Context Packer

Pinned Memory Table (PMT)

Stream ID	Tenant ID	Pinned Memory Address	Memory Size	Memory operation Phase
3215	A	0x90638	2048	H2D
…	…	…	…	…

GPU Init Request

Scheduler Init Request

GPGPU Request

Stream Creator

Manual Stream Convertor

GPU Request Stream Convertor

Sync to Asyn Memop Convertor

GPU Stream Request

GPGPU Request

GPU Stream Request

Sync Memop Request

Async Memop Request