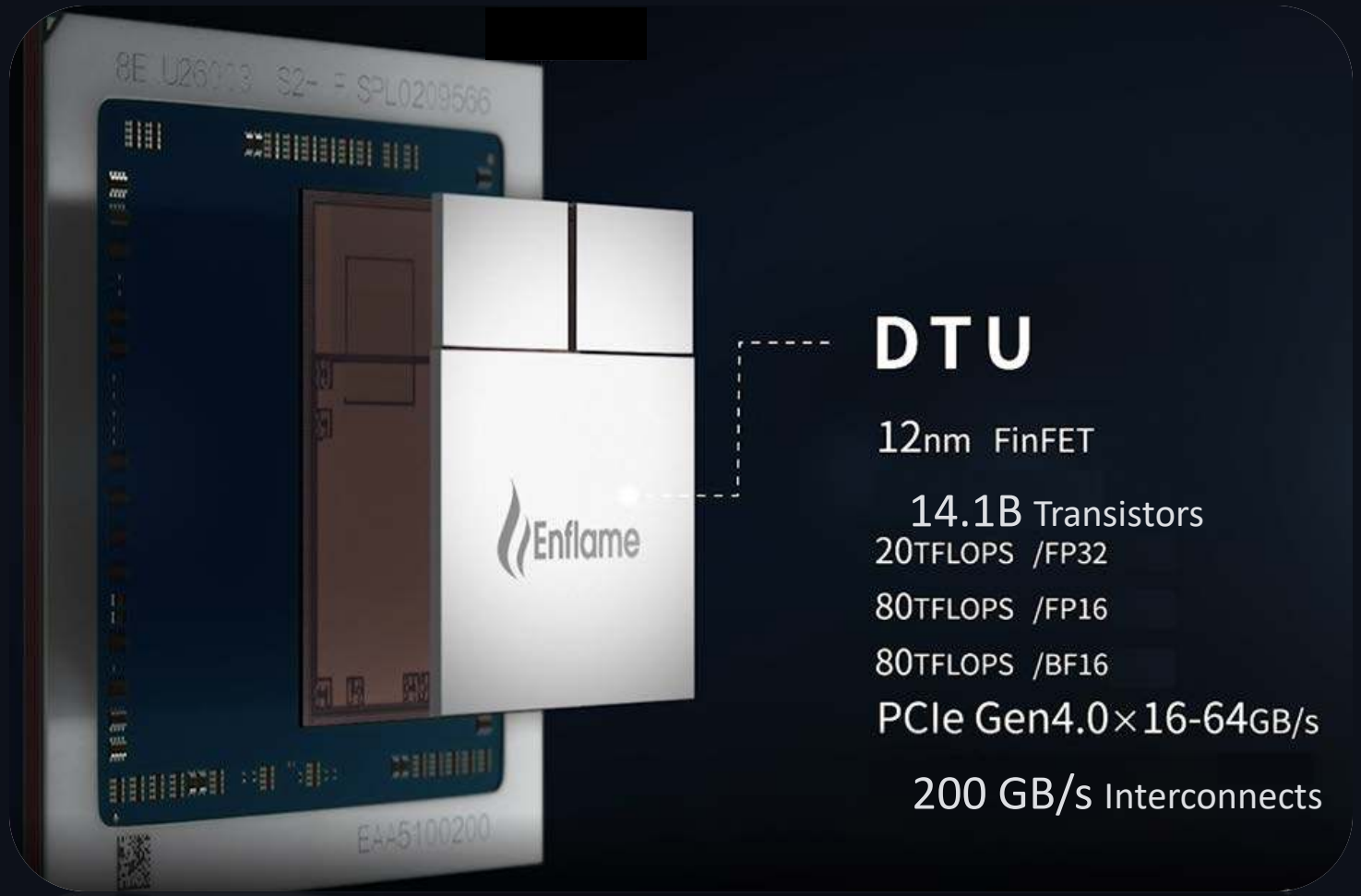
AI Compute Chip from Enflame

Hot Chips 33

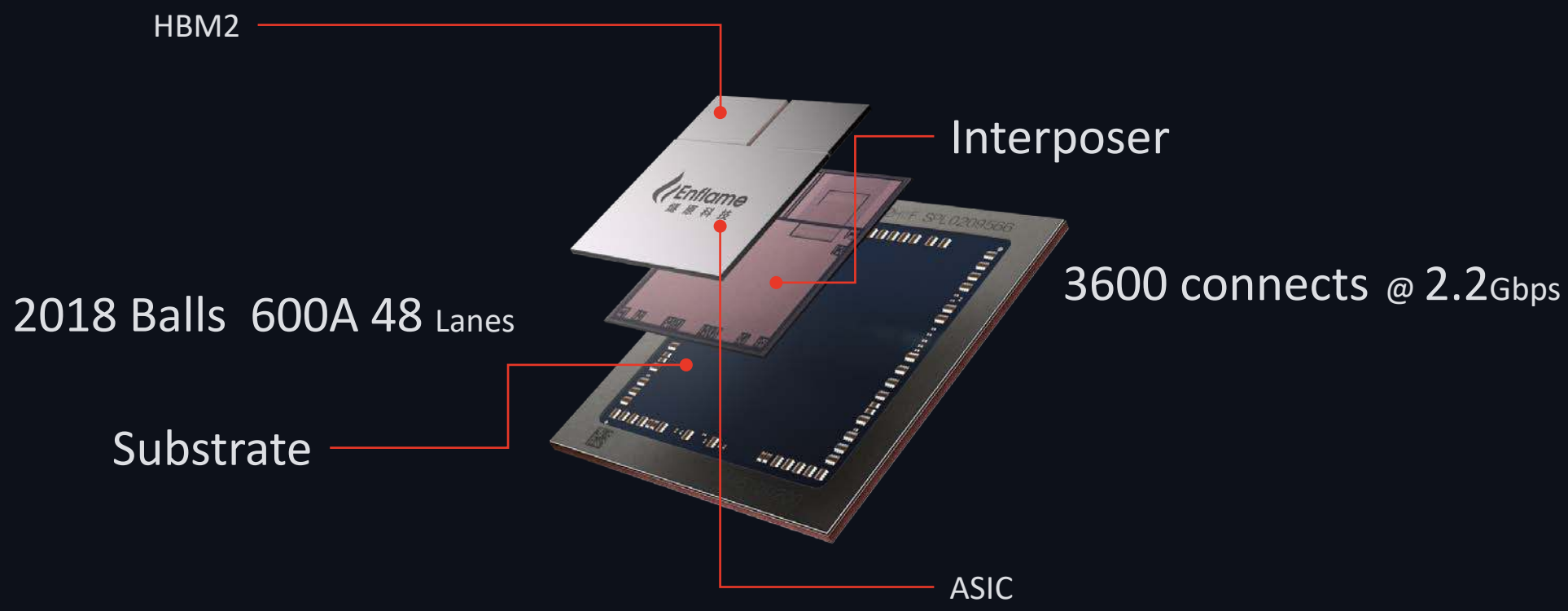
Ryan Liu / Chuang Feng

August 2021

DTU 1.0

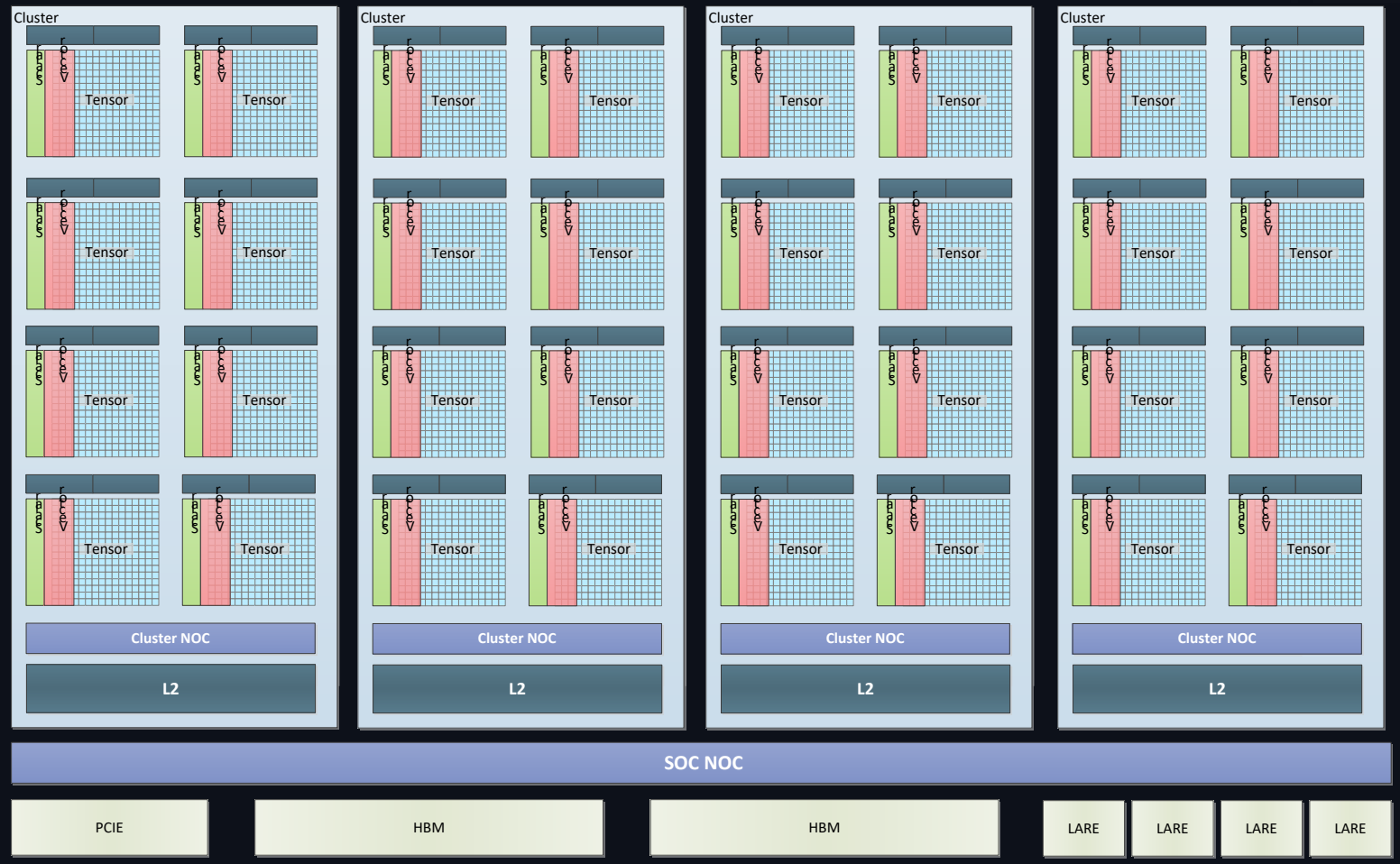


DTU 1.0 Package



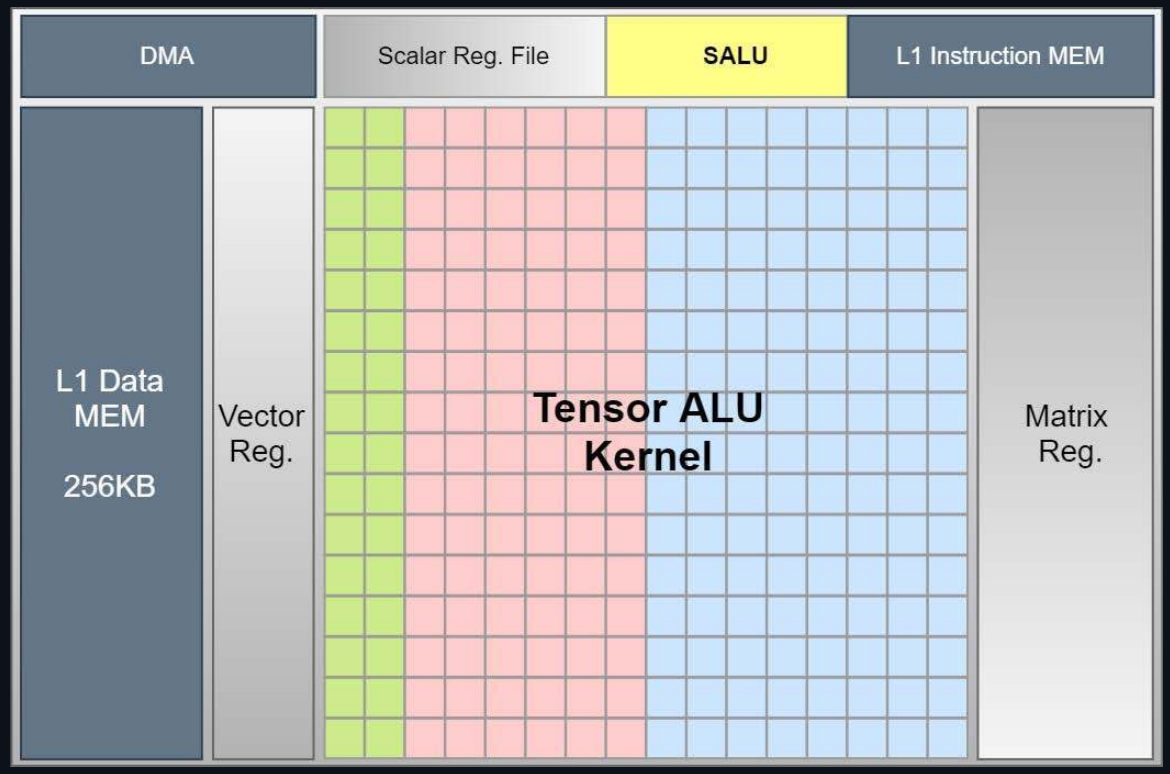
DTU 1.0 SOC

- 32 AI compute core, 4 clusters
- 40 Data transfer engines
- 4 High speed interconnects
- 2 HBM2 providing 512GB/s bandwidth



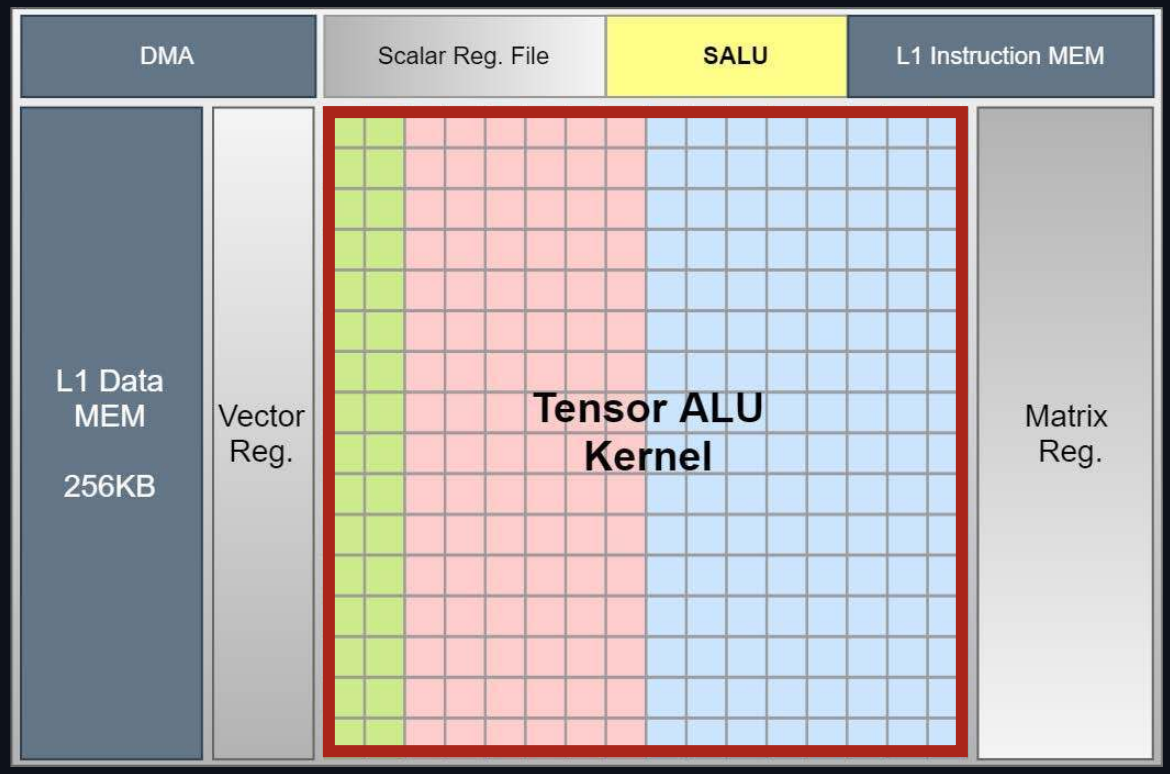
GCU-CARE 1.0

- Compute capability: 20 TFLOPS@FP32
- Bus width: 1024-bit
- Full precision support
- Fully programmable VLIW



GCU-CARE 1.0

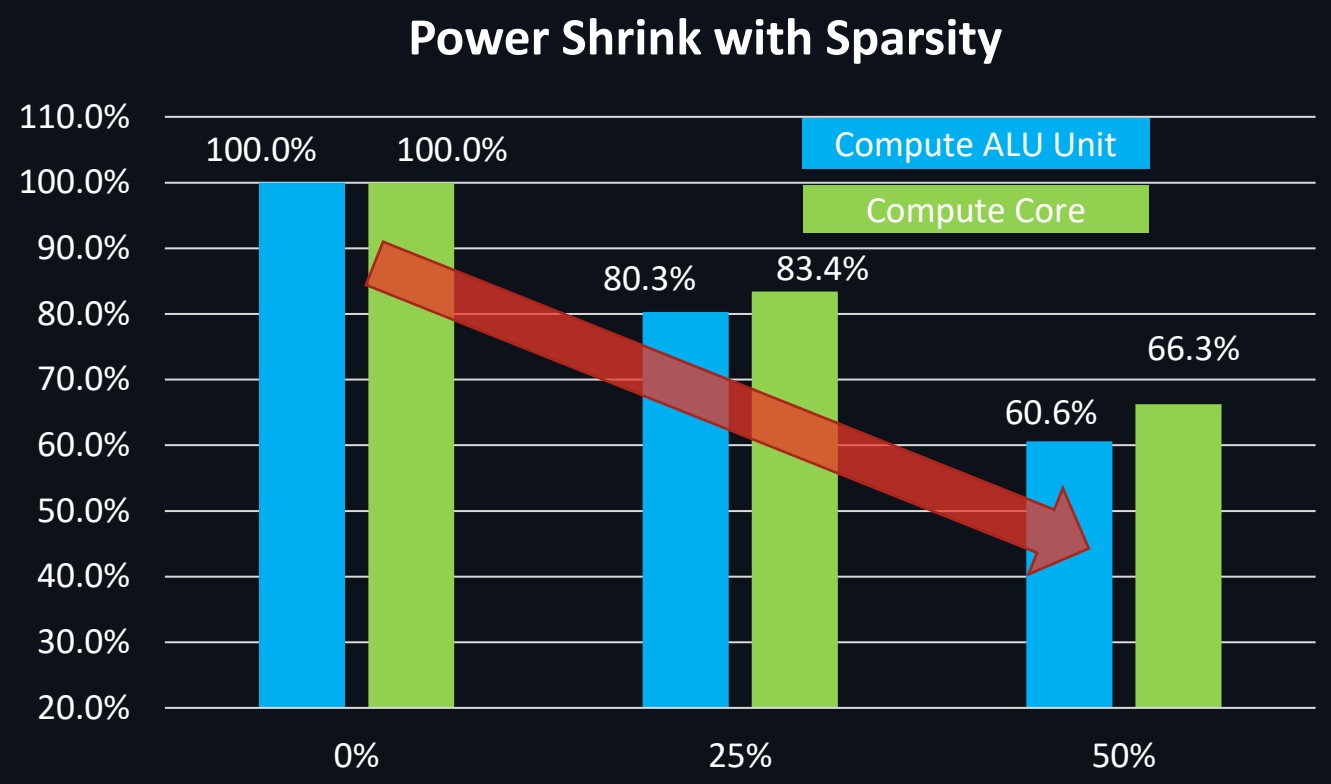
- 256 Tensor compute kernel
- Each kernel
  - Supports 1x 32-bit MAC
  - Supports 4x 16-bit/8-bit MAC
  - Supports full precision
  - Supports mixed precision
- Reuse data and bandwidth between kernels



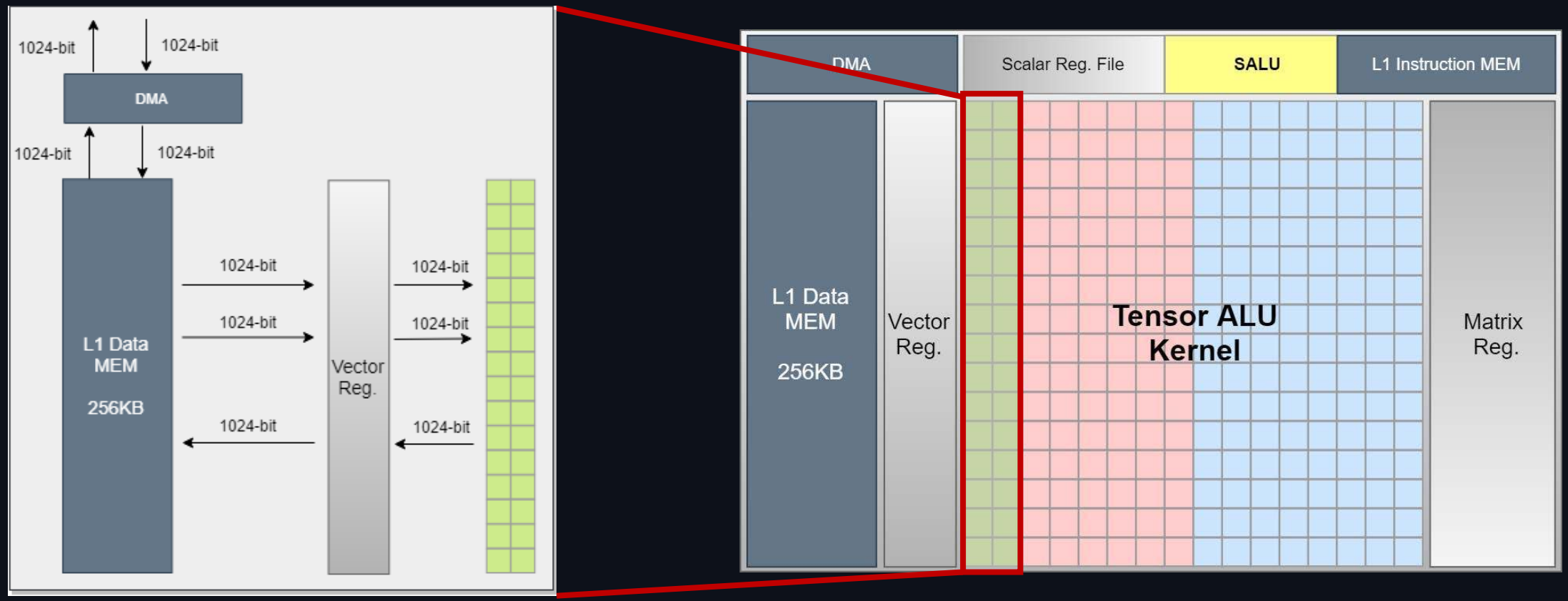
GCU-CARE 1.0

Performance Spec

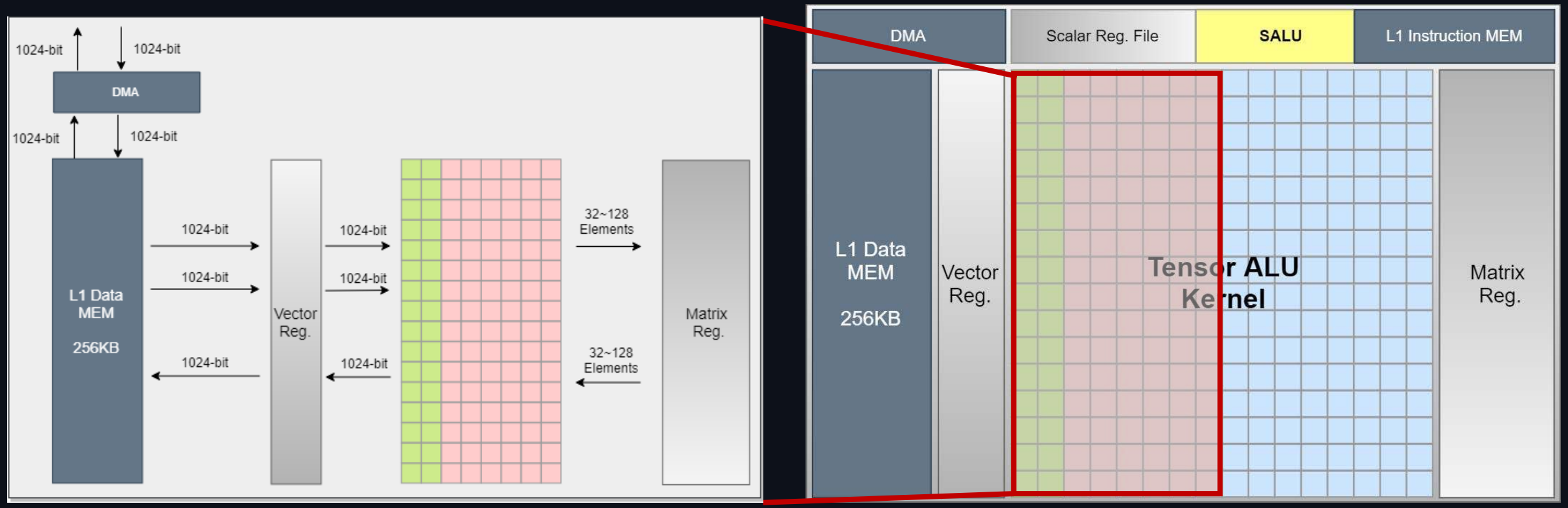
Data Type	Single Core	DTU 1.0
FP32	256 MAC	20 TFLOPS
FP16	1024 MAC	80 TFLOPS
BF16	1024 MAC	80 TFLOPS
INT32	256 MAC	20 TOPS
INT16	1024 MAC	80 TOPS
INT8	1024 MAC	80 TOPS



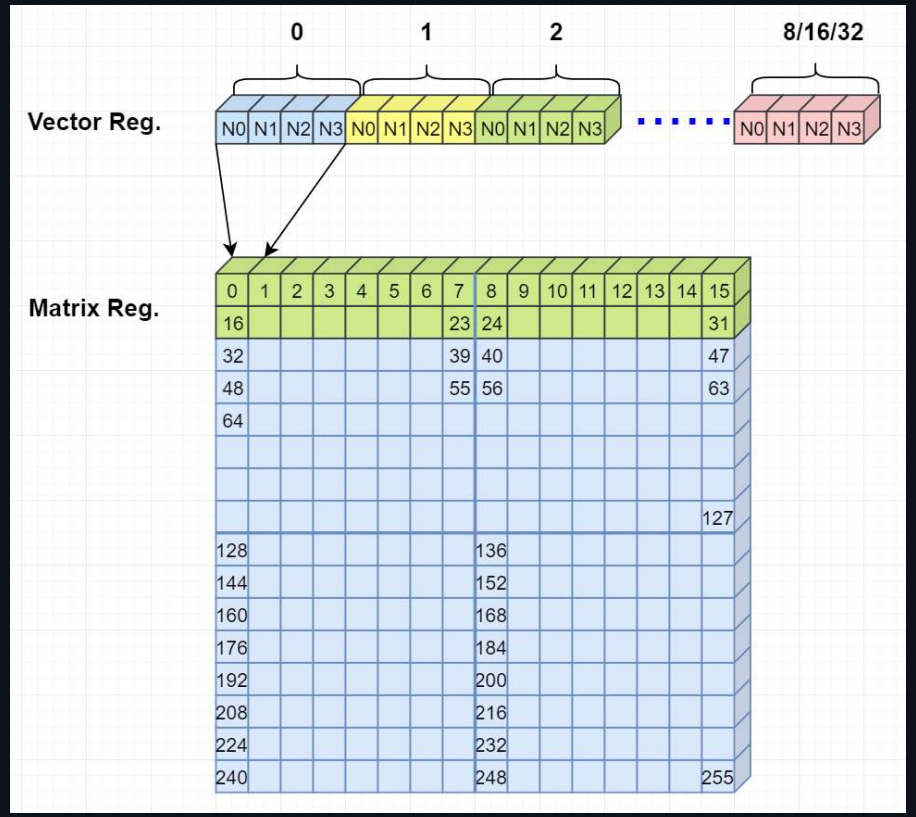
32 Kernels for 1024-bit General Vector

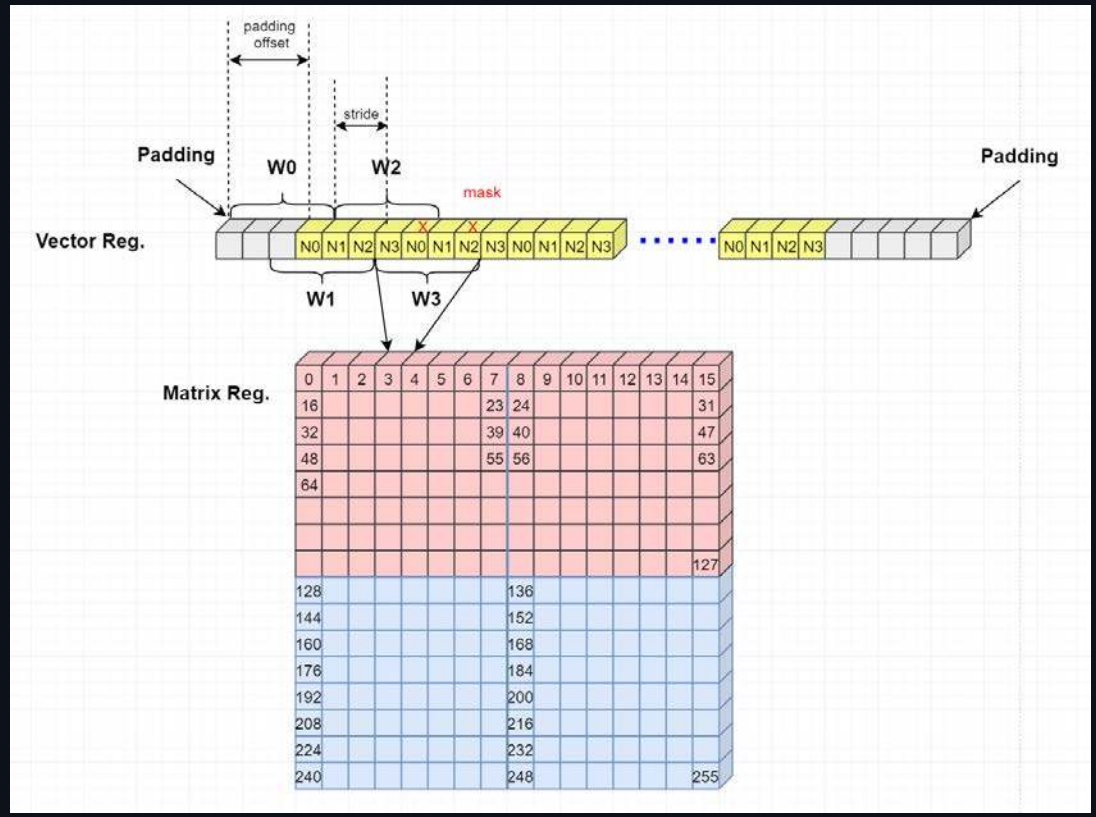


32/64/128 Kernels for Vector MAC

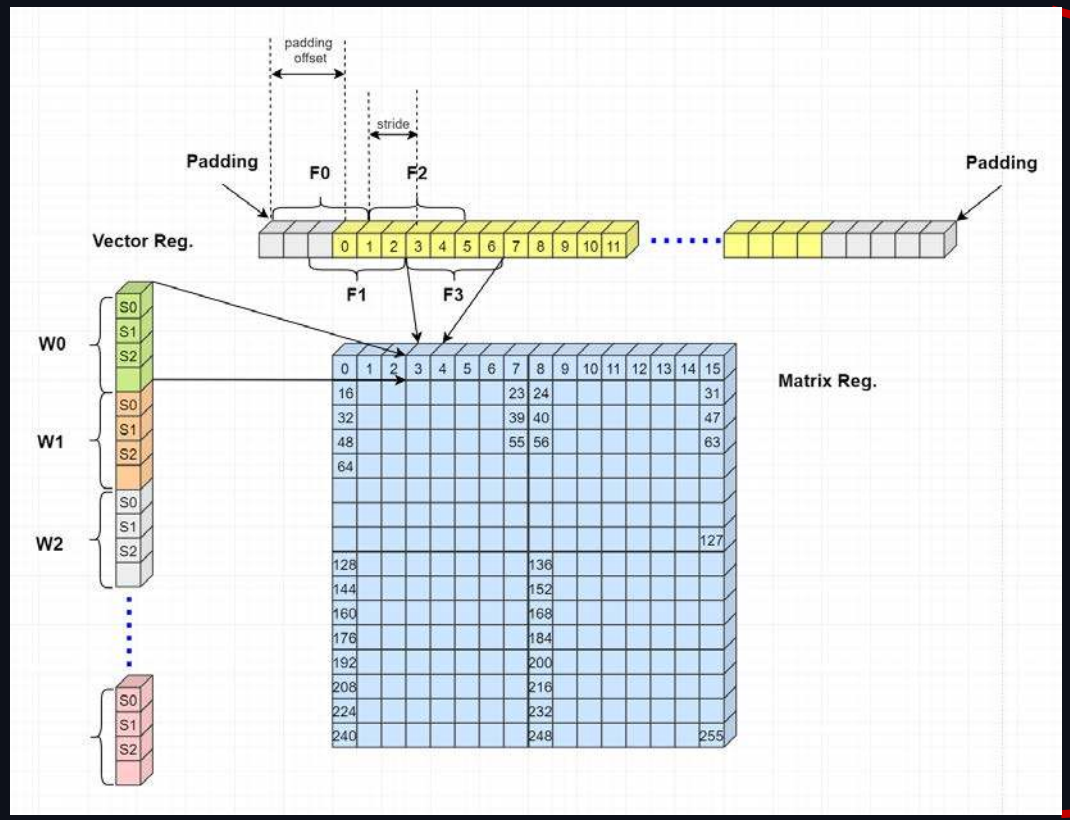


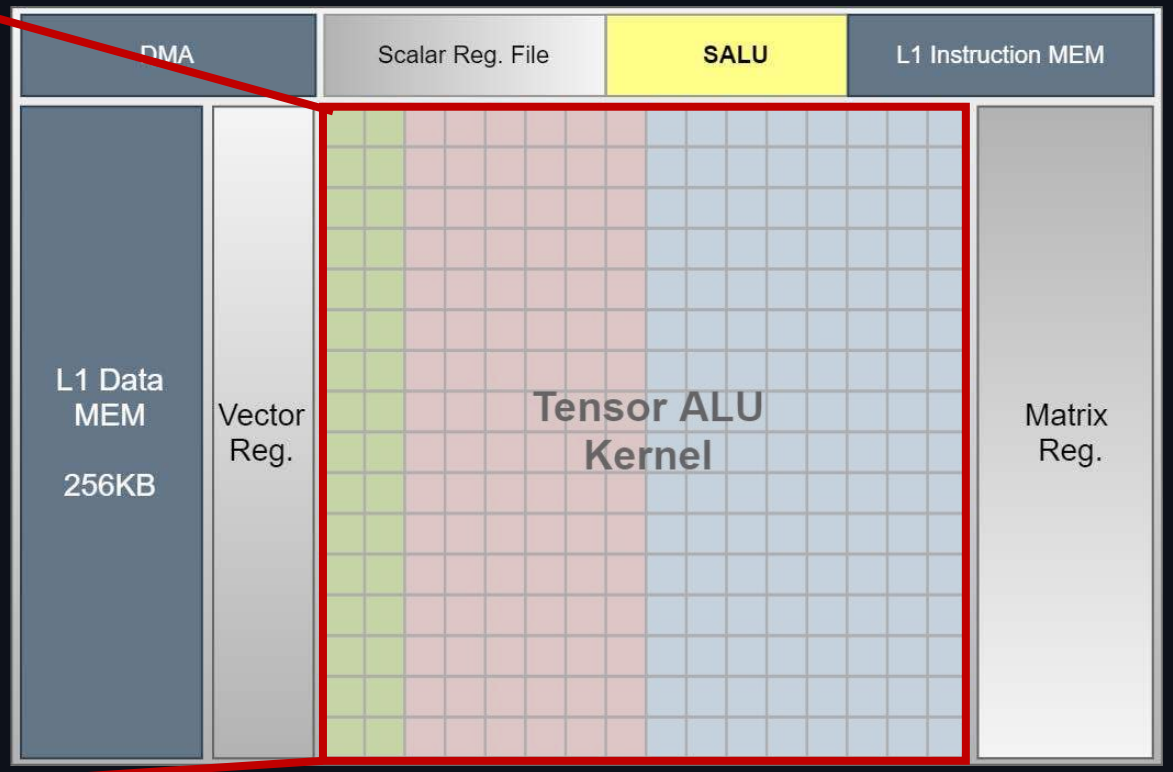
Vector Sum and Vector Pooling



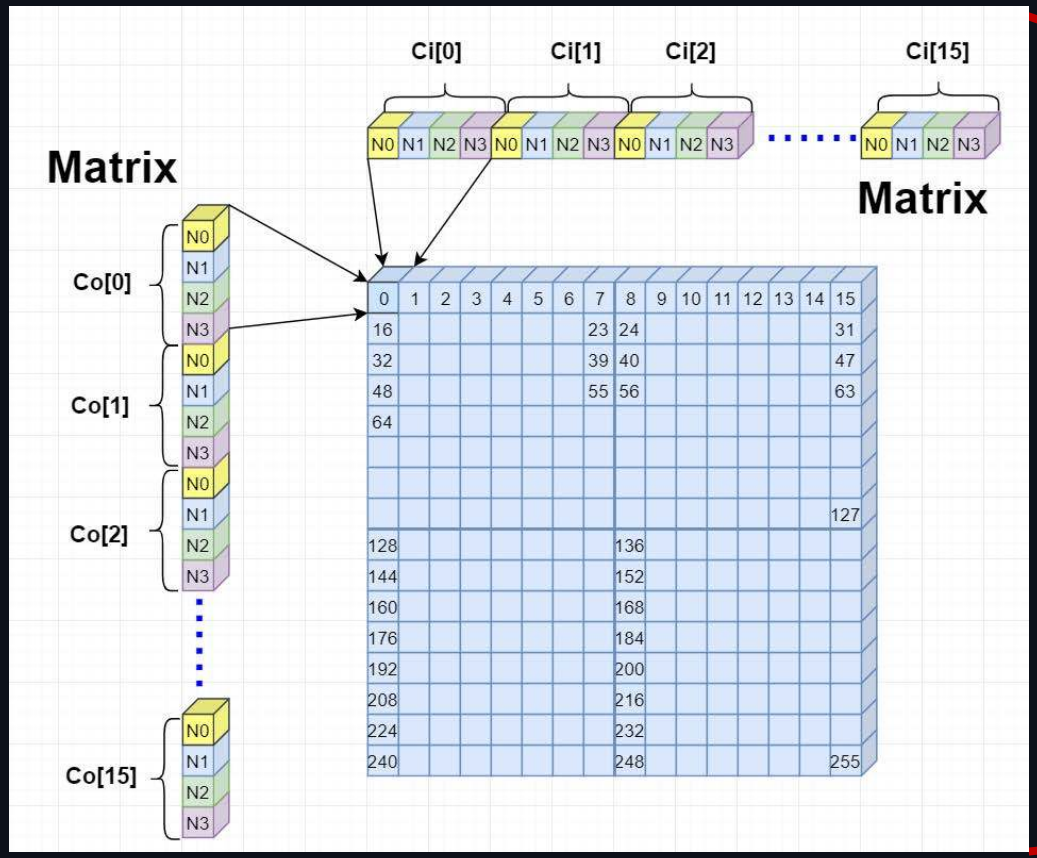


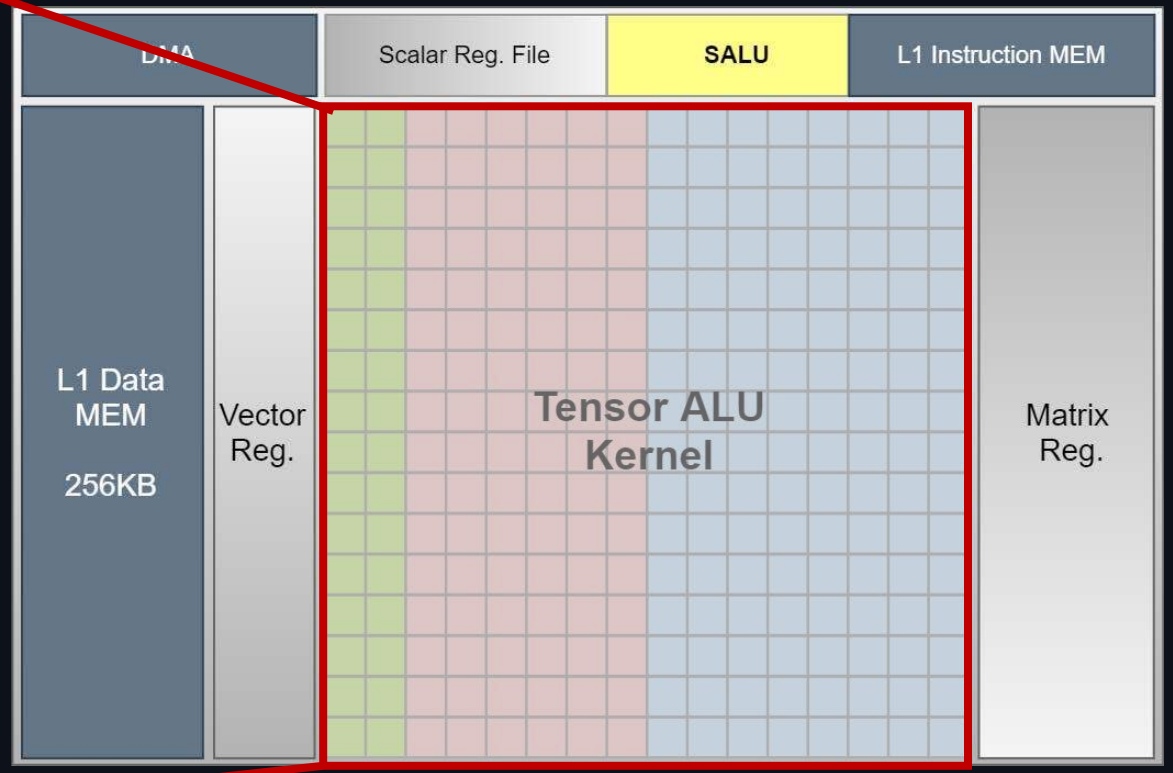
256 Kernels for Conv Operations





256 Kernels for GEMM Operations





GCU-CARE 1.0

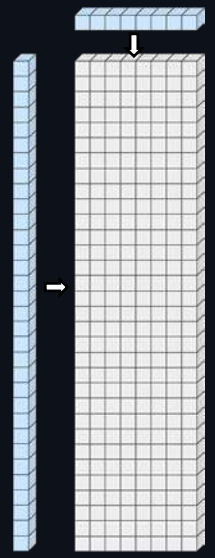
Tensor Shape	
Precision	**Tensor Shape**
FP32 (U)INT32	[32, 1] x [1, 8] [16, 1] x [1, 16] [8, 1] x [1, 32]
FP16 BF16 (U)INT16	[64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16]
(U)INT8	[128, 4] x [4, 2] [64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16] [8, 4] x [4, 32]

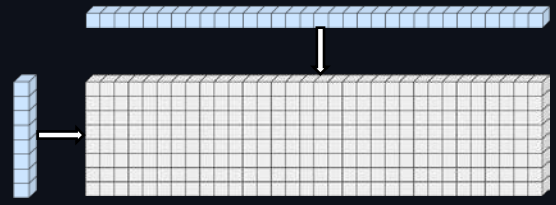
- Mixed precision support
  - FP16/BF16/FP32 input, same format or FP32 accumulation
  - INT8/16/32 input, same format or higher precision accumulation
- Various tensor shapes
  - Flexible for different operators requirements
  - Specifically accelerate convolution operation

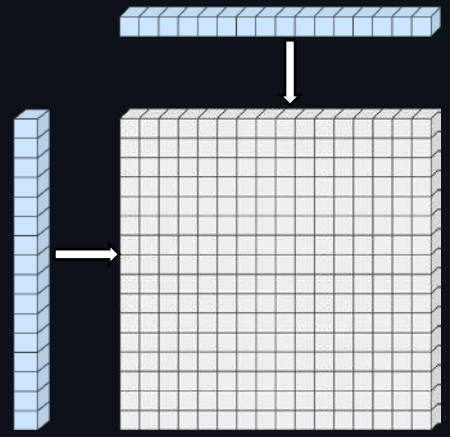
GCU-CARE 1.0

Tensor Shape

Precision	Tensor Shape
FP32 (U)INT32	[8, 1] x [1, 32] [32, 1] x [1, 8] [16, 1] x [1, 16]
FP16 BF16 (U)INT16	[64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16]
(U)INT8	[128, 4] x [4, 2] [64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16] [8, 4] x [4, 32]



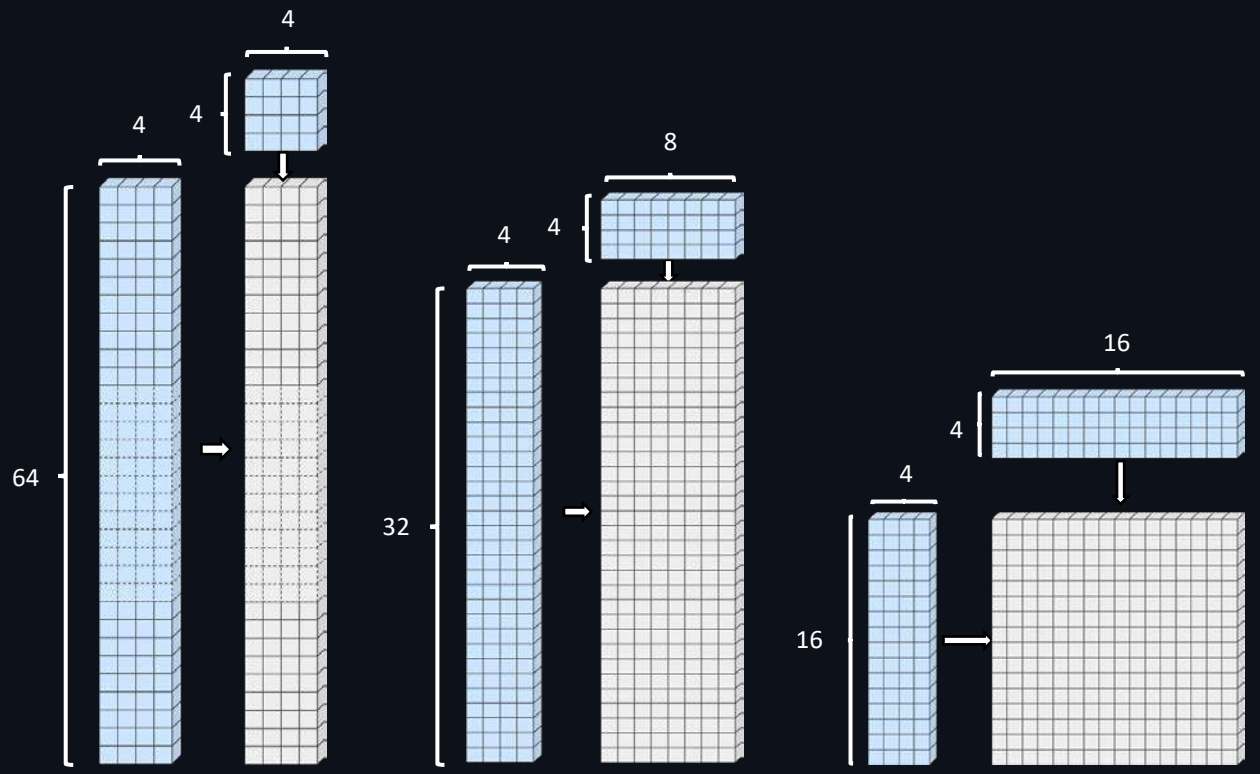




GCU-CARE 1.0

Tensor Shape

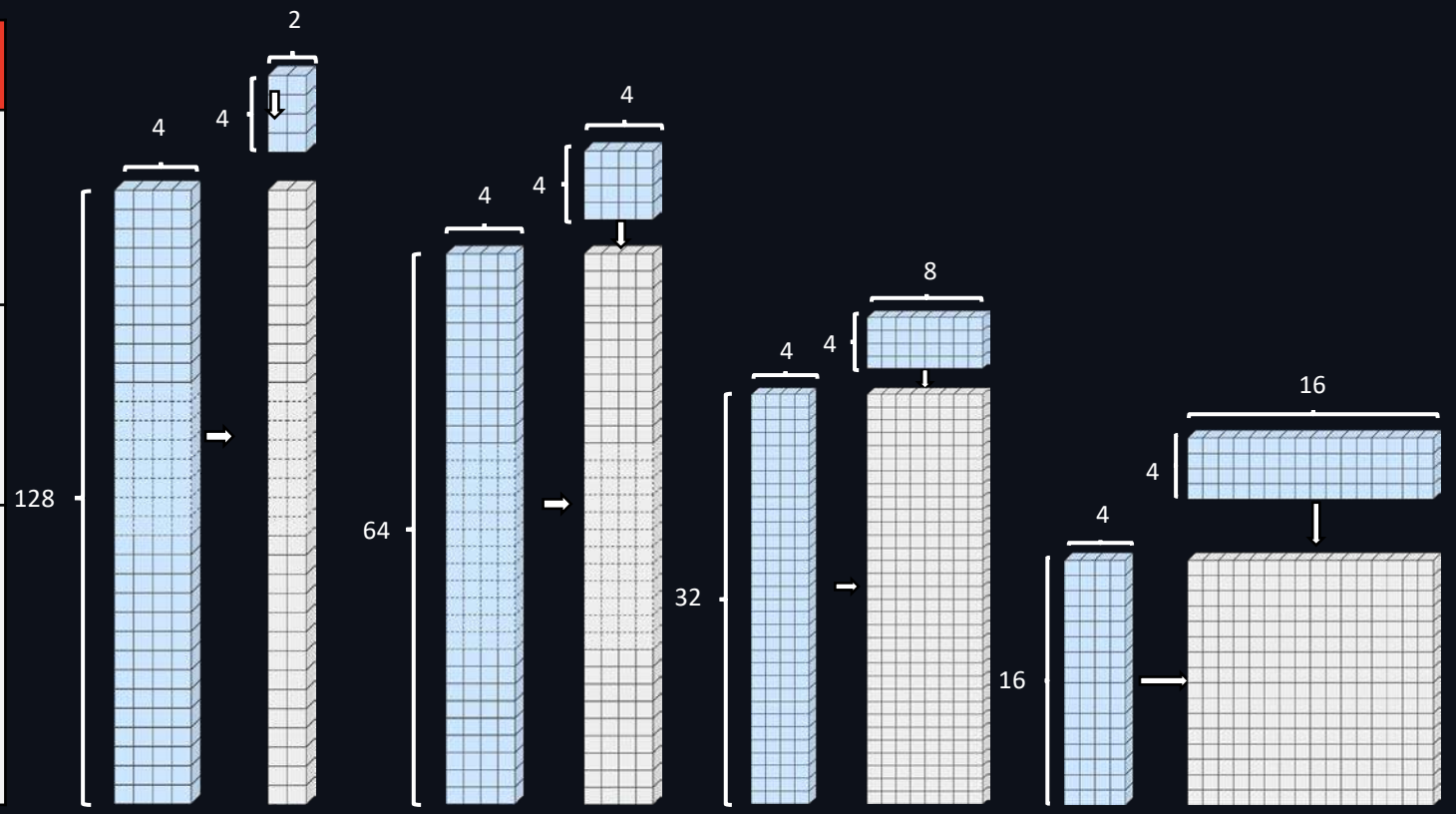
Precision	Tensor Shape
FP32 (U)INT32	[8, 1] x [1, 32] [32, 1] x [1, 8] [16, 1] x [1, 16]
FP16 BF16 (U)INT16	[64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16]
(U)INT8	[128, 4] x [4, 2] [64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16] [8, 4] x [4, 32]



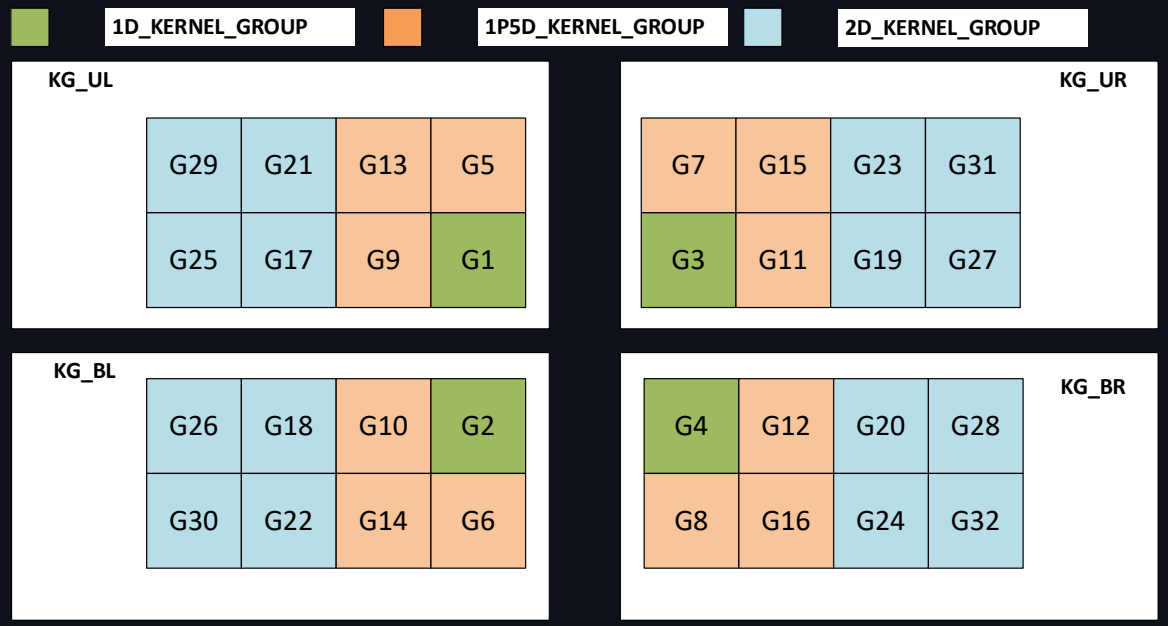
GCU-CARE 1.0

Tensor Shape

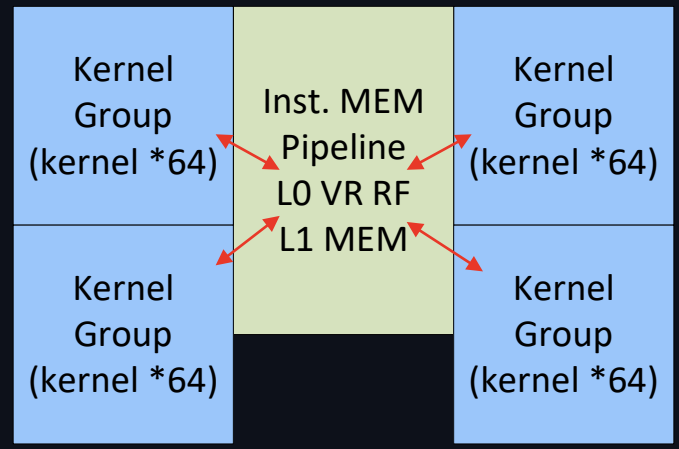
Precision	Tensor Shape
FP32 (U)INT32	[8, 1] x [1, 32] [32, 1] x [1, 8] [16, 1] x [1, 16]
FP16 BF16 (U)INT16	[64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16]
(U)INT8	[128, 4] x [4, 2] [64, 4] x [4, 4] [32, 4] x [4, 8] [16, 4] x [4, 16] [8, 4] x [4, 32]



256 Kernels in 32 Groups



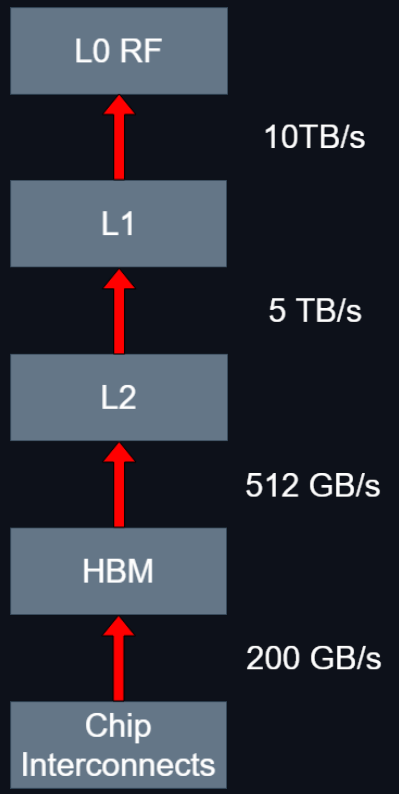
- Matrix registers/accumulators distributed inside each KG
- Each KG fed with 1k-bit data and output 512-bit
- Enable kernels based on active output point number in Inst

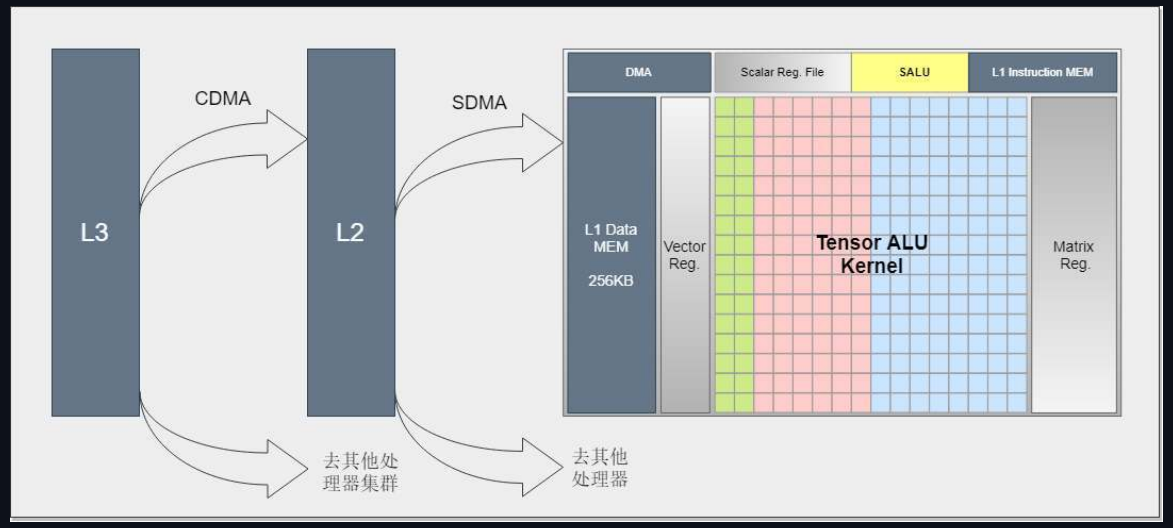


GCU-DARE 1.0

Tensor Data Transfer

- Async dataflow and compute pipeline
- Sync enabled by software programming model

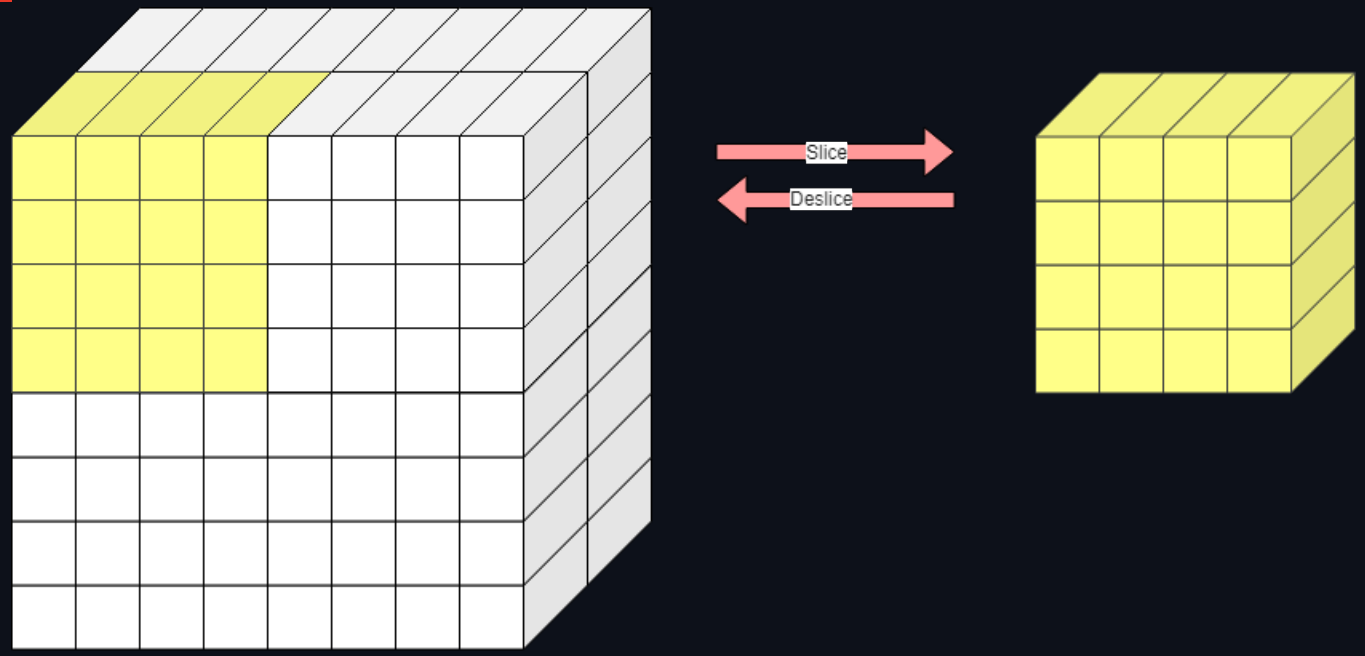




GCU-DARE 1.0

Tensor Operations in Dataflow

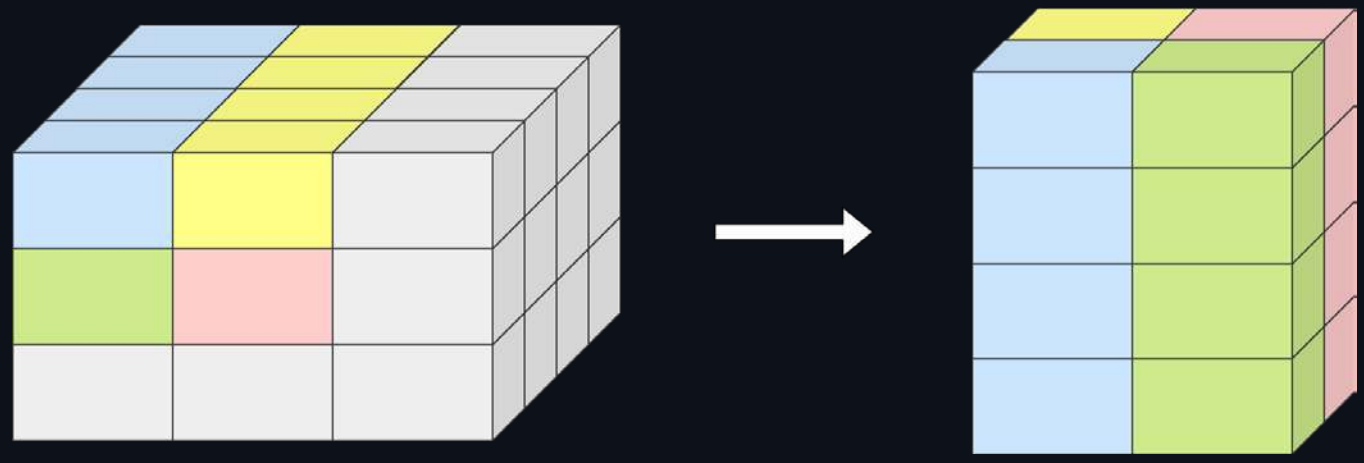
- Slice and Deslice
- Reshape
- Padding
- Concat
- broadcast
- Down-sampling
- Mirror
- Constant Filling
- Etc.



GCU-DARE 1.0

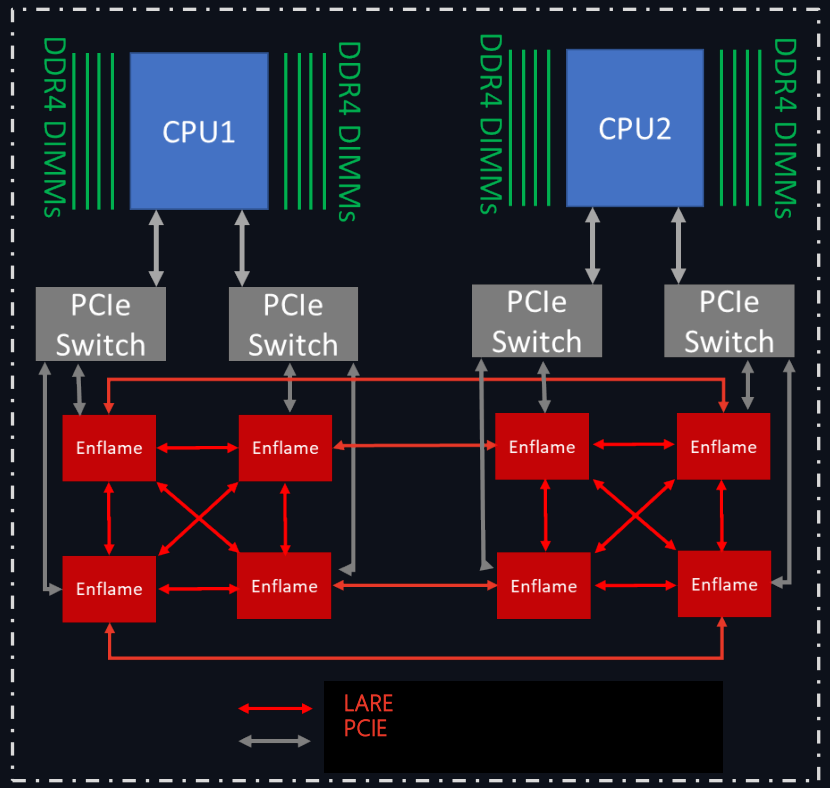
Tensor Operations in Dataflow

- Slice and Deslice
- Reshape (NHWC → HWCN, e.g.)
- Padding
- Concat
- broadcast
- Down-sampling
- Mirror
- Constant Filling
- Etc.

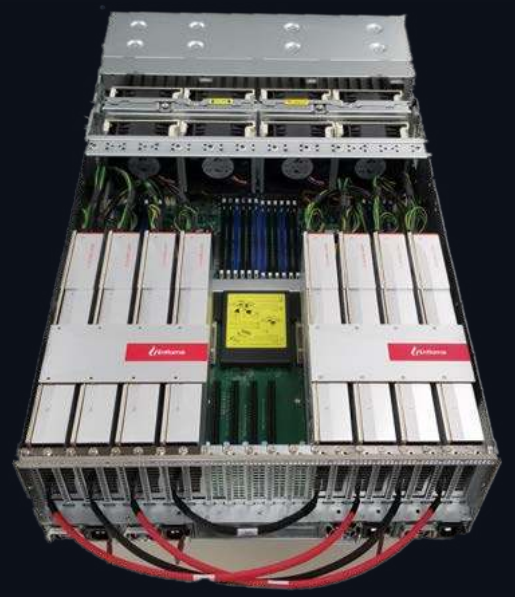


GCU-LARE 1.0

- 200GB/s bi-directional/card
- 800GB/s interconnection speed inside single server
- Latency < 1µs

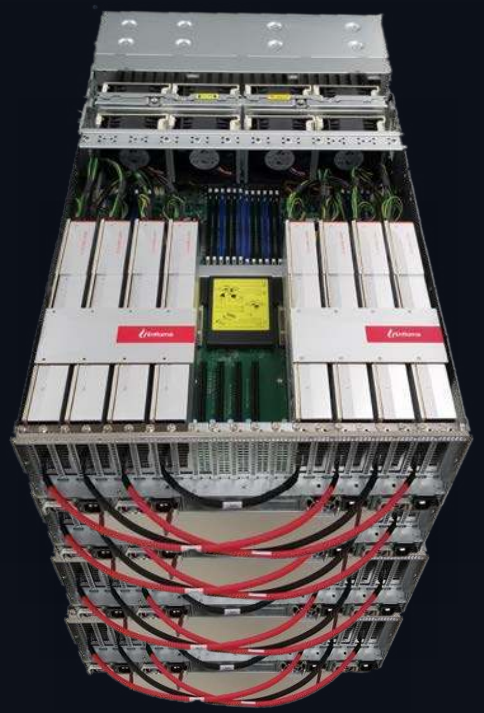


8 card in a server
LARE board + LARE Cable

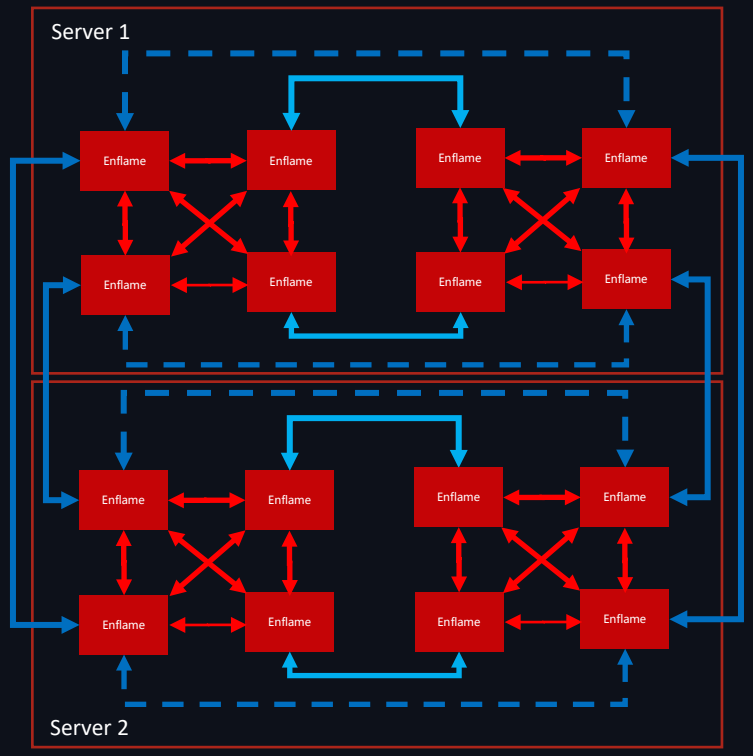


GCU-LARE 1.0

LARE Cable Connection in Rack



Multi-Server Connection in Rack



No RDMA required for inside Rack Connection

Enflame Training Accelerator Card

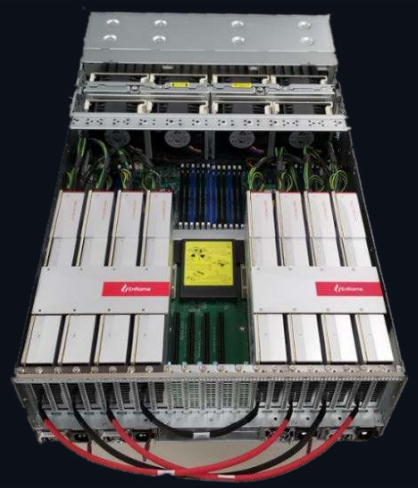
Product Feature Highlight

- Leading FP32 TFLOPS, supporting BF16
- Supports PCIe (T10) and OAM (T11)
- PCIe Gen4.0, board power 225W/300W

CloudBlazer T10

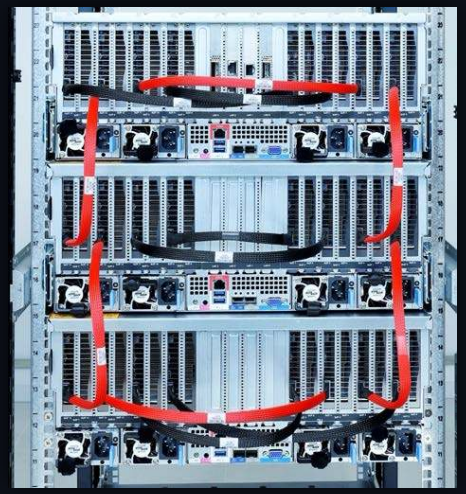
云燧 CloudBlazer T11

Enflame Training Solution



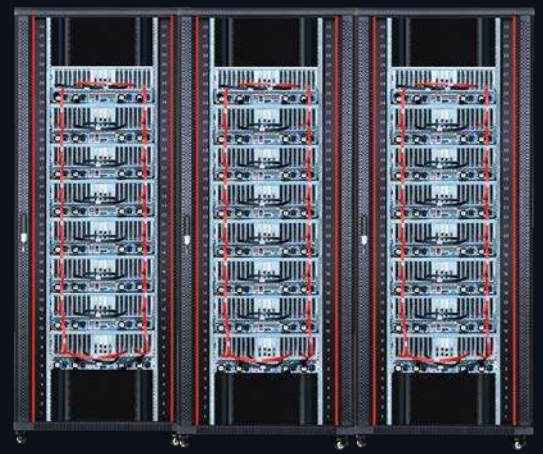
AI Server

- 8/16 cards connected in single node



AI Rack

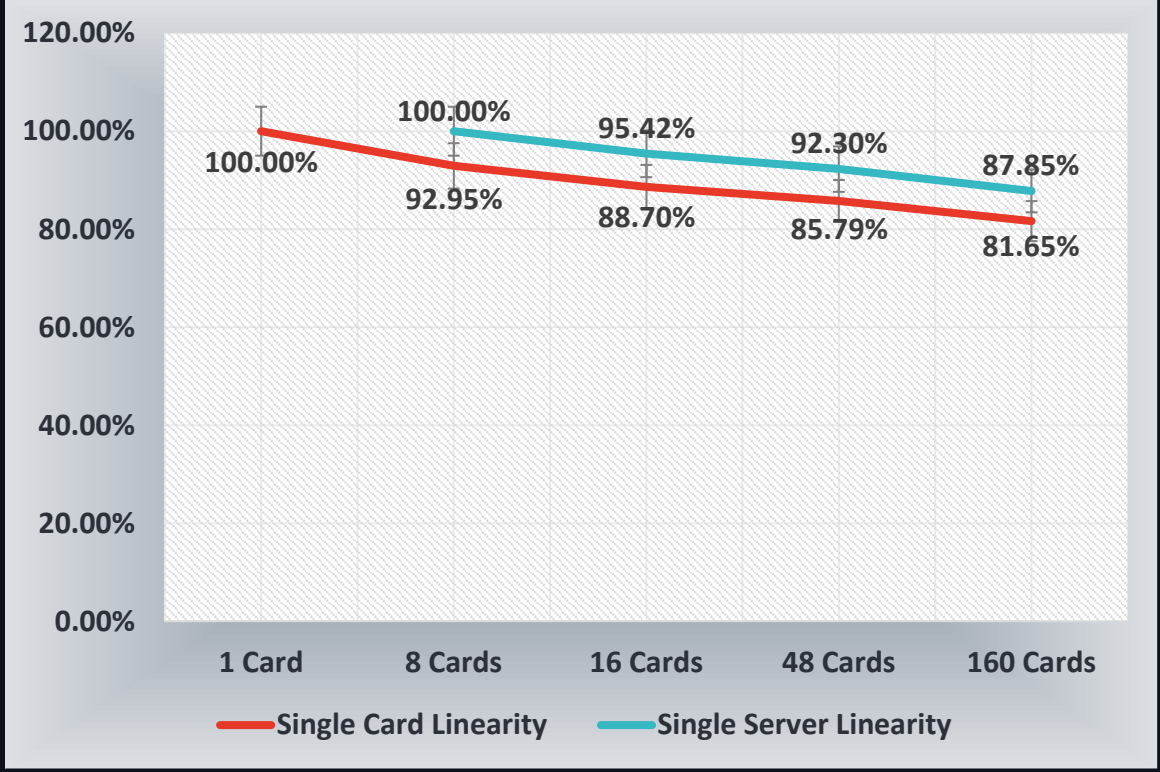
- 64 cards connected in rack

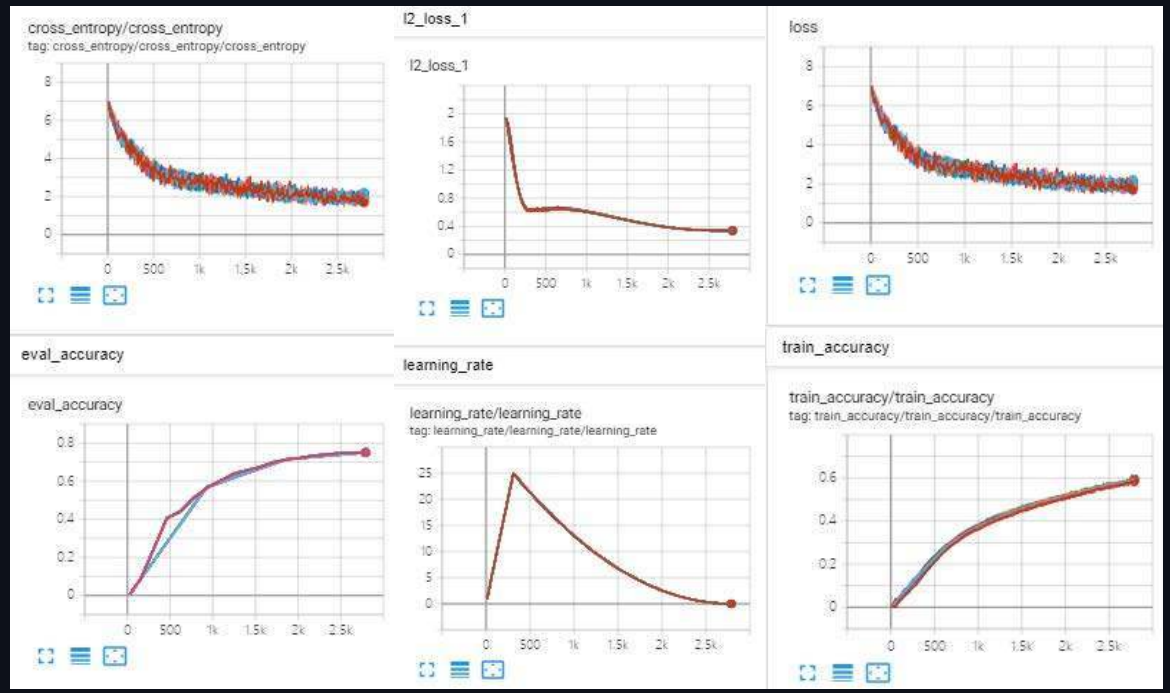


AI Pod

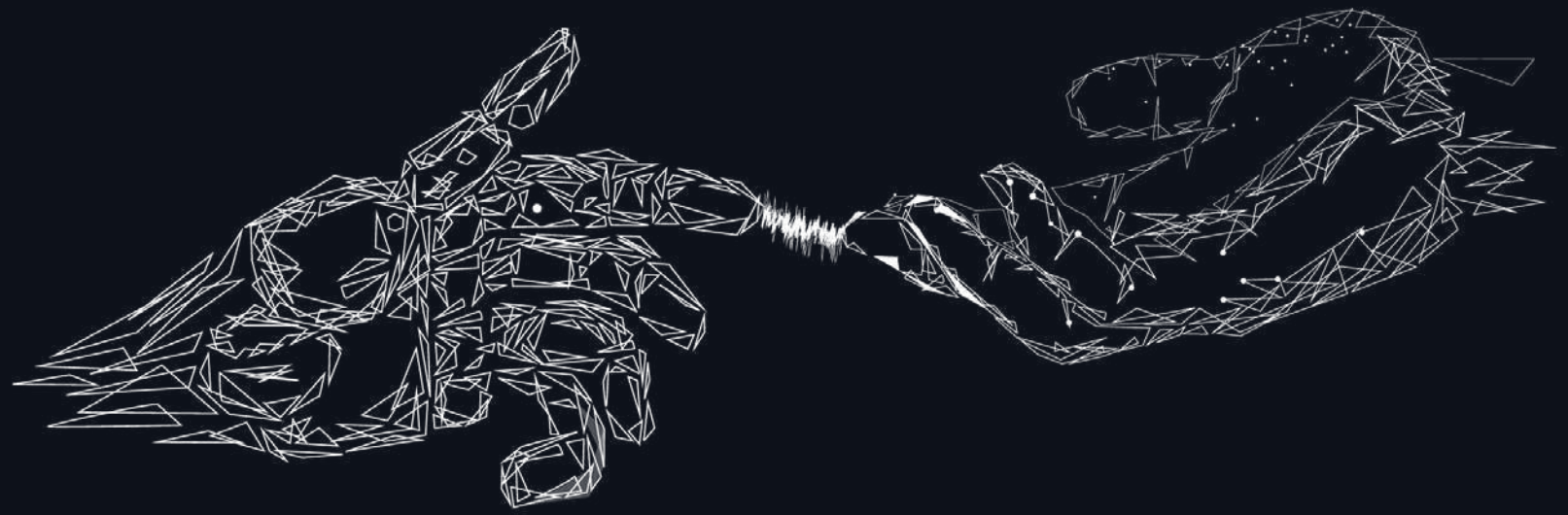
- 2D Torus Topo

Large-scale Distributed Training Cluster





What’s the next?



Thank You