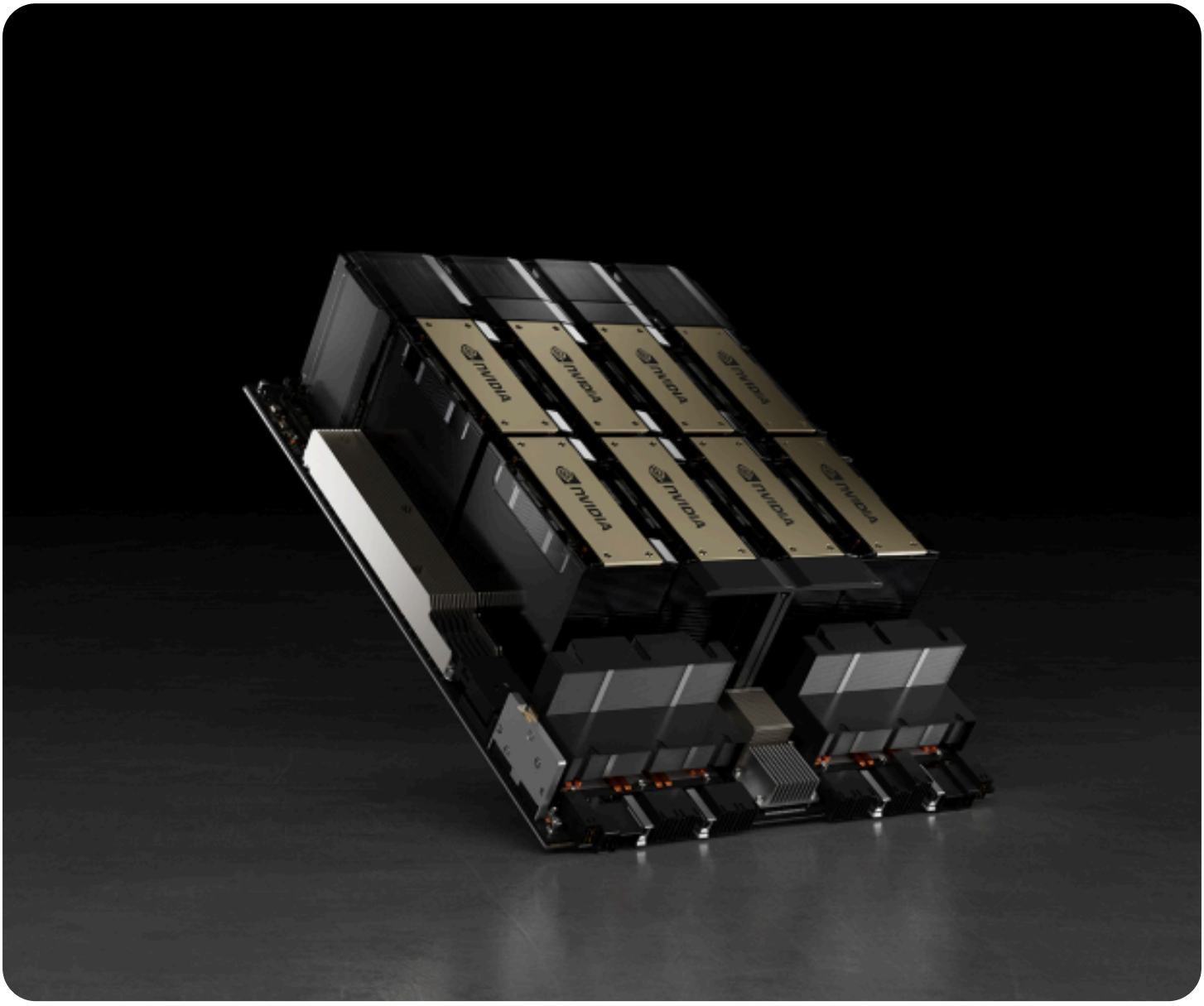
JAYLAN
SOLUTIONS
INTERNATIONAL

NEW INVENTORY LISTING

80X GPU-NVHGX-H100-494 - 94GB - LIQUID COOLING
INVENTORY CODE: GWPRCRI
LOCATION: ASIA



JAYLAN
SOLUTIONS
INTERNATIONAL

WHY 94 GB?

NVL PARTS EXPOSE AN EXTRA 14 GB HBM3 STACK (HBM3 6-HI) FOR MODEL-PARALLEL INFERENCE, AT A SMALL CLOCK TRADE-OFF (HENCE THE ≈ 15 % LOWER TFLOPS THAN THE 80 GB SXM5 BINS).

ARCHITECTURE HIGHLIGHTS - BOARD-LEVEL SPECIFICATIONS

- **TRANSFORMER ENGINE + FP8:** DOUBLES EFFECTIVE MATRIX THROUGHPUT FOR LLMS; FP8 → FP16 MIXED-PRECISION HANDLED IN HARDWARE. NVIDIA
- **FOURTH-GEN NVLINK (ON-BOARD) + NVLINK-NETWORK:** 900 GB/S ON CARD; OPTIONAL OSFP CABLES REACH EXTERNAL NVSWITCH3 FABRICS (256-GPU PODS). NVIDIA DEVELOPERNVIDIA DEVELOPER
- **DPX DYNAMIC-PROGRAMMING INSTRUCTIONS:** UP TO 40× CPU SPEED-UP FOR GRAPH/SEQUENCE ALGORITHMS (GENOMICS, ROUTING) ON HOPPER. NVIDIA
- **LIQUID COOLING DENSITY:** THE 494 LC VERSION CAN STACK 4× BOARDS IN 1U SLEDS (LENOVO SD665-N V3) GIVING 16 GPUS / 1U, 112 GPUS IN 7 U. LINKAS.RU

SOFTWARE & DRIVER STACK

- CUDA 12+, CUDNN 9, NCCL 2.20+, TENSORRT-LLM 0.9, DGX OS 6.
- HPE & CRAY FIRMWARE BUNDLES FOR 94 GB BOARDS ARE PUBLISHED (R535 BRANCH). HPE SUPPORTHPE SUPPORT
- DCGM 3.3 EUD NOW VALIDATES THE 94 GB SKU EXPLICITLY.

JAYLAN
SOLUTIONS
INTERNATIONAL

BOARD-LEVEL SPECIFICATIONS (REDSTONE-NEXT HGX H100 4-GPU, 94 GB)

Item	Per GPU	4-GPU board aggregate
Architecture	NVIDIA **Hopper H100 NVL** (TSMC 4N)	—
FP8 Tensor	3.341 PFLOPS	**13.36 PFLOPS**
FP16 / BF16 Tensor	1.671 PFLOPS	6.684 PFLOPS
TF32 Tensor	0.835 PFLOPS	3.34 PFLOPS
FP32	60 TFLOPS	240 TFLOPS
FP64	30 TFLOPS	120 TFLOPS
HBM3 capacity	**94 GB**	**376 GB**
HBM3 BW	3.9 TB/s	15.6 TB/s
NVLink 4 lanes	18 / GPU = 900 GB/s	Fully cross-barred mesh on board
Board power	≈ 2.8 kW (700 W × 4)	—
Form-factor	16-layer baseboard, 4 × SXM5 modules, 1× OCP edge connector	398 × 215 mm, ≈ 11 kg (without cold-plate manifold)

JAYLAN
SOLUTIONS
INTERNATIONAL

ARCHITECTURE HIGHLIGHTS

TRANSFORMER ENGINE + FP8: DOUBLES EFFECTIVE MATRIX THROUGHPUT FOR LLMS; FP8 → FP16 MIXED-PRECISION HANDLED IN HARDWARE. NVIDIA

FOURTH-GEN NVLINK (ON-BOARD) + NVLINK-NETWORK: 900 GB/S ON CARD; OPTIONAL OSFP CABLES REACH EXTERNAL NVSWITCH3 FABRICS (256-GPU PODS). NVIDIA DEVELOPERNVIDIA DEVELOPER

DPX DYNAMIC-PROGRAMMING INSTRUCTIONS: UP TO 40× CPU SPEED-UP FOR GRAPH/SEQUENCE ALGORITHMS (GENOMICS, ROUTING) ON HOPPER. NVIDIA

LIQUID COOLING DENSITY: THE 494 LC VERSION CAN STACK 4× BOARDS IN 1U SLEDS (LENOVO SD665-N V3) GIVING 16 GPUS / 1U, 112 GPUS IN 7 U.

ARCHITECTURE HIGHLIGHTS

Mode	GPU Power cap	Board inlet (45 °C coolant)	Notes
Training	700 W	2.8 kW	Full clocks, FP8/FP16
Mixed	500 W	2.0 kW	Typical LLM inference
Eco	400 W	1.6 kW	Data-center PUE optimisation

LIQUID-COOL COLD PLATES DROP △T BY ~15 K OVER THE 700 W AIR-COOLED 80 GB BOARD, ENABLING SUSTAINED TURBO BINS.

JAYLAN
SOLUTIONS
INTERNATIONAL

TARGET WORKLOADS & CLUSTER DESIGN TIPS

- **>70 B-PARAMETER LLM INFERENCE** IN SINGLE NODE THANKS TO 376 GB HBM3 AND NVLINK PEER-TO-PEER.
- **MULTI-NODE TRAINING:** LINK UP TO 64 GPUS IN-RACK (NVSWITCH3) OR 256 GPUS ACROSS RACKS WITH NVLINK-NETWORK OSFP CABLES.
- **HPC:** FP64 TENSOR ACCELERATION BENEFITS DENSE-MATRIX (CFD, WEATHER) CODES.
- **DATA-CENTRIC AI:** PAIR WITH BLUEFIELD-3 DPU OR CONNECTX-7 FOR 400 GB/S ROCE; NUMA BALANCE VIA PCIE 5.0 X16 CPU-GPU.

KEY DIFFERENCES VS. OTHER H100 OPTIONS

SKU	Memory	Cooling	FP8 perf	Best for
H100 PCIe 80 GB	80 GB	Air	2.0 PFLOPS	Scale-out inference, drop-in accelerator
HGX H100 4×94 GB (this)	**94 GB**	**Liquid**	**3.34 PFLOPS**	Dense LLM inference/training nodes
HGX H100 8×80 GB	80 GB	Air/Liquid	3.96 PFLOPS	Training, mainstream clusters
HGX H200 4×141 GB	141 GB	Liquid	4.7 PFLOPS	Giant-model inference w/o off-chip RAM

JAYLAN
SOLUTIONS
INTERNATIONAL

TAKE-AWAYS FOR ARCHITECTS & BUYERS

- **MEMORY FOOTPRINT MATTERS:** 94 GB SKUS LET YOU RUN 70–90 B-PARAM LLMS PER GPU W/OUT MODEL-PARALLEL SHARDING.
- **NVLINK-NETWORK IS YOUR FABRIC:** BUDGET FOR 400 G OSFP OPTICS AND NVSWITCH3 IF YOU PLAN TO SCALE PAST 4 BOARDS.
- **COOLING IS NON-NEGOTIABLE:** ENSURE 45 °C COOLANT SUPPLY, 0.20 L/S PER GPU; OTHERWISE DERATE CLOCKS.
- **LEAD-TIME PLANNING:** ALLOCATIONS SHIP QUARTERLY; PLACE NCNR POS 6-9 MONTHS AHEAD FOR 2026 CAPACITY.

JAYLAN
SOLUTIONS
INTERNATIONAL

INTERESTED IN SUBMITTING A PURCHASE ORDER?

FEEL FREE TO REPLY WITH YOUR SPEC SHEET OR PROJECT GOALS. WE'LL SEND TAILORED OPTIONS WITH PRICING AND DELIVERY SCHEDULES.

EMAIL TO: MATHEW YATES

PRESIDENT | CO-FOUNDER

MATHEW.YATES@JAYLANSOLUTIONS.COM
971-325-8537

ACCESS SURPLUS INVENTORY:
HTTPS://WWW.JAYLANSOLUTIONS.COM/ACCESS-INVENTORY