em janelas planejadas, alocar milhares de GPUs para treinamento, simulação ou processamento científico via Kubernetes/Slurm. O valor vem da utilização compartilhada e da capacidade de orquestrar workloads distintos.

3. Cases reais - Blackwell B200 / GB200 NVL72

Os sistemas Blackwell anteriores ao Rubin já demonstram como uma AI Factory é utilizada em produção. Esses casos ajudam a entender o tipo de workload e o modelo econômico que o Vera Rubin tende a ampliar.

> **NOMENCLATURA**
>
> B200 e B300 referem-se às gerações de GPU Blackwell/Blackwell Ultra. GB200 e GB300 combinam Grace CPU + Blackwell GPU no Superchip e, no NVL72, formam o sistema rack-scale com 72 GPUs.

3.1 CoreWeave: GB200 para modelos de raciocínio e agentic AI

Em abril de 2025, a NVIDIA anunciou que milhares de GPUs Grace Blackwell estavam ativas na CoreWeave. Cohere, IBM e Mistral AI passaram a utilizar sistemas GB200 NVL72 para treinar e colocar em produção modelos de próxima geração, incluindo reasoning models e agentic AI.

Fonte NVIDIA - CoreWeave GB200 NVL72

3.2 Oracle/OCI: milhares de Blackwell em cloud

A Oracle colocou em operação sua primeira onda de racks NVIDIA GB200 NVL72 liquid-cooled e anunciou milhares de GPUs Blackwell para OCI e NVIDIA DGX Cloud. O deployment combina Quantum InfiniBand, Spectrum-X Ethernet e integrações do stack OCI/NVIDIA para treinamento e inferência de modelos de raciocínio e agentes.

Fonte NVIDIA - OCI Blackwell deployment

3.3 NeoSpace: case latino-americano em finanças

A NeoSpace está entre as primeiras empresas da América Latina a implantar GB200 NVL72 em produção via OCI. Seu NeoData é usado em ambiente multi-cloud para um dos maiores bancos privados da América Latina, com mais de 60 milhões de clientes, aplicando foundation models a dados estruturados e não estruturados do setor financeiro.

Case NVIDIA - NeoSpace / Finanças

3.4 DeepL: Mixture-of-Experts sobre GB200

A NVIDIA relata que a DeepL utiliza hardware GB200 para treinar modelos Mixture-of-Experts de próxima geração, buscando maior eficiência tanto no treinamento quanto na inferência. É um exemplo direto de uma empresa de IA usando rack-scale NVL72 para um produto global de linguagem.

Fonte NVIDIA - MoE e Blackwell

4. Cases Blackwell Ultra B300 / GB300 e a transição para Rubin

O GB300 NVL72 elevou o foco em reasoning e inferência de grande escala. A NVIDIA posiciona a arquitetura para inferência em tempo real dos maiores modelos, treinamento/fine-tuning de modelos de trilhões de parâmetros, analytics massivo e HPC de precisão simples ou mista.

4.1 Microsoft / OpenAI

A NVIDIA informou que a Microsoft implantou um cluster Azure em grande escala usando GB300 NVL72 para OpenAI. O caso ilustra o papel de um supercomputador rack-scale como infraestrutura de treinamento e serving para modelos de fronteira, onde throughput, latência, memória e rede precisam escalar como um único sistema.

4.2 Lambda: AI Factory de 100+ MW

A Lambda anunciou uma AI Factory de mais de 100 MW em Kansas City, planejada inicialmente com mais de 10.000 GPUs GB300 NVL72 para acelerar trabalhos de pesquisadores, empresas e desenvolvedores. É um exemplo de transformação de energia e data center em capacidade de computação vendável.

4.3 Together AI / 5C: B200, GB200 e GB300

A Together AI, em parceria com a 5C, opera uma AI Factory em Maryland com GPUs B200 e anunciou outra instalação em Memphis com sistemas GB200 e GB300. O modelo combina infraestrutura física de alta densidade com serviços de treinamento e inferência para aplicações AI-native.

4.4 Global AI: 128 racks GB300 NVL72

A NVIDIA anunciou a compra de 128 racks GB300 NVL72 pela Global AI, totalizando mais de 9.000 GPUs, para uma implantação em Nova York. O caso mostra a escala de um projeto comercial em que o rack deixa de ser uma unidade isolada e passa a ser um bloco repetível de uma AI Factory.

Fonte NVIDIA - AI Infrastructure in America / GB300 deployments

> **O QUE ESSES CASES ENSINAM**
>
> O objetivo não é simplesmente possuir GPUs. O valor aparece quando compute, NVLink/scale-out network, storage, software, energia e cooling são tratados como uma única fábrica capaz de entregar tokens, treinamento, simulações e serviços com previsibilidade operacional.

4.5 Por que o Vera Rubin é o próximo passo

Em 2026, a NVIDIA informou que Vera Rubin NVL72 entrou em ramp de produção, com racks operando em parceiros como CoreWeave, Google Cloud, Microsoft Azure, Oracle Cloud Infrastructure e Nebius. A Microsoft também anunciou Rubin para futuros sites Fairwater, enquanto a CoreWeave passou a integrar sistemas Rubin para workloads de treinamento, inferência e agentic AI.

Foto real de referência: rack NVIDIA Vera Rubin NVL72 demonstrado na GTC 2026 (implementação de parceiro/OEM).

Fonte: NVIDIA Newsroom / Blog - Rubin platform and worldwide deployments; foto de referência de evento GTC 2026.

NVIDIA - Vera Rubin platform / deployments

NVIDIA - Rubin platform launch

> **MENSAGEM EXECUTIVA**
>
> Vera Rubin deve ser entendido como infraestrutura estratégica para AI Factory e HPC: alto CAPEX e alta densidade, mas capaz de concentrar em poucos racks uma capacidade que permite criar modelos, serviços e simulações de uma classe impossível para servidores convencionais.

5. Vera Rubin como arquitetura sistêmica de AI Factory

A Vera Rubin é uma arquitetura de infraestrutura de IA em escala de rack e POD. Ela foi desenhada para que computação, memória, comunicação, rede, armazenamento, segurança, potência e resfriamento funcionem como um sistema coordenado. Em modelos atuais, desempenho não depende apenas de FLOPS: é necessário movimentar pesos, ativações e KV cache, sincronizar aceleradores, manter baixa latência e entregar mais tokens por watt.

> **MUDANÇA ESTRATÉGICA**
>
> De servidor isolado para domínio de execução em escala de rack; de pico de FLOPS para tokens úteis por watt e por megawatt; de treinamento puro para treinamento, pós-treinamento e inferência agentiva; de interconexão periférica para memória e comunicação como parte do acelerador.

A plataforma anunciada pela NVIDIA reúne Rubin GPU, Vera CPU, NVLink 6 Switch, ConnectX-9 SuperNIC, BlueField-4 DPU e novas famílias de networking e racks especializados, incluindo compute, CPU, storage e Ethernet. O NVL72 é o núcleo de compute desse ecossistema.

6. A GPU Rubin em detalhes

A Rubin GPU utiliza uma arquitetura orientada a workloads de raciocínio, geração, recuperação de dados e uso de ferramentas. A combinação de HBM4, Transformer Engine e NVLink 6 busca manter os aceleradores ocupados mesmo em cargas com comunicação intensa e múltiplas etapas de inferência.

Item	Rubin GPU	Vera Rubin NVL72
Transistores	336 bilhões	-
SMs / Tensor Cores	224 / 896	72 GPUs
NVFP4 inferência	50 PFLOPS	3.600 PFLOPS
Memória	288 GB HBM4	20,7 TB HBM4
Largura de banda HBM4	22 TB/s	1.580 TB/s agregado
Interconexão	NVLink 6: 3,6 TB/s	260 TB/s no rack

LEITURA CORRETA

Os valores acima são especificações/picos divulgados para a plataforma. Desempenho de aplicação depende de framework, kernel, precisão, modelo, batch, contexto e estratégia de paralelismo.

7. Memória HBM4 e NVLink 6

Em inferência, especialmente na fase de decode, mover pesos e estado do modelo pode ser tão determinante quanto executar multiplicações. Por isso, Vera Rubin combina capacidade de memória, grande largura de banda e uma malha de comunicação de baixa latência.

7.1 HBM4

- Até 288 GB de HBM4 por GPU Rubin.
- Até 22 TB/s de largura de banda de memória por GPU.
- Mais espaço para pesos, ativações, KV cache e concorrência.
- Menor necessidade de descarregar estado para camadas mais lentas de memória ou storage.

7.2 NVLink 6

- Até 3,6 TB/s de scale-up por GPU para o domínio NVLink.
- Comunicação all-to-all entre GPUs do NVL72.
- NVLink-C2C de alta largura de banda para comunicação coerente CPU-GPU.
- Redução do overhead de sincronização e do tempo perdido em transferência em workloads MoE e distribuídos.

> **POR QUE ISSO IMPORTA**
>
> Modelos Mixture-of-Experts distribuem tokens entre especialistas. Quanto mais tempo o rack gasta transferindo pesos e sincronizando GPUs, menor é a fração de energia convertida em trabalho útil. HBM4 e NVLink 6 atacam diretamente esse gargalo.

8. Treinamento, inferência e IA agentiva

Um agente não responde apenas a uma pergunta: ele planeja, chama ferramentas, consulta fontes, verifica resultados e pode gerar várias sequências de tokens antes de concluir uma tarefa. Isso aumenta o volume de inferência e torna a latência por etapa crítica.

Treinamento	Inferência / Agentic AI
Modelos MoE e foundation models de grande escala.	Serviços interativos com baixa latência.
Tensor, pipeline e expert parallelism.	Long-context e grandes KV caches.
Alta comunicação para sincronização e roteamento de tokens.	Alta concorrência e batching contínuo.
HBM4 para manter parâmetros e estados próximos do compute.	Otimização por tokens/s, tokens/W e custo por milhão de tokens.

INDICADORES DIVULGADOS PELA NVIDIA

Em cenários específicos, a NVIDIA posiciona Vera Rubin para entregar ganhos expressivos em tokens por megawatt e custo por milhão de tokens frente a gerações anteriores. Trate esses números como referências de plataforma, condicionadas a modelo, software, precisão e configuração.

9. Rede, armazenamento e segurança

A plataforma não termina na GPU. Para manter 72 aceleradores ocupados, tráfego east-west, acesso ao storage e proteção do workload precisam ser tratados como partes do computador.

9.1 Rede

- ConnectX-9 SuperNIC para scale-out de altíssima largura de banda.
- Quantum-X800 InfiniBand e Spectrum-X Ethernet para clusters multi-rack.
- RDMA e caminhos de baixa latência para treinamento distribuído e inferência.
- Arquiteturas spine-leaf e ópticas de alta densidade precisam ser dimensionadas junto ao compute.

9.2 Armazenamento contextual

A arquitetura BlueField-4/STX amplia o papel do storage para além de um repositório de arquivos, aproximando camadas de contexto e KV cache do domínio de execução. Em projetos reais, isso deve ser combinado com NVMe, parallel file systems, object storage e políticas de tiering.

9.3 Segurança

- BlueField-4 para processamento de rede, offload e políticas de segurança.
- Confidential Computing e proteção de dados em uso, trânsito e repouso, conforme suporte da plataforma.
- Isolamento de tenants em ambientes cloud e GPUaaS.
- Secure boot, firmware assinado, observabilidade e RAS em escala de rack.

10. Vera Rubin versus Blackwell

A comparação deve ser feita no nível de sistema. Rubin não é apenas uma atualização de GPU: ela altera a forma como o data center entrega computação para IA, com novas exigências de memória, rede, potência e refrigeração.

Critério	Blackwell / GB200	Vera Rubin / NVL72
Foco	Treinamento e inferência em escala	IA agentiva, long-context e eficiência
Memória	HBM3e, conforme modelo	HBM4, até 288 GB/GPU
Interconexão	NVLink 5	NVLink 6
Rack de referência	GB200 NVL72	Vera Rubin NVL72
Resfriamento	Direct liquid cooling	Direct liquid cooling, rack integrado
Economia	Alto throughput	Foco em tokens por watt e por MW

10.1 Requisitos integrados para o Data Center

Domínio	Pontos a validar
Espaço	U ocupado, peso, acesso, rota de movimentação, piso e área de manutenção.
Energia	kW/kVA por rack, A/B, UPS, PDU, proteção, aterramento, harmônicas e expansão.
Térmico	CDU, vazão, ΔP, supply/return, temperatura, qualidade do fluido e leak detection.
Rede	InfiniBand/Ethernet, fibra, transceptores, latência, cabeamento e spine-leaf.
Storage	NVMe, parallel file system, object storage, throughput e metadata.
Operação	NOC, DCIM/BMS, telemetria, spare parts, treinamento, RMA e procedimentos de emergência.

CHECKLIST DE RECEBIMENTO

Confirmar modelo e revisão; obter desenho dimensional/peso; solicitar Max-P/TDP, vazão nominal/mínima/máxima e ΔP; validar supply/return e CDU; executar FAT/SAT, teste hidráulico, teste de carga e validação de rede.

11. Sumário executivo de facilities

O Vera Rubin NVL72 representa uma mudança de escala no projeto de infraestrutura física para IA. A plataforma reúne 72 GPUs NVIDIA Rubin, 36 CPUs NVIDIA Vera, NVLink 6, ConnectX-9 e BlueField-4 em um único rack-scale system. Para facilities, a referência NVIDIA DSX evolui para um Cabinet TDP de até 330 kW por rack e para uma estratégia de refrigeração líquida de alta temperatura.

> **NÚMEROS-BASE PARA PRÉ-PROJETO**
>
> Cabinet TDP DSX: 330 kW/rack | TCS design flow: >= 1,5 LPM/kW | Vazão de referência por rack: >= 495 LPM (29,7 m³/h) | CDU: liquid-to-liquid | Redundância de CDU: N+1 | Cooling design point: 45°C.

Este e-book consolida a arquitetura de Direct Liquid Cooling (DLC), explica a separação FWS/TCS, apresenta os dados necessários para seleção de CDU, mostra como interpretar ΔP e Pump Head e organiza uma RFI de engenharia para obter os dados definitivos do fabricante.

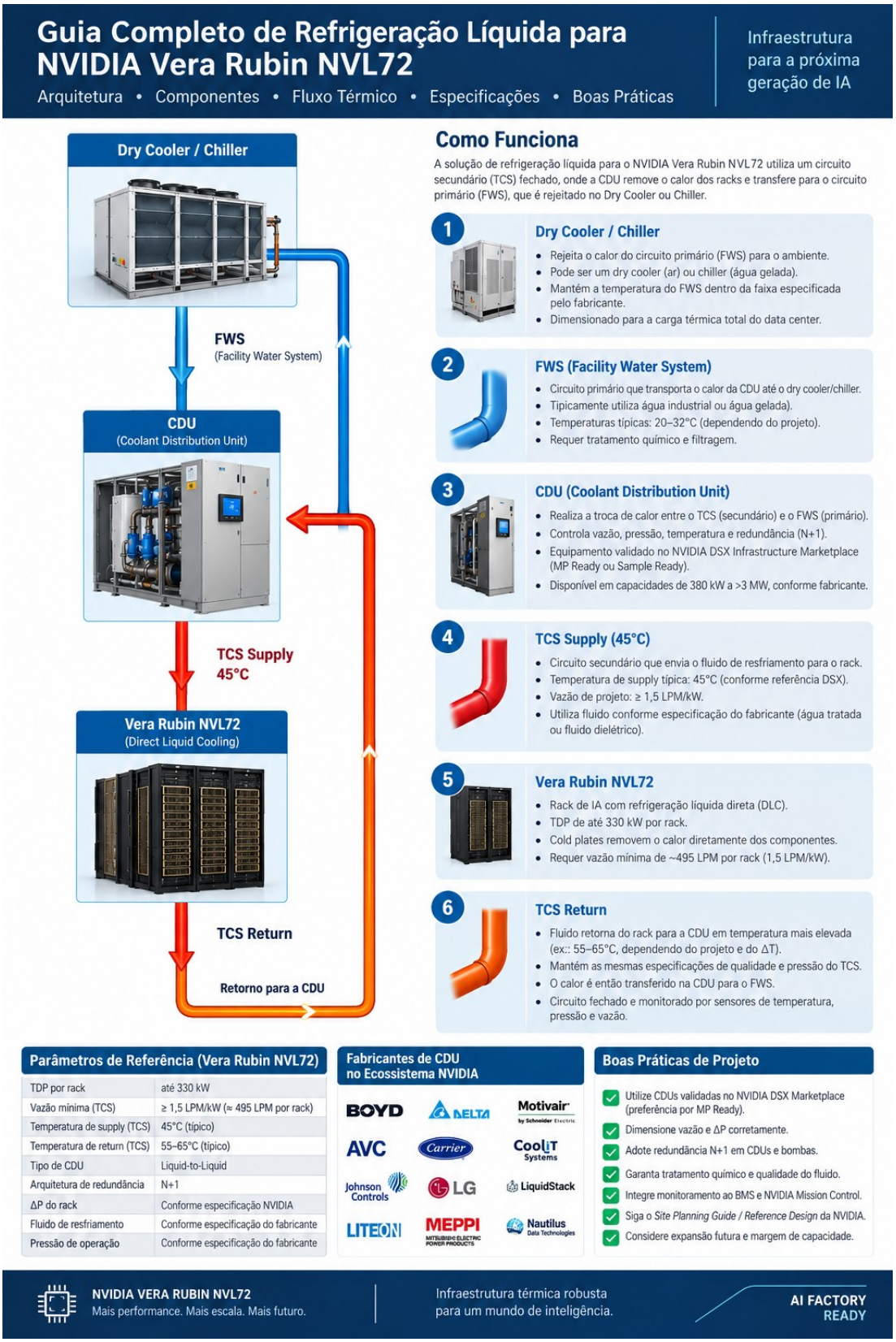
> **IMPORTANTE**
>
> A NVIDIA ainda classifica várias especificações Rubin como preliminares. O Reference Design e o Site Planning Guide da configuração efetivamente adquirida devem prevalecer sobre qualquer cálculo preliminar deste material.

Fonte: NVIDIA Vera Rubin NVL72; NVIDIA DSX Facilities Infrastructure Reference Design Overview; dados consultados em 04/09/2026.

O que você encontrará

- Arquitetura do Vera Rubin NVL72 e anatomia de compute trays e NVLink switch trays.
- Conceitos FWS, CDU, TCS, DLC, cold plates, manifolds e UQD08.
- Dimensionamento térmico e hidráulico: kW, LPM, m³/h, ΔT, ΔP e Pump Head.
- Critérios para selecionar uma CDU e interpretar MP Ready x Sample Ready.
- Fabricantes presentes no NVIDIA DSX Infrastructure Marketplace.
- Checklist/RFI com 12 informações indispensáveis antes do projeto executivo.
- Exemplos de capacidade para 1, 4, 8 e 16 racks Vera Rubin NVL72.
- Integração com BMS, Mission Control/DSX Exchange, comissionamento e operação.

12. Visão rápida da arquitetura de cooling



Fonte: Ilustração técnica consolidada. Valores de referência devem ser confirmados no projeto final NVIDIA/OEM.

13. NVIDIA Vera Rubin NVL72 - visão técnica do rack

O NVIDIA Vera Rubin NVL72 é um sistema rack-scale da terceira geração MGX. A NVIDIA informa que o rack integra 72 GPUs Rubin e 36 CPUs Vera, interligadas por NVLink 6, com networking ConnectX-9 e BlueField-4. O objetivo é fazer o rack operar como um grande acelerador único dentro da AI Factory.

Parâmetro	Referência
GPUs	72 x NVIDIA Rubin
CPUs	36 x NVIDIA Vera
HBM4 total	20,7 TB
Bandwidth HBM4	até 1.580 TB/s
CPU memory	54 TB LPDDR5X
Memória rápida total DGX	75 TB
NVLink	6ª geração
NVLink switch system	9 x L1 NVLink Switch Units
Scale-out	ConnectX-9 / BlueField-4
Cabinet TDP de facilities	até 330 kW - referência DSX

Fonte: NVIDIA Vera Rubin NVL72 / DGX Vera Rubin NVL72; NVIDIA DSX Facilities Infrastructure Reference Design Overview.

13.1 Estrutura física do rack

- 18 compute trays.
- 9 NVLink 6 switch trays.
- Cada compute tray integra 2 Vera Rubin Superchips.
- Cada compute tray totaliza 4 GPUs Rubin + 2 CPUs Vera.
- Design do compute tray: modular, cable-free, hose-free e fanless, com manifold interno redesenhado.
- Rack MGX de terceira geração com rack manifolds UQD08 e liquid-cooled busbars de alta corrente.

Fonte: NVIDIA Technical Blog - Vera Rubin POD e Inside the NVIDIA Vera Rubin Platform.

14. Especificação de cada compute tray ("lâmina")

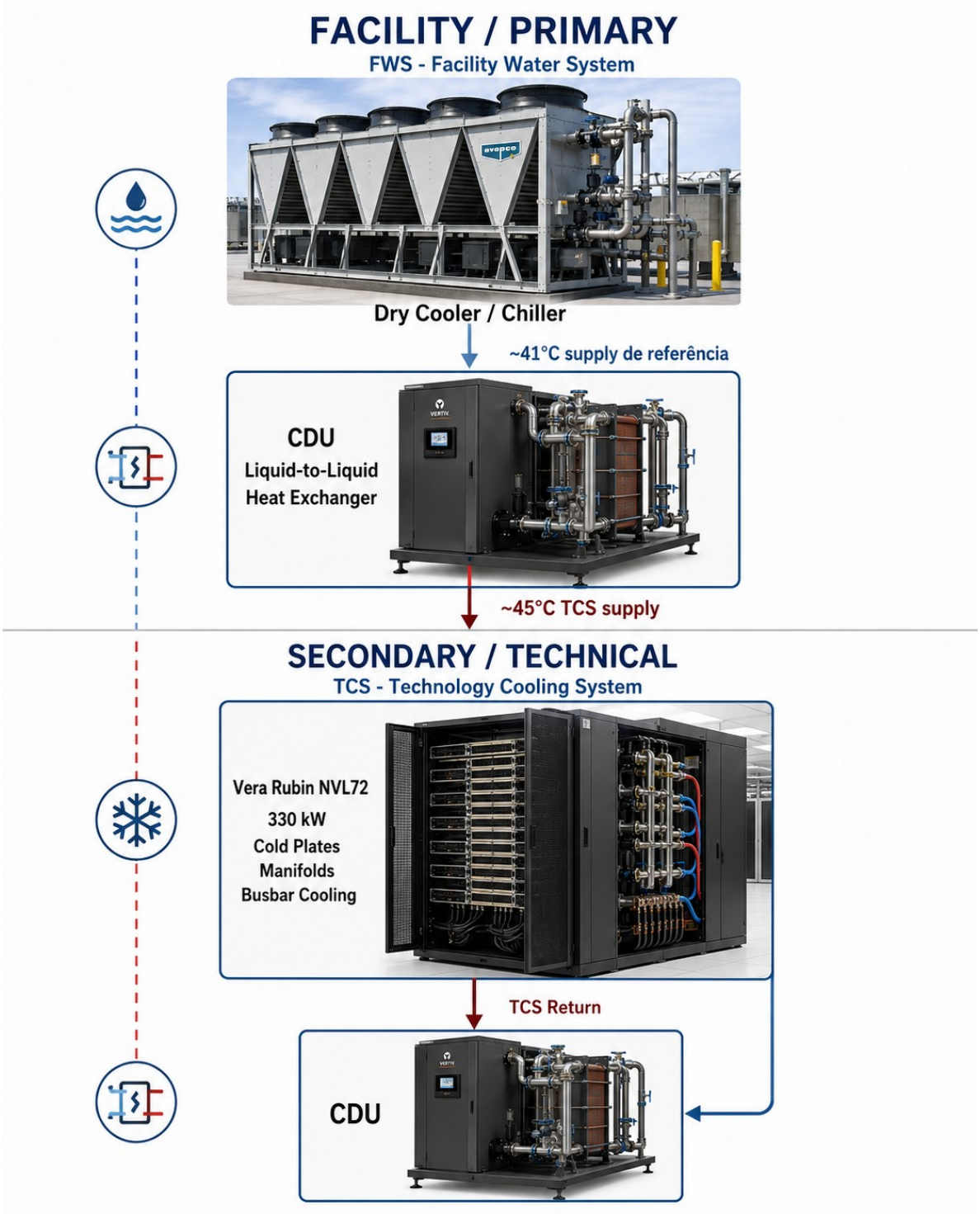
O compute tray é a unidade de serviço e de cálculo repetida 18 vezes no rack. Cada tray integra dois Vera Rubin Superchips e concentra compute, networking, cooling, gerenciamento e power delivery.

Item	Por compute tray
Vera Rubin Superchips	2
Rubin GPUs	4
Vera CPUs	2
HBM4 total	1.152 GB (4 x 288 GB)
HBM4 bandwidth	88 TB/s agregado
CPU cores	176 Olympus cores
CPU memory	até 3 TB LPDDR5X
NVFP4 inference	200 PFLOPS por tray
NVLink 6	14,4 TB/s agregado
Networking	ConnectX-9 + BlueField-4
Cooling	Direct Liquid Cooling, tray fanless

> **ATENÇÃO À POTÊNCIA POR LÂMINA**
>
> A documentação pública detalha capacidade e arquitetura do tray, mas o Max Power/TDP elétrico definitivo por compute tray deve ser obtido no Site Planning/Reference Design final. Não derive o circuito elétrico apenas dividindo 330 kW por 18.

Fonte: NVIDIA Technical Blog - Inside the NVIDIA Vera Rubin Platform: Six New Chips, One AI Supercomputer.

15. Arquitetura DLC: FWS -> CDU -> TCS -> Rack



Dois loops separados: Facility Water System (primário) e Technology Cooling System (secundário). Ilustração técnica.

16. O que são FWS, CDU e TCS?

16.1 FWS - Facility Water System

É o circuito primário do Data Center. Transporta calor entre as CDUs e a planta de rejeição térmica - normalmente chillers, dry coolers, bombas e trocadores da Central Utility Building (CUB). O FWS não deve ser confundido com o fluido técnico que circula dentro dos racks.

16.2 CDU - Coolant Distribution Unit

É a interface térmica e hidráulica entre o FWS e o TCS. Em uma CDU liquid-to-liquid, um heat exchanger mantém os dois circuitos separados. A CDU controla temperatura, vazão, pressão/differential pressure e normalmente inclui bombas redundantes, filtragem, sensores, controles e integração ao BMS.

16.3 TCS - Technology Cooling System

É o circuito secundário, de qualidade técnica, que sai da CDU e alimenta diretamente os manifolds e cold plates do equipamento de TI. Para o projeto Vera Rubin, o TCS precisa ser dimensionado pelo operating point térmico e hidráulico do rack, e não apenas pela capacidade nominal em kW da CDU.

> **REGRA PRÁTICA**
>
> FWS = água da instalação | CDU = separação e controle | TCS = fluido técnico que chega ao rack | DLC = remoção de calor diretamente no chip.

17. 45°C cooling: por que isso muda o Data Center

A NVIDIA descreve os racks MGX como projetados para operar com warm-water inlet de 45°C. Em um cenário de referência, água de facility na ordem de 41°C alimenta a CDU, que fornece aproximadamente 45°C ao rack. O objetivo é ampliar a janela para rejeição de calor sem depender continuamente de compressão mecânica.

- Maior possibilidade de dry cooling/free cooling em vários climas.
- Menor consumo de compressores e potencial redução de PUE.
- Mais do orçamento elétrico do site pode ser direcionado ao compute.
- Arquitetura liquid-to-liquid permite separar a qualidade do fluido de TI da água do facility.
- A temperatura de 45°C é um design point; temperaturas, approach e limites finais devem vir do projeto contratado.

Fonte: NVIDIA Technical Blog - Vera Rubin POD; NVIDIA DSX Facilities Infrastructure Reference Design Overview.

> **NÃO CONFUNDIR**
>
> 45°C de TCS supply não significa que o Facility Water obrigatoriamente entra na CDU a 45°C. A CDU precisa de uma diferença térmica (approach) e o design FWS/TCS deve ser verificado pelo fabricante.

18. Dimensionamento térmico e hidráulico

18.1 Carga térmica de referência

A NVIDIA DSX informa que o Cabinet TDP escala até 330 kW para Vera Rubin NVL72. Para pré-projeto de cooling, esse é o valor de referência mais consistente para dimensionar heat rejection, CDU e distribuição, até que o Site Planning final indique valores específicos de MaxP/MaxQ.

18.2 Vazão de TCS

O DSX define TCS design flow de pelo menos 1,5 LPM/kW. Portanto:

> **CÁLCULO POR RACK**
>
> 330 kW x 1,5 LPM/kW = 495 LPM = 29,7 m³/h por Vera Rubin NVL72.

Racks	Carga térmica	TCS mínimo	Vazão
1	330 kW	495 LPM	29,7 m³/h
2	660 kW	990 LPM	59,4 m³/h
4	1,32 MW	1.980 LPM	118,8 m³/h
8	2,64 MW	3.960 LPM	237,6 m³/h
16	5,28 MW	7.920 LPM	475,2 m³/h

Fonte: NVIDIA DSX Facilities Infrastructure Reference Design Overview: TCS >=1,5 LPM/kW; Cabinet TDP até 330 kW.

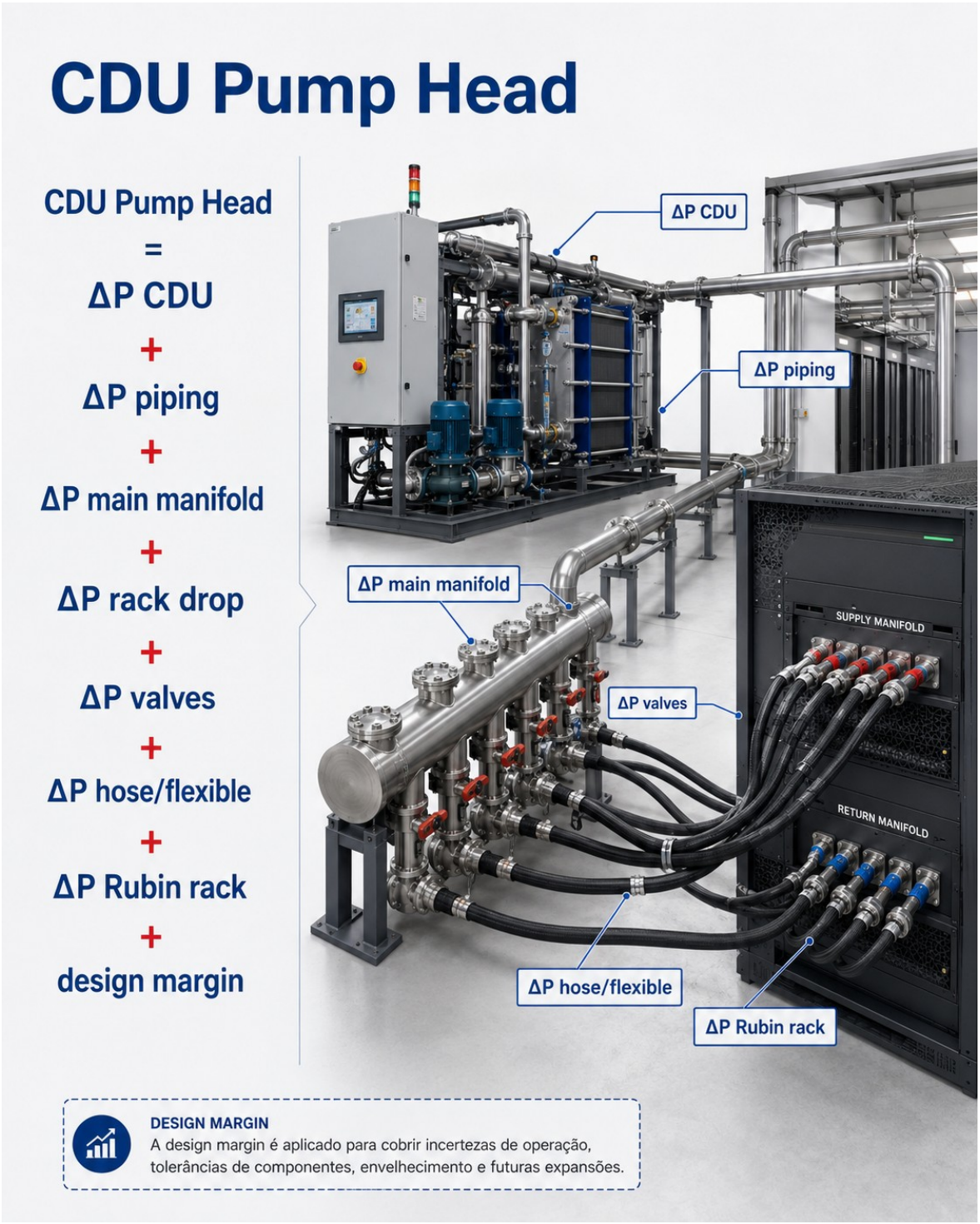
19. ΔT: transformando kW e vazão em temperaturas

O balanço térmico de um circuito de água pode ser aproximado por Q = m_dot x Cp x ΔT. Em água, 495 LPM equivalem aproximadamente a 8,25 kg/s. Para 330 kW, o ΔT teórico é da ordem de 9,6°C.

> **EXEMPLO DIDÁTICO**
>
> Se o TCS entrar a 45°C e remover 330 kW a ~495 LPM com propriedades próximas às da água, o retorno teórico ficaria próximo de 54-55°C. Esse valor é apenas um cálculo indicativo; o coolant e as temperaturas finais devem seguir o OEM/CDU.

- Maior ΔT pode reduzir a vazão exigida para a mesma carga, mas deve respeitar os limites do rack.
- Maior concentração de glicol altera densidade, viscosidade e capacidade térmica.
- A capacidade publicada de uma CDU depende do approach temperature e do operating point.
- Use as curvas do fabricante, não apenas a potência nominal de catálogo.

20. ΔP e CDU Pump Head



Elementos que compõem a perda de pressão total do loop secundário. Ilustração técnica.

21. Como interpretar ΔP

ΔP é lido "Delta P" e significa diferença de pressão. Em um circuito hidráulico, representa a perda de pressão entre dois pontos. Se o rack recebe 3,0 bar e retorna 2,2 bar, o ΔP do rack é 0,8 bar.

> **PUMP HEAD TOTAL**
>
> ΔP CDU + ΔP piping + ΔP main manifold + ΔP rack drop + ΔP valves + ΔP hose/flexible + ΔP Rubin rack + design margin.

A bomba da CDU deve fornecer vazão suficiente no Pump Head requerido. Uma CDU de muitos megawatts pode não atender ao projeto se sua curva de bomba não entregar a vazão necessária no ΔP real do circuito.

- Obter do OEM o ΔP do rack na vazão nominal e máxima.
- Calcular perdas de carga do supply e return headers.
- Somar válvulas, filtros, conexões, flexíveis, manifolds e acessórios.
- Aplicar margem coerente para tolerâncias, envelhecimento, fouling e expansão.
- Verificar NPSH, cavitação e operação das bombas em paralelo quando aplicável.

22. Manifolds, cold plates e interface do rack

Na terceira geração MGX, a NVIDIA descreve novos manifolds internos de tray, rack manifolds UQD08 e liquid-cooled busbars capazes de suportar correntes muito elevadas. O compute tray foi redesenhado para um conjunto cable-free, hose-free e fanless.

- Supply manifold distribui o coolant pelos trays.
- Return manifold coleta o fluido aquecido.
- Universal Quick Disconnects (UQD08) facilitam conexão e manutenção.
- Cold plates transferem calor diretamente de GPU/CPU e outros componentes ao fluido.
- A interface externa do rack - diâmetro, padrão de conexão, working pressure e posição - deve vir do desenho mecânico do OEM.

> **INFORMAÇÃO OBRIGATÓRIA PARA O PROJETO**
>
> Não assuma que o conector interno UQD08 é o mesmo conector externo do rack. Solicite o Rack Supply/Return Interface Drawing com quantidade de conexões, diâmetro, tipo, altura, orientação, pressão e requisitos de isolamento.

Fonte: NVIDIA Technical Blog - Vera Rubin POD; Inside the NVIDIA Vera Rubin Platform.

23. Como selecionar a CDU para Vera Rubin

A seleção não deve ser feita apenas por "MW da CDU". É necessário validar simultaneamente capacidade térmica, vazão, pressão disponível, approach, fluido, materiais, redundância, alimentação e controles.

Critério	O que validar
Capacidade térmica	>= carga dos racks no operating point e ATD definidos
Vazão secundária	>= 1,5 LPM/kW como referência DSX
Pump Head	suficiente para o ΔP total calculado
Tipo	Liquid-to-liquid na referência DSX
Redundância	N+1 em grupos de CDU
Bombas	redundância/failover e curva validada
Fluido/materiais	compatibilidade química e wetted materials
Filtragem	rating compatível com OEM
Controles	constant flow / constant ΔP conforme estratégia
Integração	BMS / DSX Exchange / Mission Control
Manutenção	isolamento, bypass e troca sem impacto indevido

EXEMPLO

Para 1 rack: 330 kW e >=495 LPM. Uma CDU nominal de 380 kW/600 LPM pode parecer suficiente, mas a decisão só é válida se ela entregar esses valores no ΔP e no approach reais do projeto.

24. NVIDIA DSX Marketplace: MP Ready x Sample Ready

O NVIDIA DSX Infrastructure Marketplace publica CDUs submetidas à CDU Self-Qualification Suite e indica o status da cadeia de fornecimento. MP Ready não é sinônimo isolado de "homologação"; é um indicador de prontidão para produção em massa, enquanto a coluna Validation Type mostra os testes técnicos realizados.

Status	Significado prático
MP Ready	Mass Production Ready: produto em estágio de fornecimento em escala.
Sample Ready	Produto disponível em estágio de amostra/qualificação, ainda sem o mesmo status de produção em massa.
Validation Type	Lista os ensaios realizados - hydraulic constant flow/DP, thermal capacity, pump failover, flow accuracy, group control etc.

- Hydraulic Test - Constant Flow.
- Hydraulic Test - Constant DP.
- Pumping Capacity.
- Thermal Test - Nominal Capacity / Low Load.
- Flow Sensor Accuracy.
- Cold Start Test.
- Pump Failover.
- Group Control.
- Wetted Materials Compatibility.

Fonte: NVIDIA DSX Infrastructure Marketplace; lista dinâmica consultada em 04/09/2026.

25. Fabricantes de CDU no ecossistema NVIDIA DSX

A lista pública é dinâmica. Os valores abaixo refletem a consulta de 04/09/2026 e devem ser reconfirmados antes da compra.

Fabricante	Modelo	Capacidade @ 4°C ATD	Flow @ 35 psi	Status
AVC	CDU1000-LTL-RW	1,2 MW	1.600 LPM	Sample Ready
Boyd	ROL2300	1,1 MW	2.600 LPM	MP Ready
Carrier	65LL	1,2 MW	2.500 LPM	Sample Ready
CoolIT	CHx1500	1,5 MW	1.950 LPM	-
Delta	RDF106CDT5192	1,0 MW	1.500 LPM	MP Ready
Delta	CDU3000	2,0 MW	3.200 LPM	-
Johnson Controls	SACDU-1050	1,0 MW	-	-
LG Electronics	LGE	600 kW	850 LPM	-
LiquidStack	L2L CDU800	800 kW	1.200 LPM	-
LiquidStack	D1PM20	2,5 MW	3.750 LPM	-
LITEON	LC-LL-WCDU-6011(S)	380 kW	600 LPM	Sample Ready
MEPPI	ME-CDU 1200	1,25 MW	-	-
Motivair	MCDU50	1,7 MW	1.136 LPM	MP Ready
Motivair	MCDU55	1,3 MW	1.616 LPM	Sample Ready
Nautilus	EcoCore FCD	3,6 MW	3.300 LPM	-

Fonte: NVIDIA DSX Infrastructure Marketplace. A capacidade publicada é condicionada ao ATD e ao operating point.

26. Estratégia de aquisição no Brasil

Para implantação no Brasil, é recomendável separar três critérios: presença local do fabricante, disponibilidade do modelo exato e status/validação do modelo no ecossistema NVIDIA. Uma empresa pode ter operação brasileira sem manter em estoque local a CDU específica do projeto Rubin.

Fabricante/ecossistema	Evidência pública no Brasil	Observação
Schneider Electric / Motivair	Portfólio brasileiro de liquid cooling e CDUs	A Schneider publica CDUs Motivair e serviços locais; validar o modelo DSX específico.
Delta Electronics	GoCool LTL no site brasileiro	GoCool-1000/1200/1500/3000; validar equivalência/status do modelo exato no DSX.
Johnson Controls / Silent-Aire	Página brasileira de Data Centers e liquid cooling	Boa presença de serviços; validar SACDU/modelo e lead time.
Carrier	Presença regional/global e CDU 65LL	Confirmar disponibilidade local e status de supply chain.
LG	Operação local	Confirmar modelo CDU, suporte e BOM para Rubin.

RECOMENDAÇÃO DE COMPRA

Emitir RFI/RFQ para pelo menos 3 fornecedores e exigir: curva de bomba, thermal map por ATD, materiais molhados, qualidade do coolant, redundância, integração BMS, FAT/SAT, peças de reposição e compromisso de suporte no Brasil.

27. Potência elétrica, kVA e consumo operacional

O DSX publica 330 kW como Cabinet TDP para Vera Rubin NVL72. Para uma aproximação de infraestrutura elétrica, kVA = kW / fator de potência. Entretanto, o dimensionamento final deve usar o electrical site planning e os limites de alimentação do rack, não apenas o TDP térmico.

Fator de potência	Cálculo	Equivalência
PF 0,95	330 kW / 0,95	347,4 kVA
PF 0,98	330 kW / 0,98	336,7 kVA
PF 0,99	330 kW / 0,99	333,3 kVA

RESERVA DE PROJETO

Para uma posição de rack, a infraestrutura elétrica deve ser confirmada pelo Site Planning final. Como pré-projeto, 330 kW representa ~347 kVA a PF 0,95, antes de considerar arquitetura A/B, derating, seletividade e requisitos de redundância.

27.1 Consumo médio

A NVIDIA não publica um "consumo médio universal" do Rubin. Training, inference, comunicação e perfis MaxP/MaxQ produzem cargas diferentes. Por isso, use cenários de utilização somente para orçamento de energia, nunca para reduzir a capacidade física instalada.

Cenário	Potência média	Energia em 730 h
50% do envelope	165 kW	120,5 MWh/mês
70%	231 kW	168,6 MWh/mês
80%	264 kW	192,7 MWh/mês
90%	297 kW	216,8 MWh/mês
100%	330 kW	240,9 MWh/mês

Fonte: Cenários de utilização são cálculos de planejamento, não especificação NVIDIA de consumo médio.

28. BMS, Mission Control e observabilidade

O cooling de uma AI Factory precisa ser observável em tempo real. A documentação DSX publica um catálogo de pontos para integração entre BMS e o ecossistema NVIDIA/DSX Exchange.

Ponto	Uso operacional
CDULiquidSupplyTemperature	Temperatura TCS supply
CDULiquidReturnTemperature	Temperatura TCS return
CDULiquidDifferentialPressure	ΔP do secundário
CDULiquidFlow	Vazão em LPM
CDULiquidSystemPressure	Pressão do sistema/return
Leak detection / alarms	Eventos de vazamento e anormalidade
CDU availability/status	Disponibilidade e estado operacional

- Alarmes por temperatura, vazão, ΔP e pressão fora da faixa.
- Trend histórico para detectar fouling, filtro saturado e degradação de bomba.
- Interlock e lógica de isolamento conforme matriz de causa e efeito.
- Integração dos eventos de facility com Mission Control/DSX para operação coordenada de power e cooling.

Fonte: NVIDIA Mission Controls to BMS Data Catalog; NVIDIA DSX Exchange.

29. RFI técnico - 12 informações obrigatórias

1	**Quantidade exata de Vera Rubin NVL72** Define a escala total e impacta potência, cooling, vazão, CDU e distribuição. Também determina redundância e expansão.
2	**Max-P/TDP confirmado por rack** Base para energia e carga térmica. Deve ser o valor oficial da configuração e do perfil operacional contratado.
3	**TCS design flow por rack** Vazão nominal necessária para remover a carga térmica e dimensionar bombas, headers e CDU.
4	**TCS minimum/maximum flow** Limites operacionais para evitar subfluxo, excesso de vazão e operação fora do envelope.
5	**TCS supply/return temperatures** Determina ΔT, approach da CDU, eficiência térmica e capacidade de rejeição.
6	**ΔP do rack na vazão nominal e máxima** Necessário para calcular Pump Head e selecionar bombas e válvulas.
7	**Working pressure e maximum allowable pressure** Define limites de segurança do circuito interno, hoses, válvulas e conexões.
8	**Coolant / water-quality specification** Especifica fluido, química, pH, condutividade, filtragem e materiais compatíveis.
9	**Rack Supply/Return connection type e dimensions**

	Define diâmetro, tipo, posição, quantidade e padrão das interfaces hidráulicas.
10	**CDU approved/validated models** Permite alinhar a seleção ao NVIDIA DSX Marketplace e à configuração OEM.
11	**BMS/Mission Control telemetry requirements** Define sensores, protocolos, alarmes e pontos necessários para integração IT/OT.
12	**Site Planning Guide / Reference Design final** Documento de autoridade para validar o projeto executivo antes da implantação.

Pacote mínimo esperado na resposta da RFI

- Tabela preenchida com valores nominais, mínimos e máximos e respectivas unidades.
- Curvas hidráulicas do rack e da CDU para os operating points relevantes.
- Desenhos mecânicos das interfaces Supply/Return e requisitos de instalação.
- Especificação química do coolant e materiais molhados compatíveis.
- Lista de alarmes, sensores, protocolos e pontos de integração BMS/DSX.
- Documentos finais de Site Planning, FAT/SAT e critérios de aceite.

30. Exemplo de pré-dimensionamento - 8 racks Rubin

Este exemplo serve apenas para sizing inicial. O projeto executivo deve fechar ΔP, curvas das bombas, topologia hidráulica, diversificação, redundância, expansão e envelope elétrico.

Parâmetro	Valor de pré-projeto
Racks Vera Rubin	8
GPUs Rubin	576
CPUs Vera	288
Cabinet TDP agregado	2,64 MW
TCS design flow mínimo	3.960 LPM
Vazão equivalente	237,6 m³/h
Cooling architecture	Liquid-to-liquid CDU group
CDU redundancy	N+1
TCS supply design point	45°C class
BMS	Supply/return temp, flow, ΔP, pressure, alarms

> **SELEÇÃO DA CDU**
>
> Não use apenas 2,64 MW / capacidade nominal da CDU. Verifique também se o grupo de CDUs entrega >=3.960 LPM no Pump Head total e no approach térmico do projeto, mantendo N+1.

Para clusters maiores, a distribuição deve ser concebida como uma planta hidráulica de alta capacidade, com headers principais, isolamento por rack/ramal, balanceamento, instrumentação e estratégia de manutenção concorrente.

31. Comissionamento e testes

Um sistema DLC de alta densidade precisa ser comissionado como infraestrutura de missão crítica. O FAT da CDU e o SAT do sistema devem validar não só a potência térmica, mas o comportamento hidráulico e de controle.

- Flushing e limpeza das tubulações antes da conexão ao rack.
- Análise química do coolant e verificação de materiais compatíveis.
- Teste hidrostático conforme pressão admissível e procedimento do fabricante.
- Verificação de vazão por ramal e balanceamento dos racks.
- Teste de ΔP em carga nominal e máxima.
- Teste de pump failover e perda de alimentação A/B.
- Teste de válvulas de isolamento e bypass.
- Leak detection e causa/efeito no BMS.
- Validação de sensores de temperatura, pressão e vazão.
- Teste de perda do FWS, recuperação e sequência de restart.
- Teste de operação N+1 com uma CDU indisponível.
- Registro de baseline para operação e manutenção preditiva.

> **CRITÉRIO DE ACEITE**
>
> A entrega deve incluir curvas as-built, parâmetros de controle, setpoints, matriz de alarmes, water-quality baseline, lista de peças críticas e procedimento de resposta a vazamentos.

32. Principais riscos de projeto

Risco	Impacto
CDU escolhida apenas por MW	Pode faltar vazão ou Pump Head no operating point real.
ΔP do rack desconhecido	Impossibilita dimensionar corretamente as bombas.
Conector do rack assumido	Pode gerar incompatibilidade mecânica e retrabalho.
Water quality indefinida	Risco de corrosão, fouling e falha de cold plates.
Sem N+1 real	Manutenção ou falha de CDU pode afetar múltiplos racks.
Sem isolamento por rack	Aumenta blast radius de manutenção/vazamento.
Sem BMS/telemetria adequada	Perda de detecção precoce de degradação.
Usar consumo médio para sizing	Cria risco de insuficiência em MaxP.
Ignorar ATD/approach	Capacidade nominal da CDU pode não existir nas temperaturas do projeto.
Não validar disponibilidade local	Risco de lead time, spare parts e suporte insuficiente no Brasil.

33. Glossário rápido

Termo	Definição
DLC	Direct Liquid Cooling - refrigeração direta por líquido.
FWS	Facility Water System - circuito primário da instalação.
TCS	Technology Cooling System - circuito secundário que atende o equipamento de TI.
CDU	Coolant Distribution Unit - interface FWS/TCS com trocador, bombas e controles.
ΔP	Delta P - diferença/perda de pressão entre dois pontos.
ΔT	Delta T - diferença de temperatura entre supply e return.
Pump Head	Pressão/altura manométrica que a bomba precisa fornecer para vencer perdas do circuito.
Cold plate	Trocador em contato térmico com chip/componente.
Manifold	Coletor/distribuidor de supply e return.
UQD08	Universal Quick Disconnect usado na arquitetura MGX de terceira geração.
ATD	Approach Temperature Difference do heat exchanger/CDU.
MP Ready	Mass Production Ready - status de supply chain no NVIDIA DSX Marketplace.
Sample Ready	Estágio de amostra/qualificação no Marketplace.
BMS	Building Management System.
DSX	Arquitetura/ecossistema NVIDIA para AI Factory, incluindo facilities, hardware e IT/OT.

34. Conclusão executiva - da GPU à AI Factory

A principal inovação da Vera Rubin é sistêmica. O valor não está em uma GPU isolada, mas na combinação de compute, HBM4, NVLink 6, rede scale-out, storage, segurança, energia e Direct Liquid Cooling operando como uma única fábrica de computação.

Para empresas e data centers, a oportunidade está em transformar energia e infraestrutura física em capacidade útil de IA: treinamento de modelos, inferência agentiva, GPUaaS, pesquisa científica, simulação e serviços soberanos. O desafio também é sistêmico: elétrica, hidráulica, conectividade, software, observabilidade, operação e suporte precisam evoluir no mesmo ritmo do acelerador.

> **A PERGUNTA CORRETA**
>
> Em vez de perguntar apenas quantas GPUs cabem no rack, o projeto deve responder quantos tokens úteis, agentes concluídos, simulações e workloads de negócio o data center consegue produzir por megawatt - com disponibilidade, segurança e previsibilidade operacional.

35. Referências técnicas e créditos

NVIDIA Vera Rubin NVL72 - especificações oficiais: https://www.nvidia.com/pt-br/data-center/vera-rubin-nvl72/

NVIDIA DGX Vera Rubin NVL72: https://www.nvidia.com/pt-br/data-center/dgx-vera-rubin-nvl72/

NVIDIA DSX Facilities Infrastructure Reference Design Overview: https://docs.nvidia.com/dsx/facilities-infra/reference-design-overview

NVIDIA DSX Documentation: https://docs.nvidia.com/dsx

NVIDIA DSX Infrastructure Marketplace - CDU: https://marketplace.nvidia.com/en-us/enterprise/dsx-infrastructure/

NVIDIA Technical Blog - Vera Rubin POD: https://developer.nvidia.com/blog/nvidia-vera-rubin-pod-seven-chips-five-rack-scale-systems-one-ai-supercomputer/

NVIDIA Technical Blog - Inside the Vera Rubin Platform: https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/

NVIDIA Mission Controls to BMS Data Catalog: https://docs.nvidia.com/datacenter/dsx/bms-datacatalog-interactive.html

NVIDIA - CoreWeave GB200 NVL72 production deployment: https://blogs.nvidia.com/blog/coreweave-grace-blackwell-gb200-nvl72/

NVIDIA - Oracle OCI GB200 NVL72 deployment: https://blogs.nvidia.com/blog/oracle-cloud-infrastructure-blackwell-gpus-agentic-ai-reasoning-models/

NVIDIA Case Study - NeoSpace / Finance: https://www.nvidia.com/pt-br/case-studies/neospace/

NVIDIA - Blackwell / Mixture-of-Experts deployments: https://blogs.nvidia.com/blog/mixture-of-experts-frontier-models/

NVIDIA - GB300 deployments and AI Infrastructure in America: https://nvidianews.nvidia.com/news/nvidia-partners-ai-infrastructure-america

NVIDIA - Vera Rubin worldwide deployments: https://blogs.nvidia.com/blog/vera-rubin/

NVIDIA - Rubin platform launch and customer plans: https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer

Schneider Electric Brasil - Liquid Cooling / Motivair: https://www.se.com/br/pt/work/products/product-launch/liquid-cooling-products/

Delta Electronics Brasil - GoCool Liquid Cooling: https://delta-electronics.com.br/infraestrutura-para-data-center/resfriamento-liquido-e-a-ar/

Johnson Controls Brasil - Data Centers: https://www.johnsoncontrols.com.br/industries/data-centers

Conteúdo incorporado do e-book EnQ Digital - NVIDIA Vera Rubin: arquitetura Rubin GPU, Vera CPU, HBM4, NVLink 6, rede, armazenamento, segurança e galeria de imagens reais.

NVIDIA - Vera Rubin NVL72

NVIDIA Technical Blog - Inside NVIDIA Rubin GPU Architecture

NVIDIA Newsroom - Vera Rubin platform

Fonte: Créditos de imagens reais incorporadas: NVIDIA Corporation; Geeknetic; Pisapapeles; Data Center Dynamics, conforme créditos do e-book EnQ Digital de origem.

DATA DE CORTE

Conteúdo técnico e lista de fornecedores verificados em 04/09/2026. O NVIDIA DSX Marketplace é dinâmico; revalide status, capacidade, flow e disponibilidade no momento da contratação.

Documentos NVIDIA que devem ser solicitados ao fabricante/representante

- NVIDIA Vera Rubin NVL72 Reference Design - NVOnline #1151654.
- NVIDIA DSX - Vera Rubin Facilities Infrastructure Reference Design - NVOnline #1145739.
- NVIDIA DSX Facilities Infrastructure Design Guide - NVOnline #1152370.
- Site Planning / Power & Cooling Specification final da configuração adquirida.
- Rack Supply/Return Interface Drawing e coolant/water-quality specification.