

Kunlunxin R480-X8

8x R200-8F

XPU-R

rack

OVERVIEW

1.02 FP16 PFLOPS Dense	0.3 TB Aggregate Memory	4.1 TB/s Aggregate Bandwidth	— Aggregate Power
----------------------------------	-----------------------------------	--	-----------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP32	0.3	32.0	—
FP16	1.0	128.0	—
INT8	2.0	256.0	—
INT16	1.0	128.0	—

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	R200-8F
Count	8
Memory per GPU	32 GB
Bandwidth per GPU	512 GB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	160 W
FP16 Efficiency	—
Aggregate Bandwidth	4.1 TB/s

SYSTEM DETAILS

Vendor: Kunlunxin	Family: R480	Architecture: XPU-R
Form Factor: rack	Release Year: —	Process Node: 7nm

INTERCONNECT

Chip-to-chip interconnect at 200 GB/s

NOTES

The Kunlunxin R480-X8 is an eight-card accelerator group built from the 32 GB R200-8F, the second-generation XPU-R part. Kunlunxin rates it at 2,048 TOPS of INT8, 1,024 TOPS of INT16, 1,024 TFLOPS of FP16 and 256 TFLOPS of FP32, with 256 GB of GDDR6 across the eight cards, 4,096 GB/s of aggregate memory bandwidth and 200 GB/s of chip-to-chip interconnect, on a 7nm process with ECC throughout. Kunlunxin lists it for both inference and training. Every figure is published per-card times eight on Kunlunxin's own page, and each per-card term matches the R200-8F exactly, which is what identifies the eight cards as the 32 GB variant rather than the 16 GB R200. Kunlunxin labels none of these figures dense or sparse and publishes no sparse rate. It also publishes no power

figure for the group; summing eight 160 W cards would count only the accelerators and exclude CPUs, fans and supply losses, so total system power is left blank rather than understated.