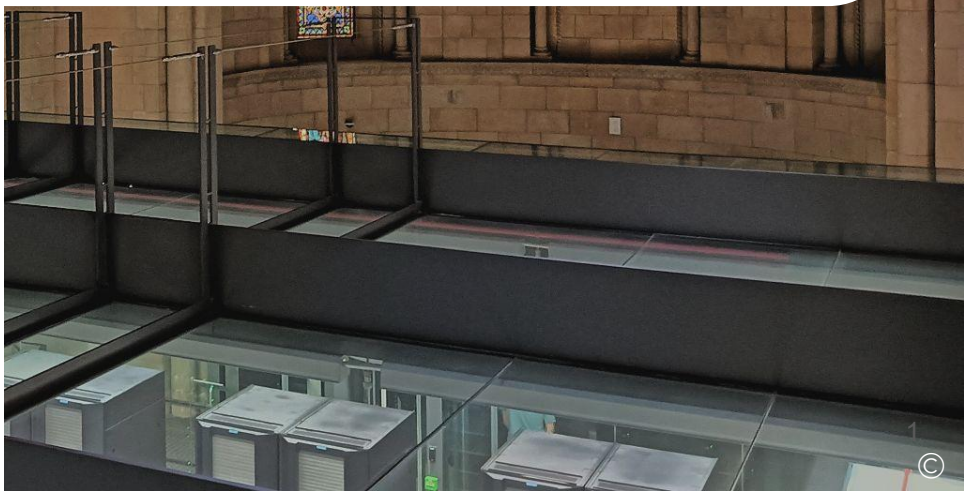
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

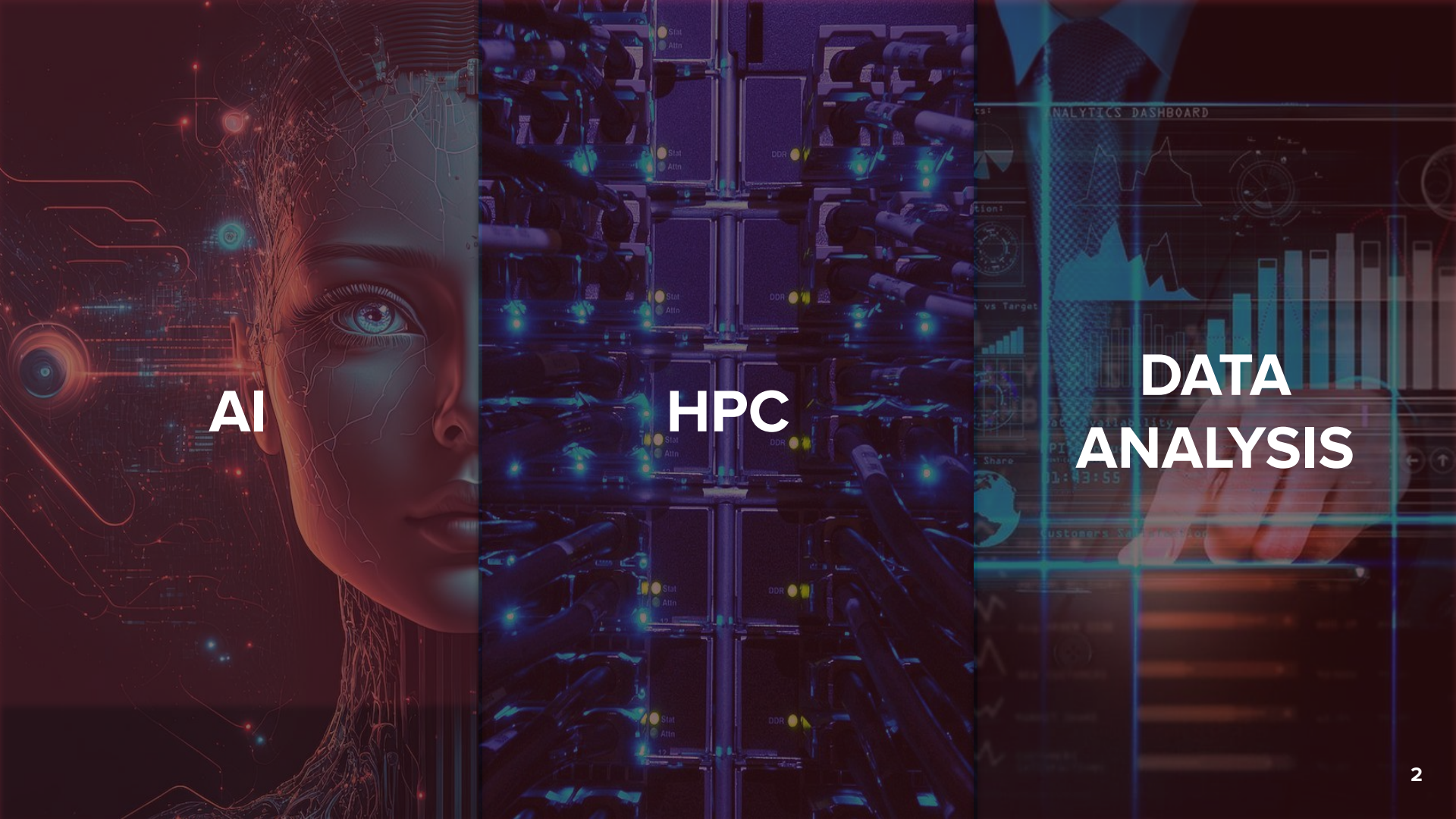
FIB

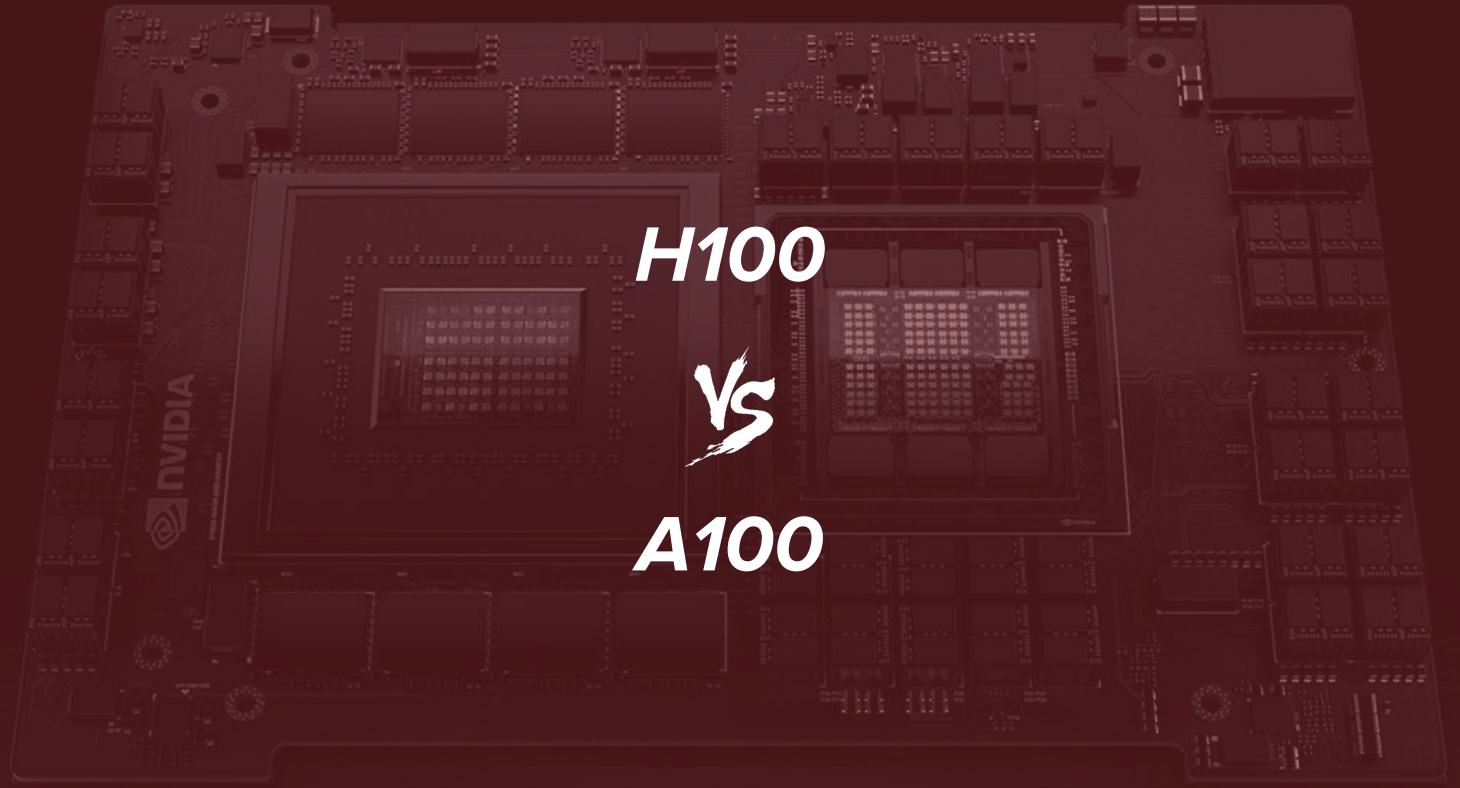
NVIDIA H100
Tensor Core GPU Architecture

9th October 2023



Huilin Ni
huilin.ni@estudiantat.upc.edu
Sergi Guimerà Roig
sergi.guimera@estudiantat.upc.edu

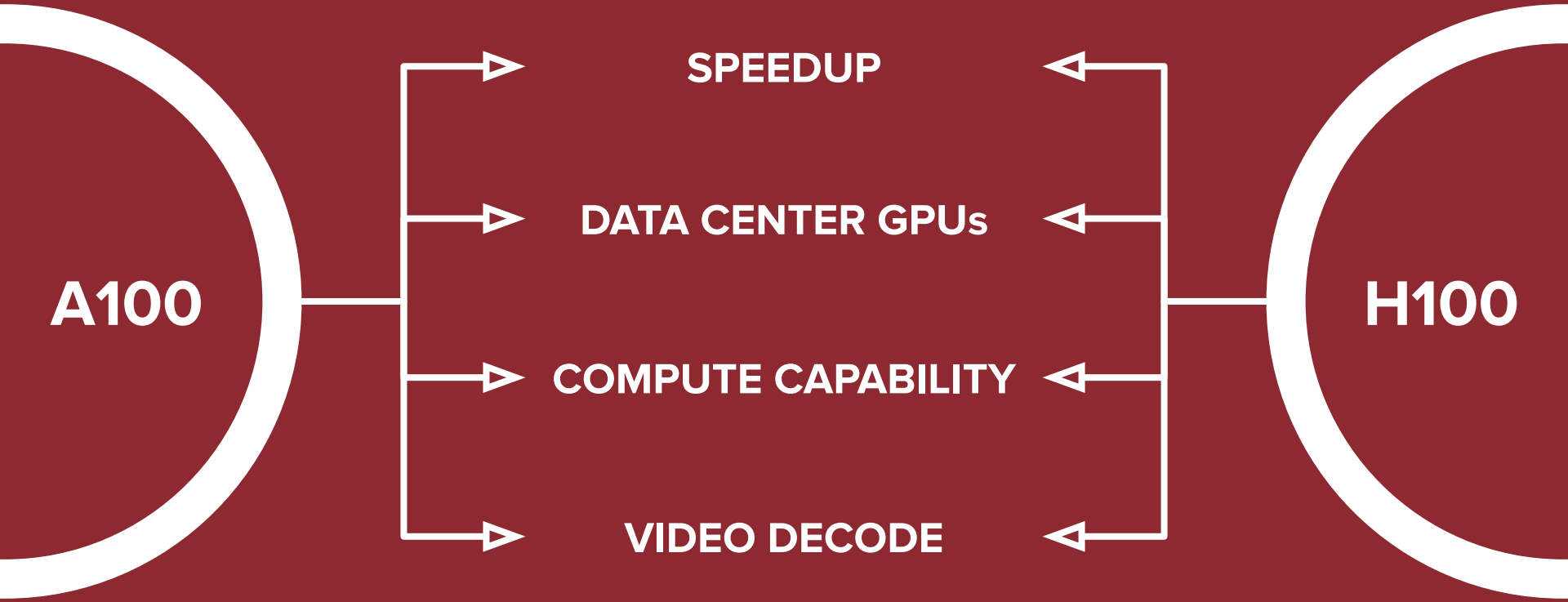

AI
HPC
DATA
ANALYSIS



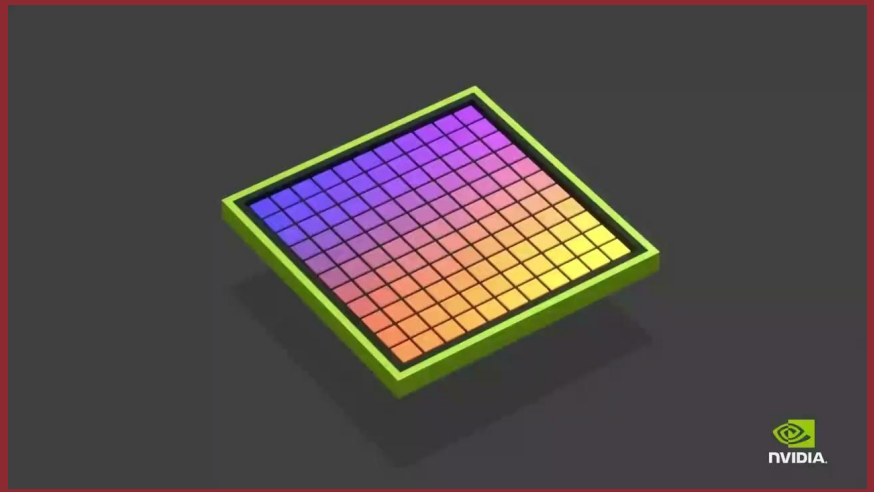
H100

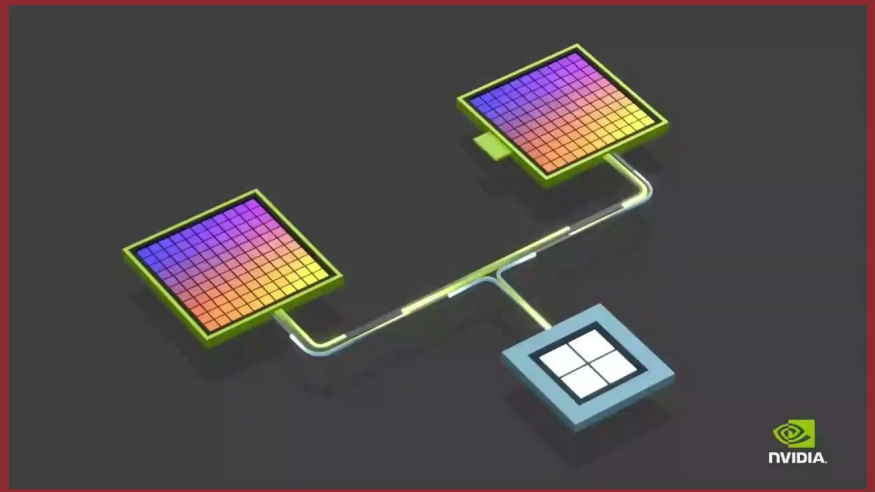
VS

A100

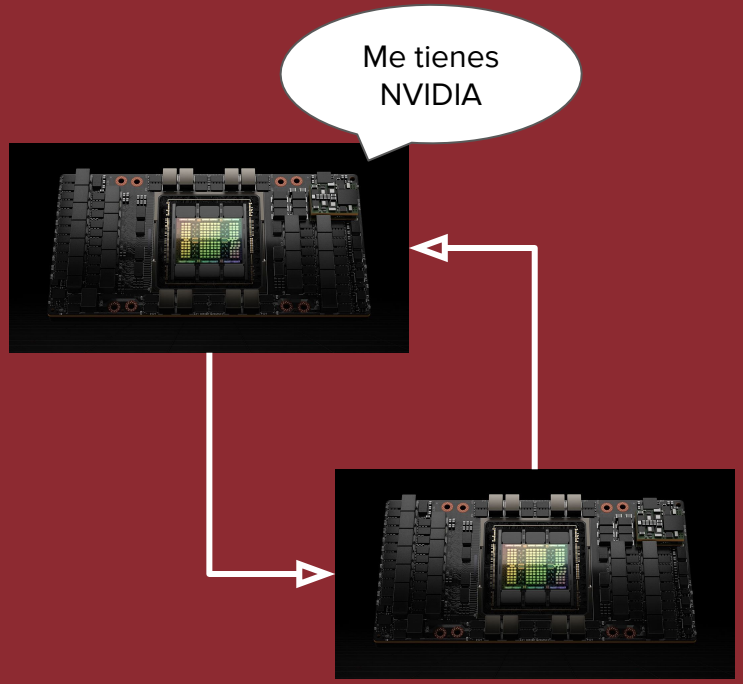

A100
SPEEDUP
DATA CENTER GPUs
COMPUTE CAPABILITY
VIDEO DECODE
H100

PCIe

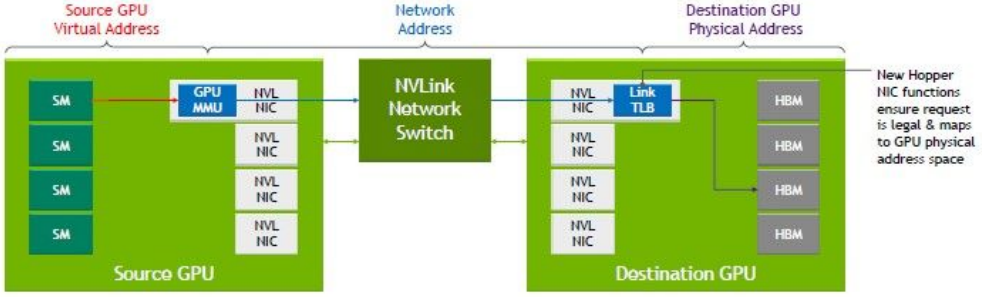




(4th Gen) NVLink



NVLINK NETWORK



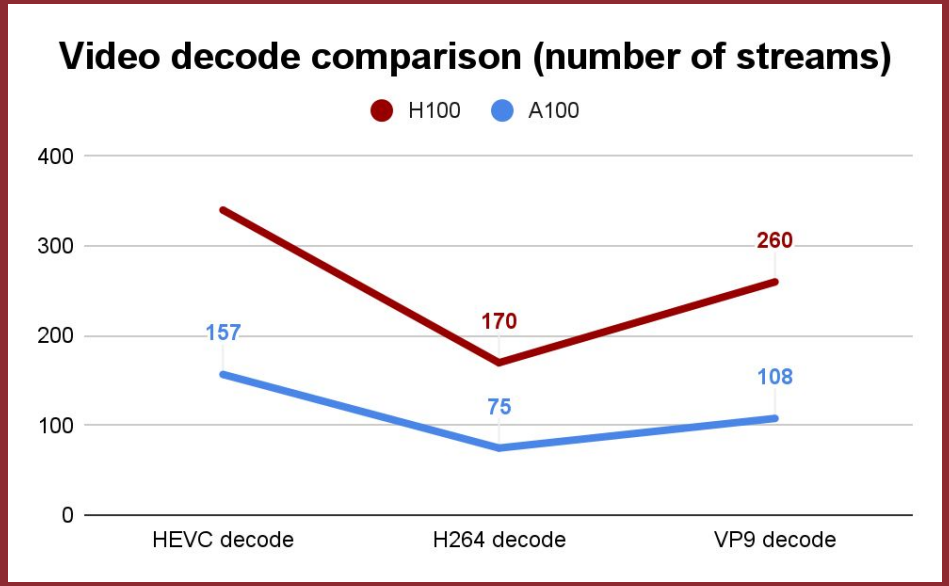
	NVLink	NVLink Network
Address Spaces	1 (shared)	N (independent)
Request Addressing	GPU physical address	Network address
Connection Setup	During boot process	Runtime API call by software
Isolation	No	Yes

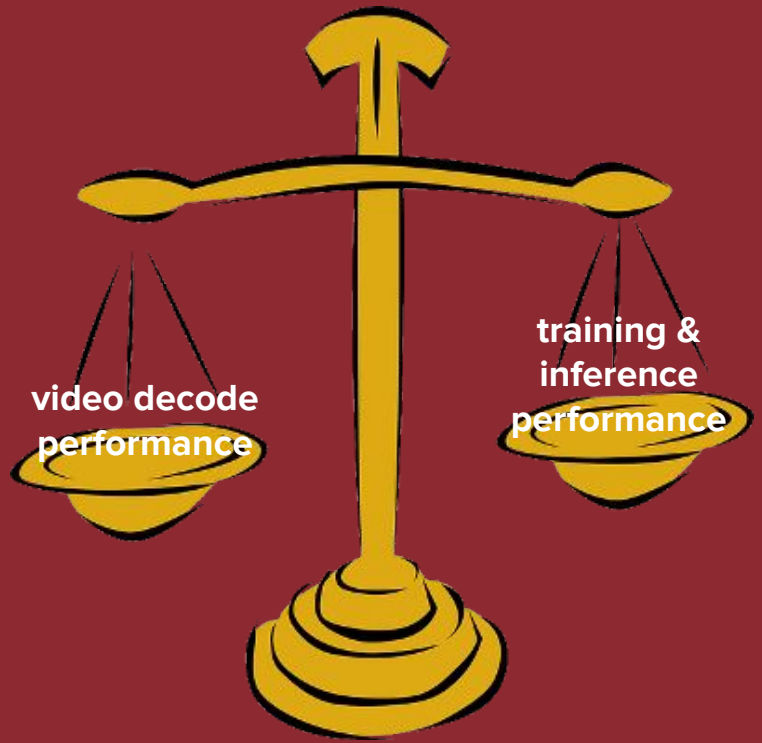
Security Enhancements

- Data Protection & Isolation
- Content Protection
- Physical Damage Protection

Confidential Computing

H100 Video IO Features





DALI

LMDB (Caffe, Caffe2)

RecordIO (MXNet)

TFRecord (TensorFlow)

List of JPEGs Separate files

Video files