Université de Paris
Paris Centre for
Cosmological Physics

Comparison of Graphcore IPUs and Nvidia GPUs for cosmology applications

Bastien Arcelin (APC, Paris)

IN2P3/IRFU
Machine Learning Workshop

Future photometric surveys

Rubin Observatory LSST

- First light in 2021
- 10 years of operation

↓

60 Petabytes

Euclid

- Send to space in 2022
- 6 years of operation

↓

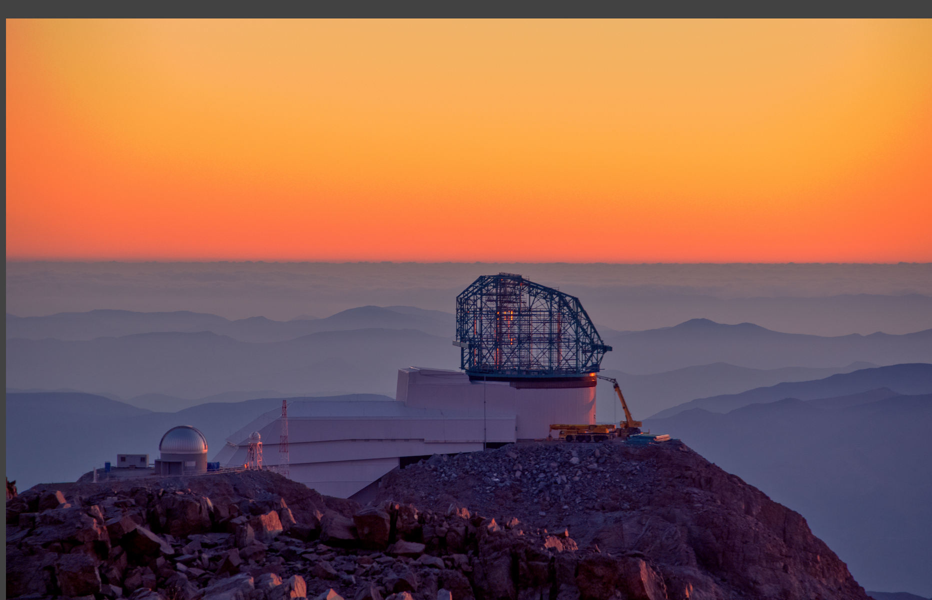
156 Petabytes

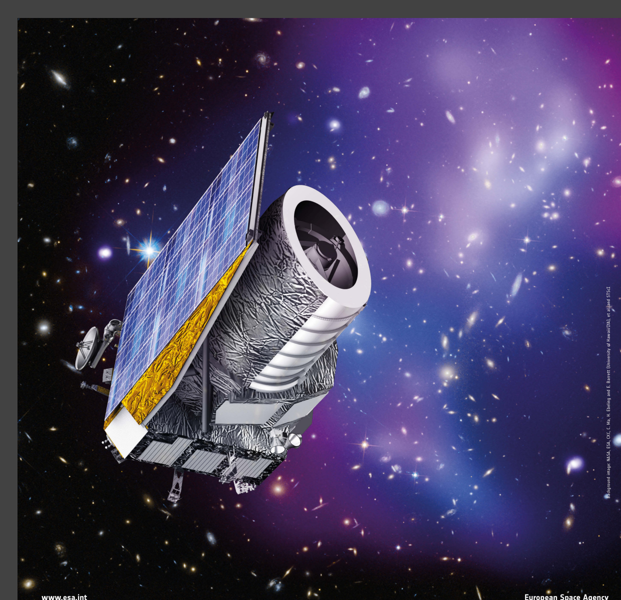
Nancy-Grace Roman Space Telescope

- Send to space mid-2020s
- 5 years of operation (+5)

↓

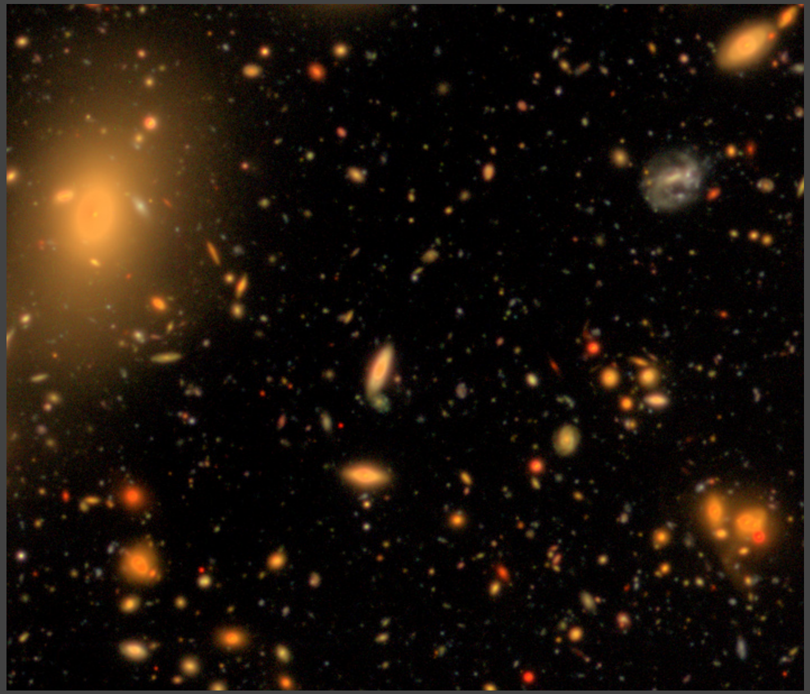
20 (40) Petabytes





Future photometric surveys

- Nature of the data:
  sky images in several filters (or colors)



Credit: HSC

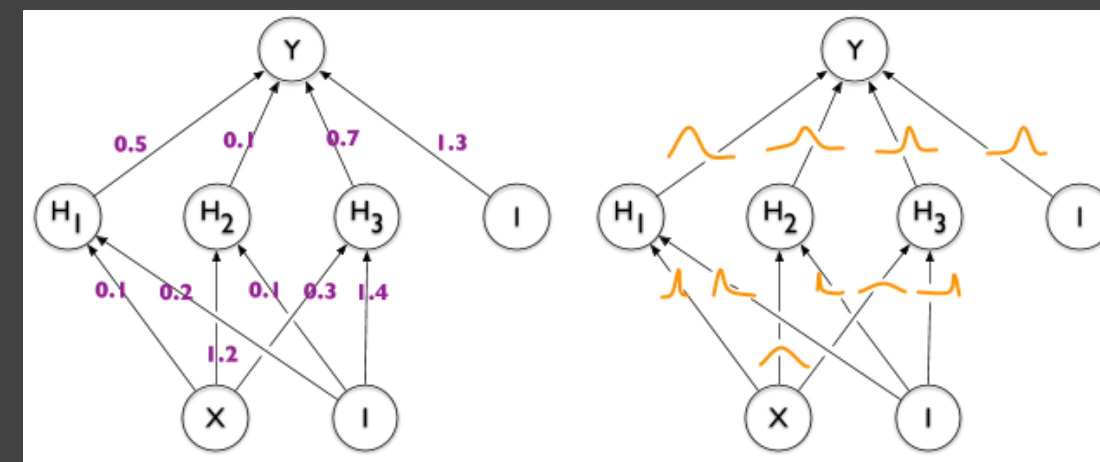
- Example for LSST:
  look for transient (asteroids, Supernovae…):
  10 millions alerts per night ⟶ Deep learning for classification of events
  (Möller+2020)

Credit: Rick Fienberg

Cosmological applications

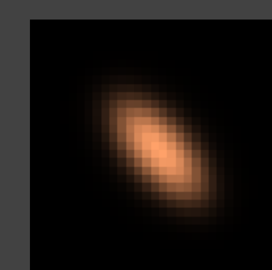
Training data and use cases

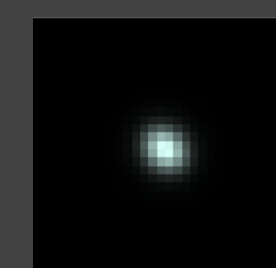
- Cosmological use cases:
  - Inference: galaxy image simulation
  - Training: galaxy shape parameter estimation:
    - Deterministic neural network
    - Bayesian neural network: characterize epistemic uncertainty, i.e. from the model, learning approximate posterior distribution over the weights and biases (via reparametrization trick, Kingma+2015)

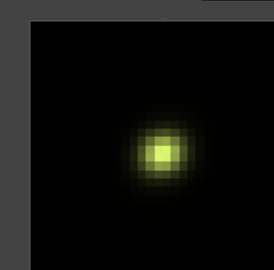


Credit: Sanjay Thakur

- Training and generated data:
  - 64 x 64 pixels stamps of galaxy images (without noise) in 6 different filters (or colors)
  - galaxies generated using fitted parametric models on real data (GREAT3 Challenge, Mandelbaum et al.+2014)
  - Networks fed with or generating arrays of size (batch size, 64, 64, 6)



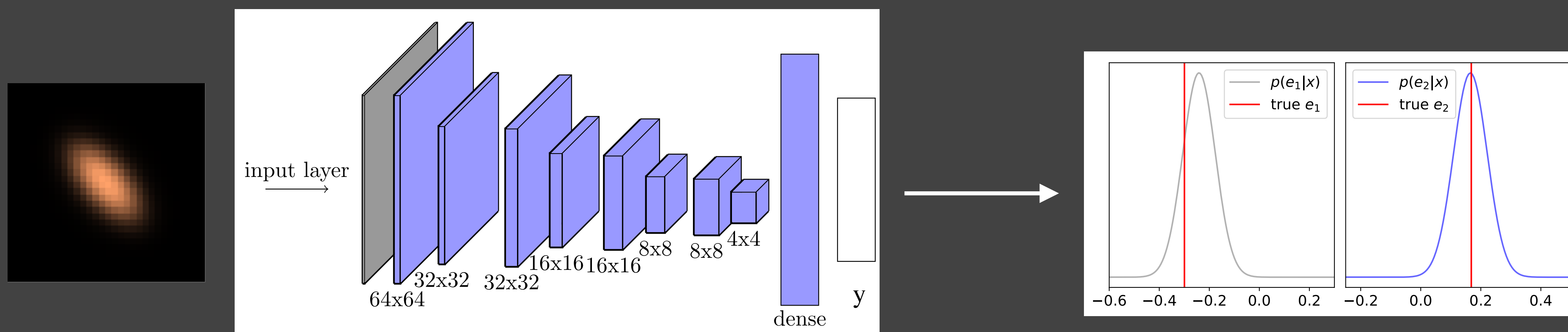




Cosmological applications

Training data and use cases

- Shape parameter estimation:
  - deterministic convolutional neural network (CNN): 1.5M parameters
  - Bayesian CNN: 2.7M parameters



- Galaxy image generation (inspired from Lanusse+2020):
  - Variational AutoEncoder (VAE, Kingma+2014) trained on galaxy images (already trained on GPUs)
  - Generation via sampling the latent space using normalizing flows (MADE, Germain+2015)

Variational AutoEncoder

Normalizing flow + trained decoder

IPU/GPU

Hardware description

	Processing Unit	Cores	Memory	Single precision performance	Max Power Consumption
GPU IPU	Nvidia Tesla V100 PCIe Graphcore Colossus MK1 GC2	5120 1216	32000 Mb 286 Mb	14 TFLOPS 31.1 TFLOPS	250 W 120 W

MK1 IPU

- MIMD (Multiple Instruction Multiple Data)
- 1216 cores (tile = core + 256KiB of memory), 6 threads per tile ⟶ 7.296 threads in parallel
- Poplar SDK v.1.4.0
- accessed through Azure IPU preview

V100 GPU

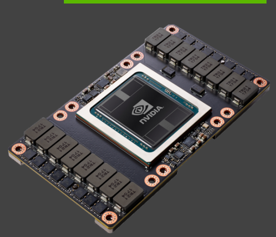
- SIMD (Single Instruction Multiple Data)
- 5120 cores
- CUDA v.10.1.105
- accessed through CC IN2P3

GRAPHCORE

TensorFlow

Version 2.1.0

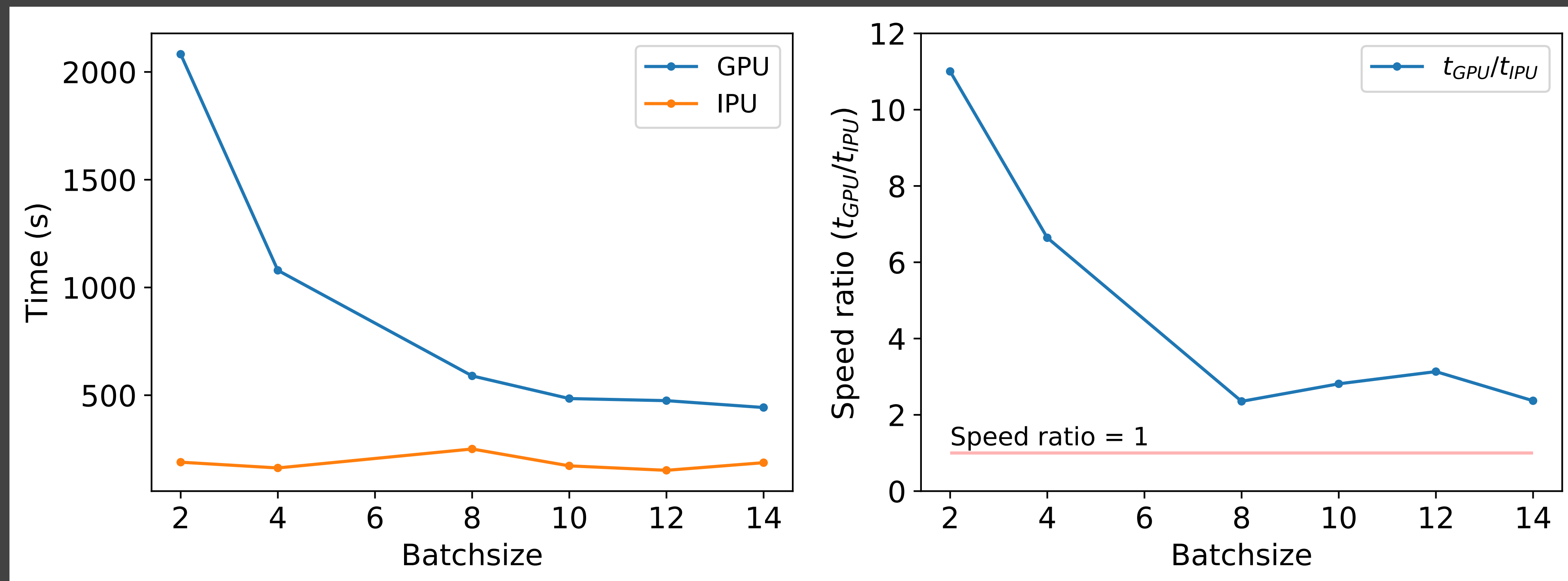
NVIDIA



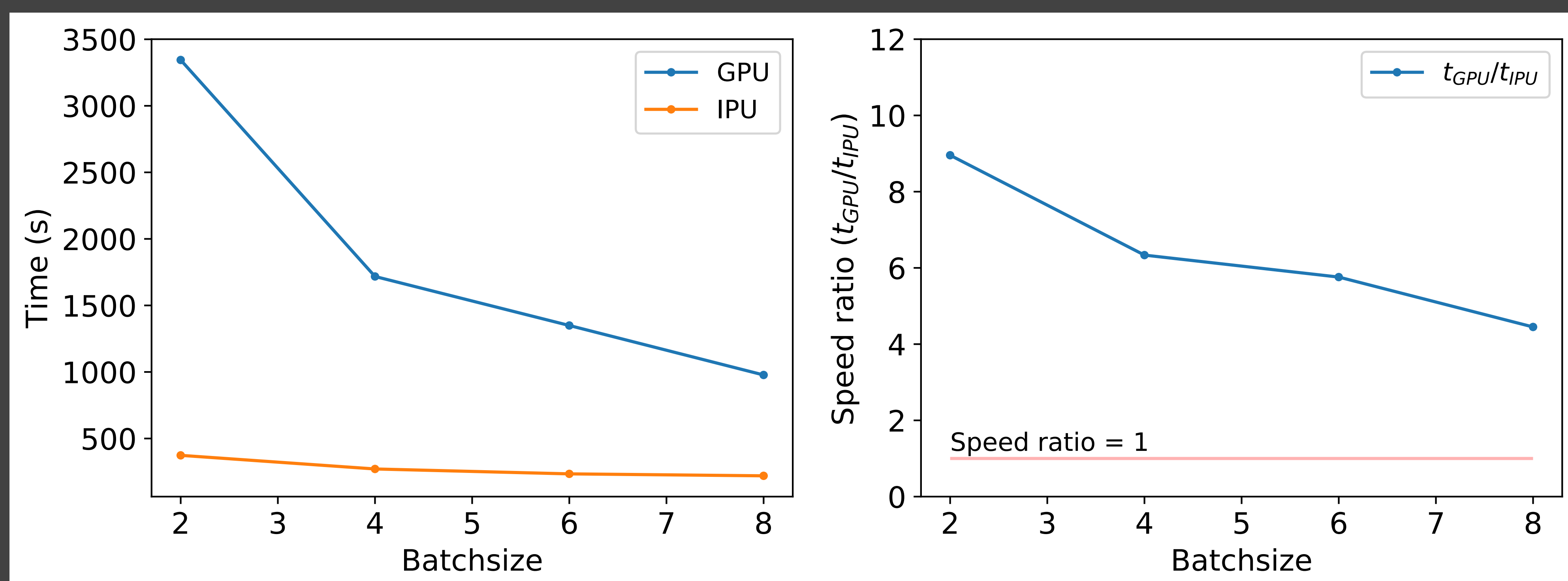
Results

Training

Deterministic
CNN
1.5M parameters



Bayesian
CNN
2.7M parameters

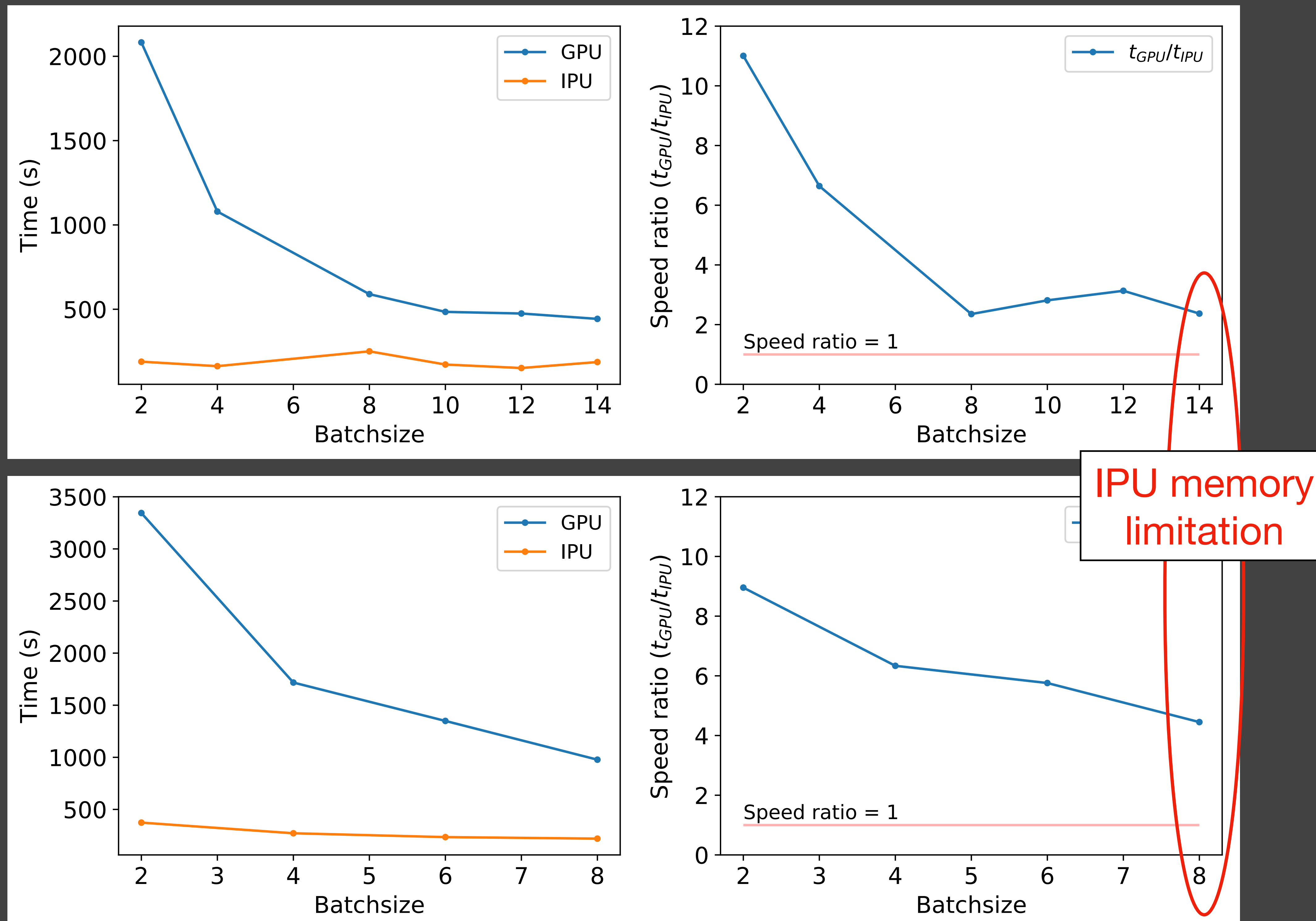


Results

Training

Deterministic
CNN
1.5M parameters

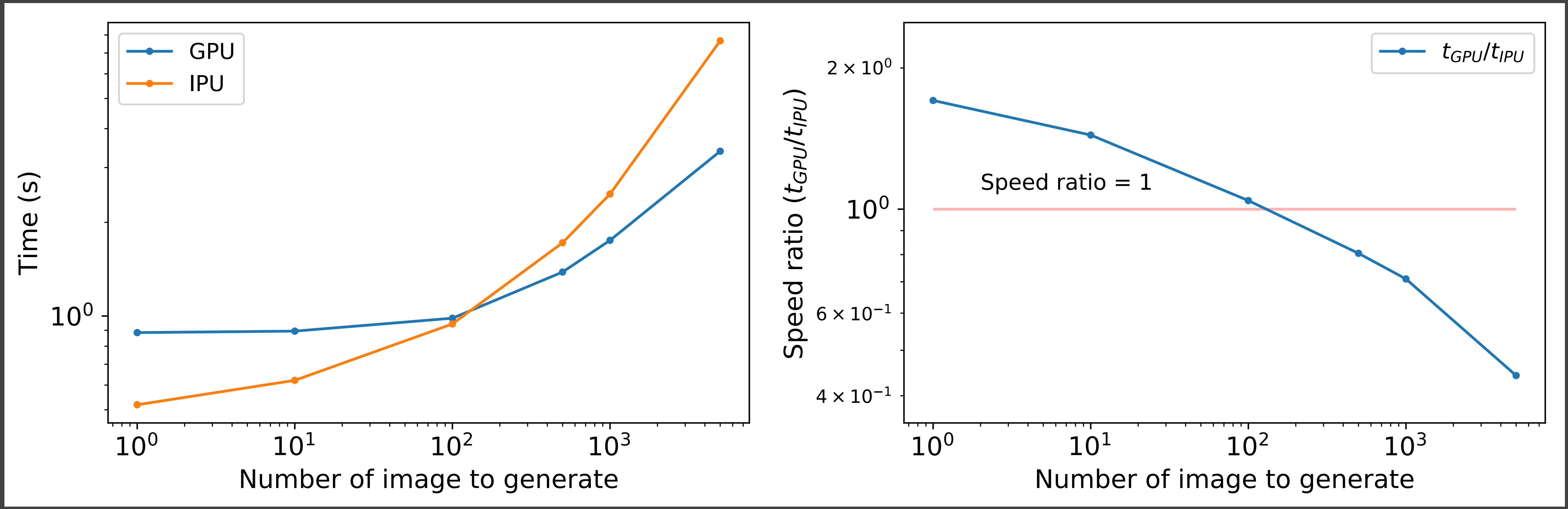
Bayesian
CNN
2.7M parameters



Results

Inference

Sampling latent space + generation image with VAE decoder



➡ Train a network generating samples on the fly: IPU

➡ Generate large amount of data: GPU

<u>Example</u> : CosmoDC2 (Korytov+2019) catalog contains around 2.26 billion galaxies: < 18 days of computing for a single V100 GPU.

Conclusion

Training NN:
IPUs performed at least twice as fast as GPUs
- But : Restriction to small batch sizes (IPU’s small memory size)

Inference:
IPUs perform better than GPUs at small batches but are outperformed for larger sample sizes.
- Hardware choice depends on the task

Important note:
In an analysis or simulation pipeline: scale the processes on several IPUs/GPUs (no more memory limitation ?)

Next:
- Same test on new generation hardware: MK2 IPUs (increased in-processor memory and exchange memory, and higher single precision performance) and A100 GPUs (more memory, and higher single precision performance).
- This work on arxiv soon.