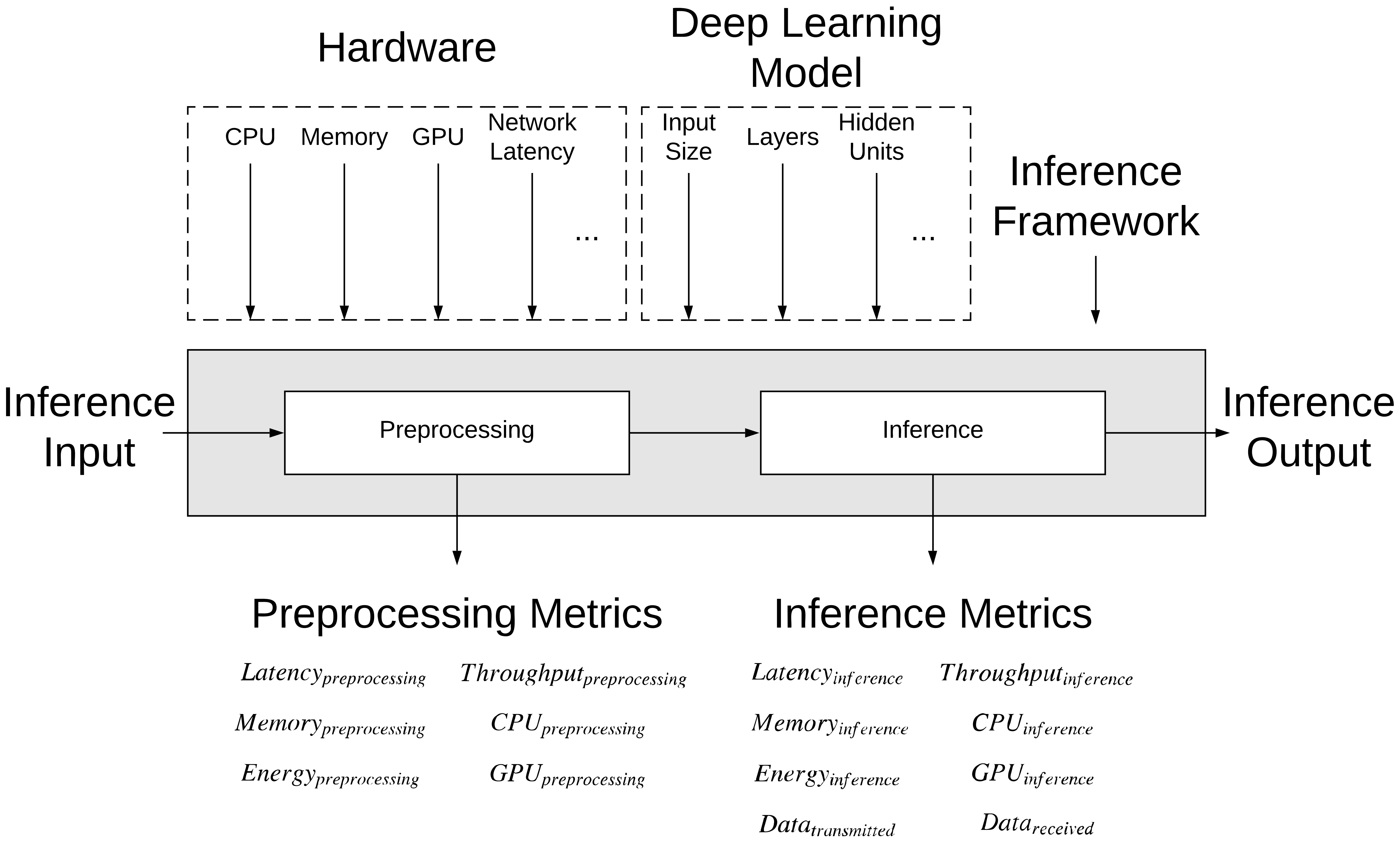
Hardware
Deep Learning Model
CPU
Memory
GPU
Network Latency
...
Input Size
Layers
Hidden Units
...
Inference Framework
Inference Input
Preprocessing
Inference
Inference Output
Preprocessing Metrics
$Latency_{preprocessing}$
$Throughput_{preprocessing}$
$Memory_{preprocessing}$
$CPU_{preprocessing}$
$Energy_{preprocessing}$
$GPU_{preprocessing}$
Inference Metrics
$Latency_{inference}$
$Throughput_{inference}$
$Memory_{inference}$
$CPU_{inference}$
$Energy_{inference}$
$GPU_{inference}$
$Data_{transmitted}$
$Data_{received}$