IT UNIVERSITY OF COPENHAGEN

DASYA
www.dasya.dk
@dasyaITU

DAPHNE
daphne-eu.github.io/

RAD
rad.itu.dk

INDEPENDENT
RESEARCH FUND
DENMARK
dff.dk

www.itu.dk

TPCxAI on NVIDIA Jetsons

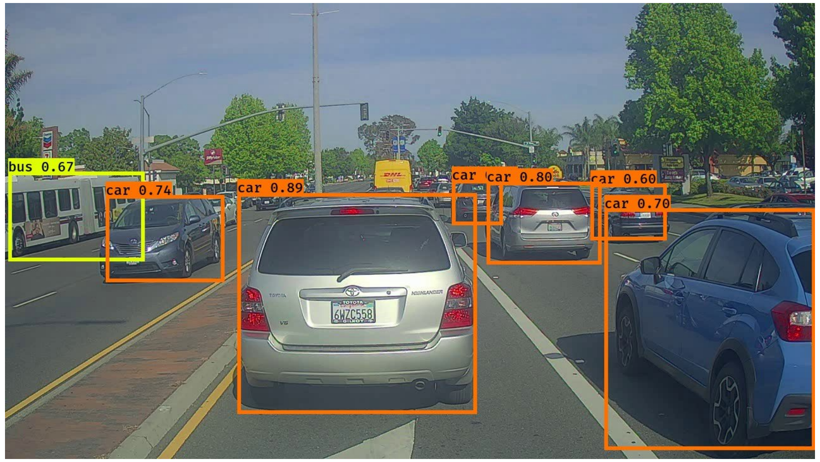
Robert Bayer, Jon Voigt Tøttrup, Pınar Tözün

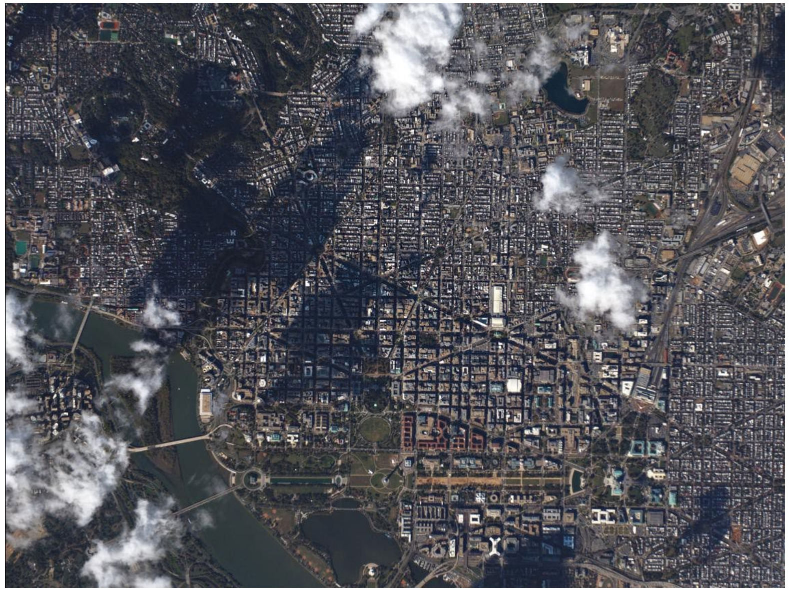
IT University of Copenhagen

roba@itu.dk, @robertbayer_

ML@Edge

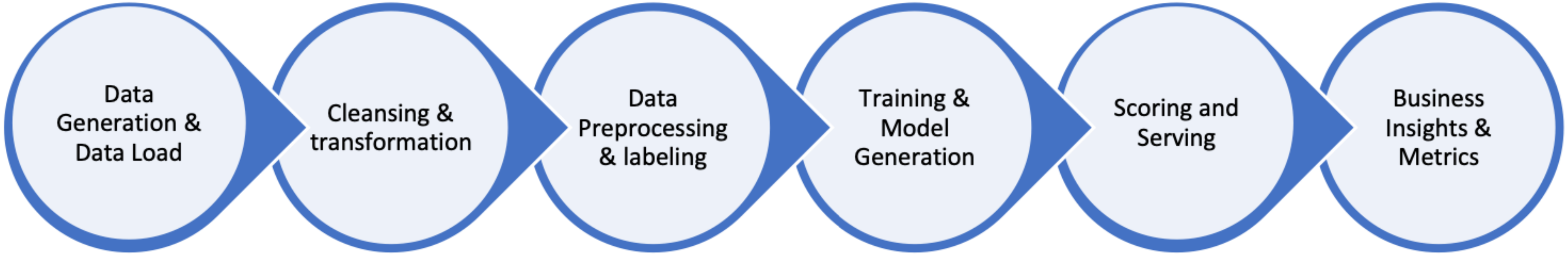
- Low-latency & real-time applications
- Poor / non-existing connectivity
- Legal restrictions & privacy
- Large amount of data – need for smart preprocessing





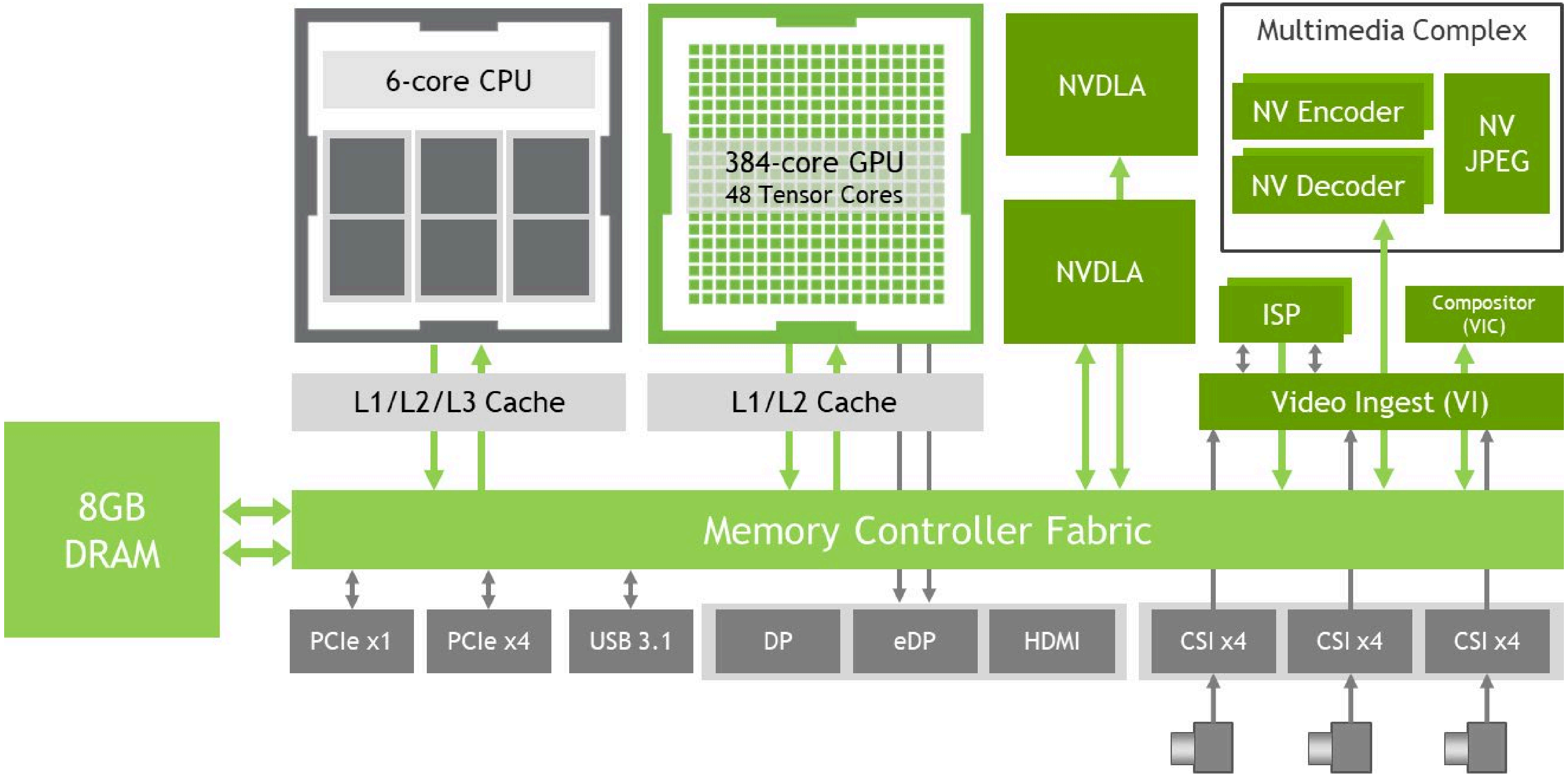
TPCxAI

- Benchmark for machine learning or data science systems
- 10 use cases modeled on retail datacenter
- End-to-end
- Scaling factor


Data Generation & Data Load
Cleansing & transformation
Data Preprocessing & labeling
Training & Model Generation
Scoring and Serving
Business Insights & Metrics

Source: TPC Express AI (TPCx-AI) Standard Specification Revision 1.0.1

Example: System on Chip devices

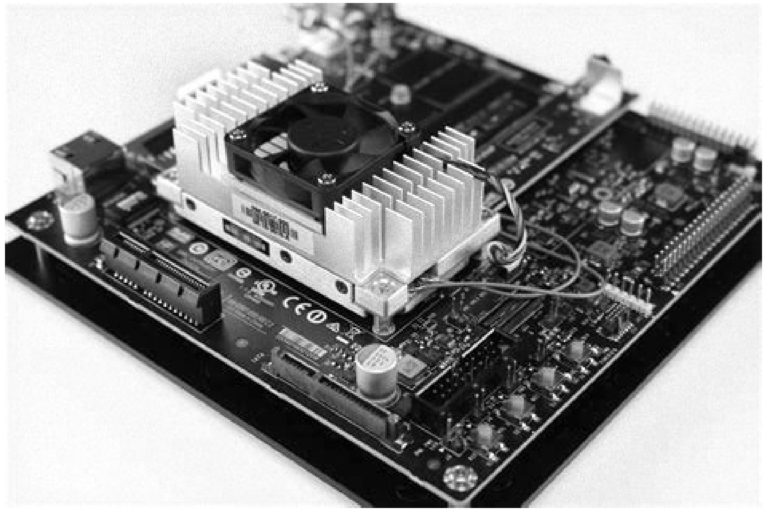


Common at edge

Heterogeneous

Memory shared between CPU and GPU

Experimental setup - hardware



Device	GPU	CPU	RAM	PWR	Price
TX2	NVIDIA Pascal, 256 CUDA Cores	NVIDIA Denver (2 Cores) & Arm Cortex A57 (4 Cores) @ 2.0 GHz	8 GB	15W	$399
Xavier	NVIDIA Volta, 512 CUDA Cores, 64 Tensor Cores	8 Cores ARM v8.2 64-bit @ 2.2 GHz	32 GB	30W	$699
Desktop	NVIDIA RTX 2070, 2304 CUDA Cores, 288 Tensor Cores	8 Cores Intel Core i7-6700K @ 4.0 GHz	16 GB (CPU), 8 GB (GPU)	∼ 550W	-

Experimental setup - workload

- Out-of-the-box TPCxAI
- Scaling factors: 1, 3
- Modifications:
    - Offloading data generation to an x86 system
    - Fixed parameter propagation error
    - Parallelised preprocessing stage of use case 8

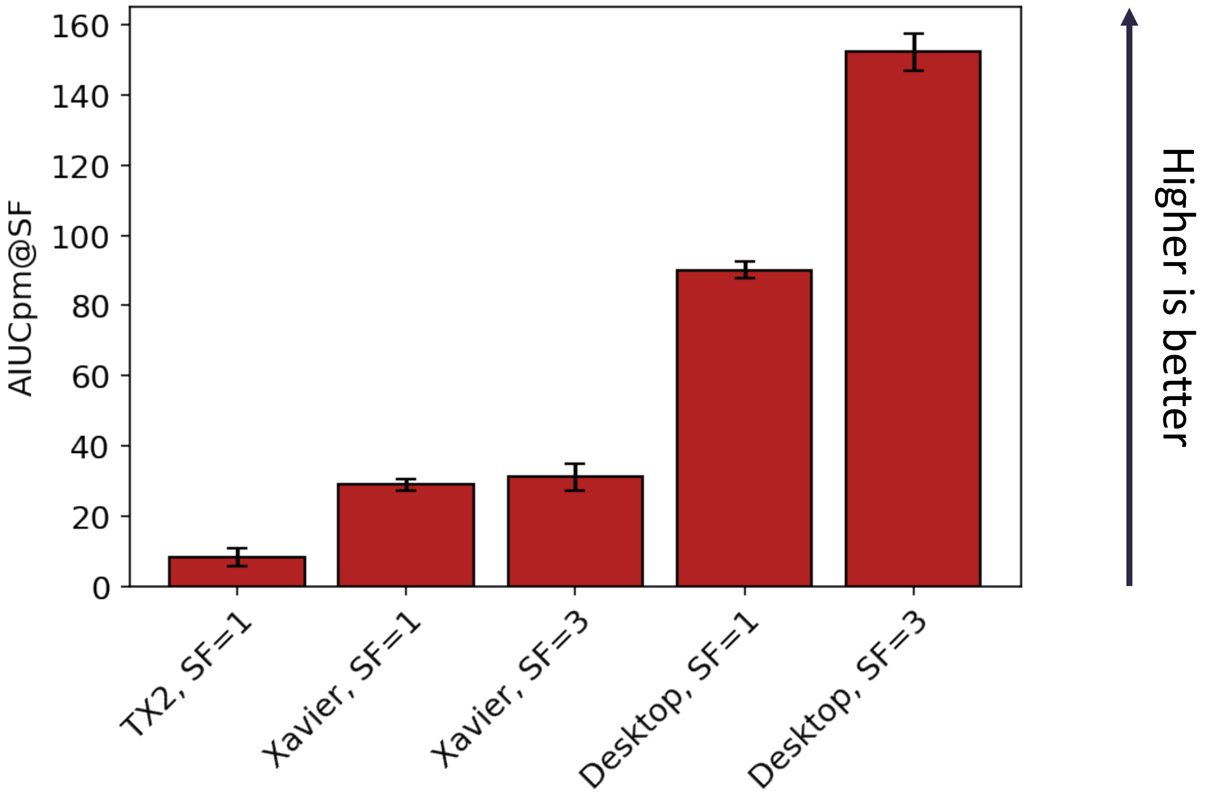
Experimental setup - metrics

- TPCxAI metrics
  - AIUCpm@SF – primary metric, workload-to-latency ratio
  - $/AIUCpm@SF

$$\text{AIUCpm@SF} = \frac{\mathrm{SF} * \mathrm{N} * 60}{\sqrt[4]{\mathrm{T}_{\mathrm{LD}} * \mathrm{T}_{\mathrm{PTT}} * \mathrm{T}_{\mathrm{PST}} * \mathrm{T}_{\mathrm{TT}}}}$$

- Power consumption
  - Tegrastats on Jetsons (CPU, GPU)
  - NVIDIA-SMI on Desktop (GPU only)

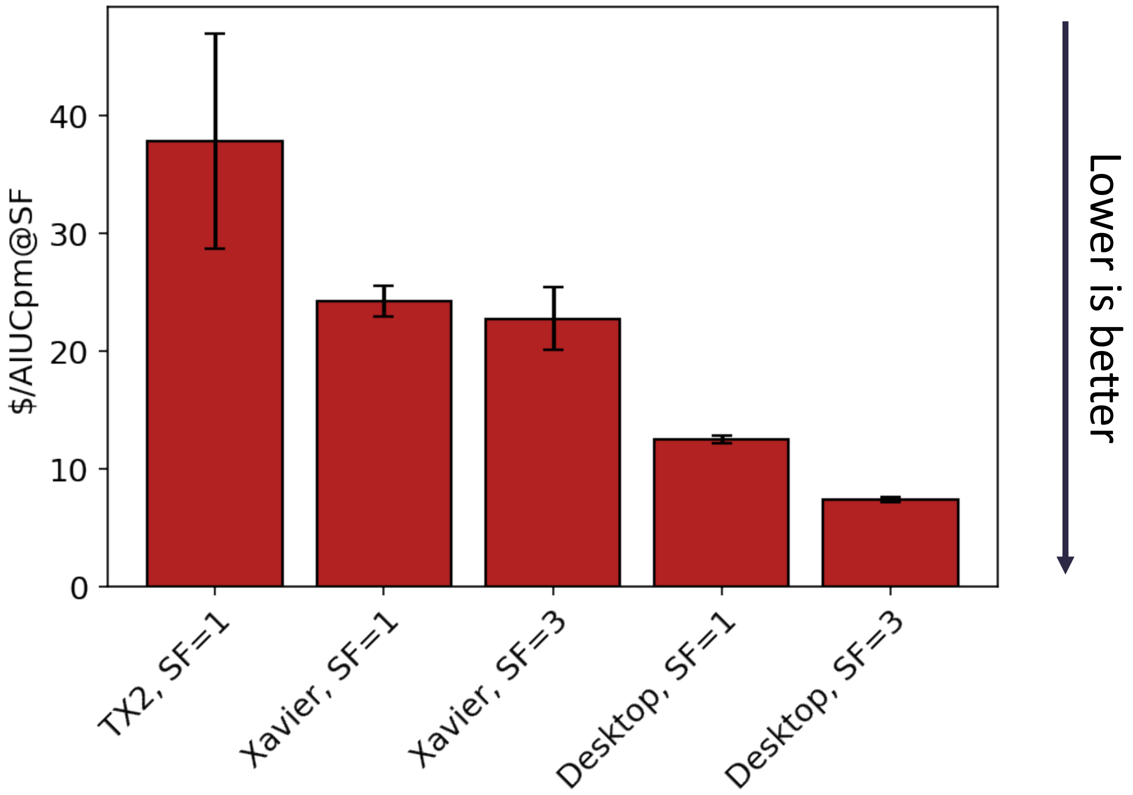
TPCxAI results



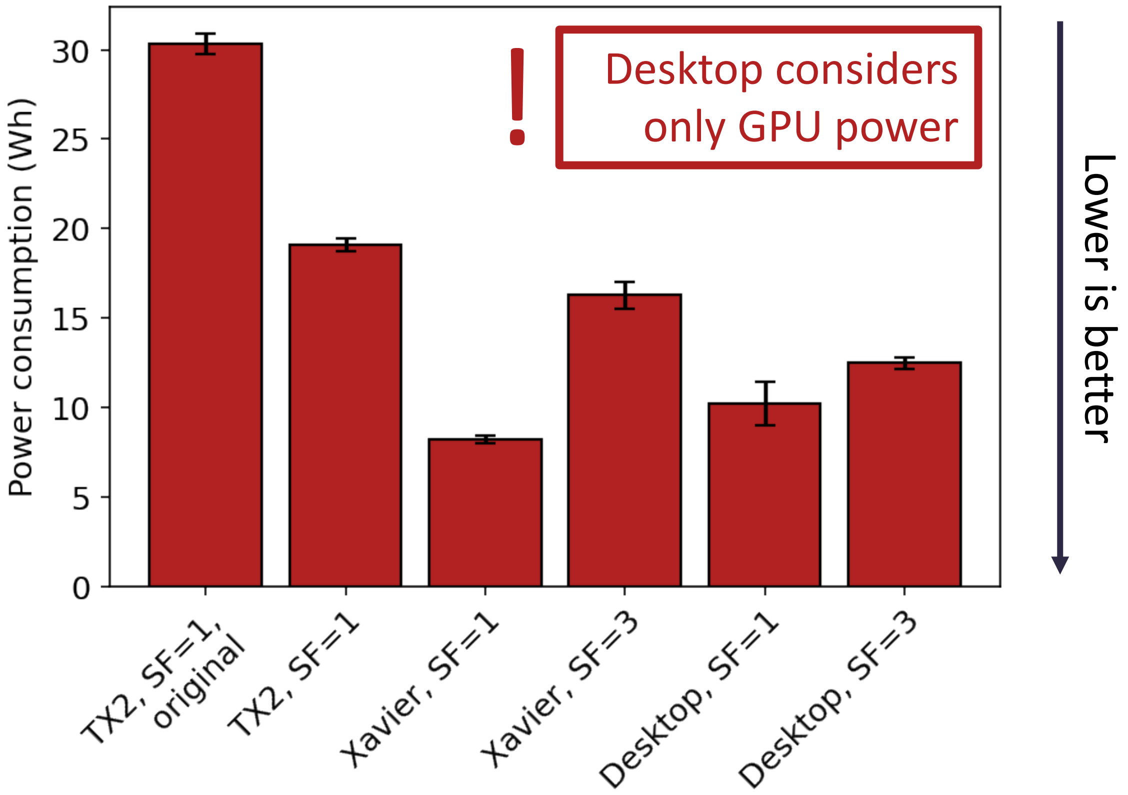
➔**Desktop outperforms Jetsons**

➔**But does not include network latencies**

What about cost- and power-efficiency?



Lower is better



! Desktop considers only GPU power

Lower is better

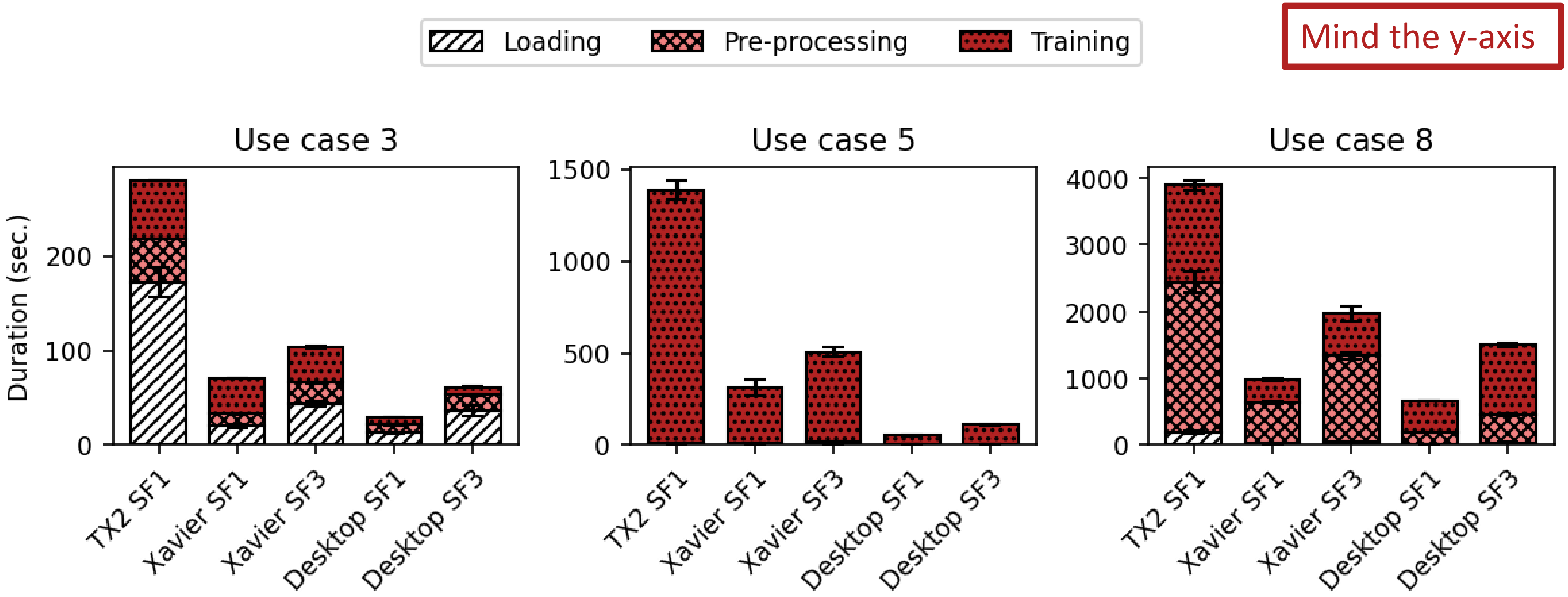
➔ **Desktop provides better price-to-performance ratio**

➔ **Xavier has better power consumption**

benchmark relies mostly on CPU

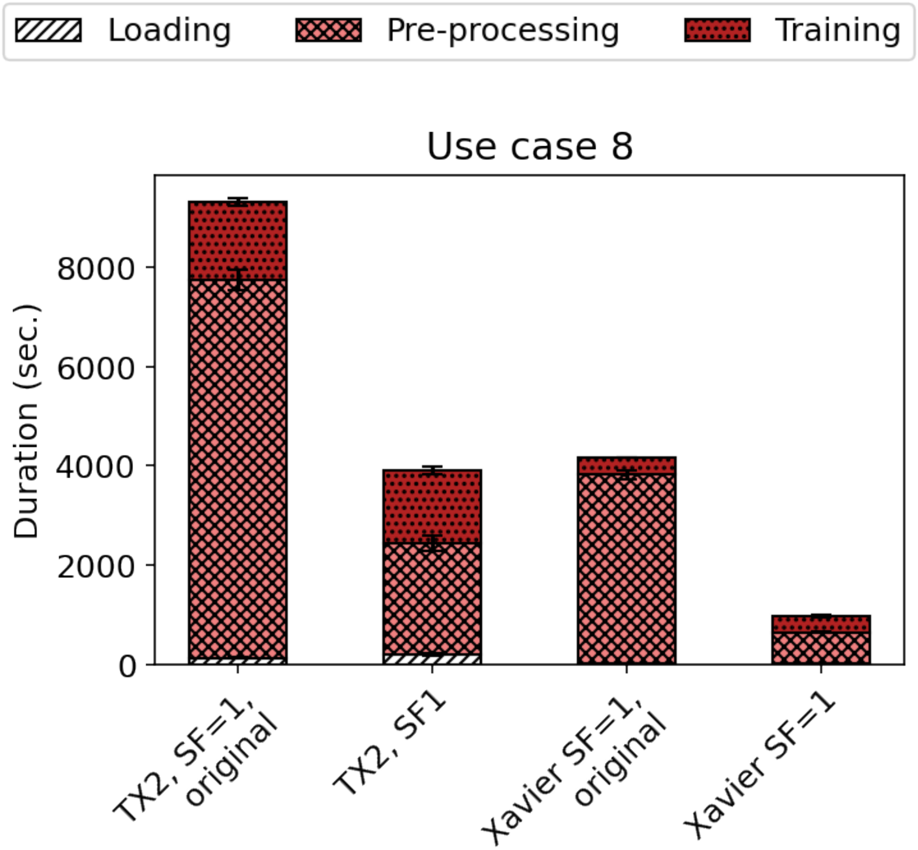
Zooming in on use cases

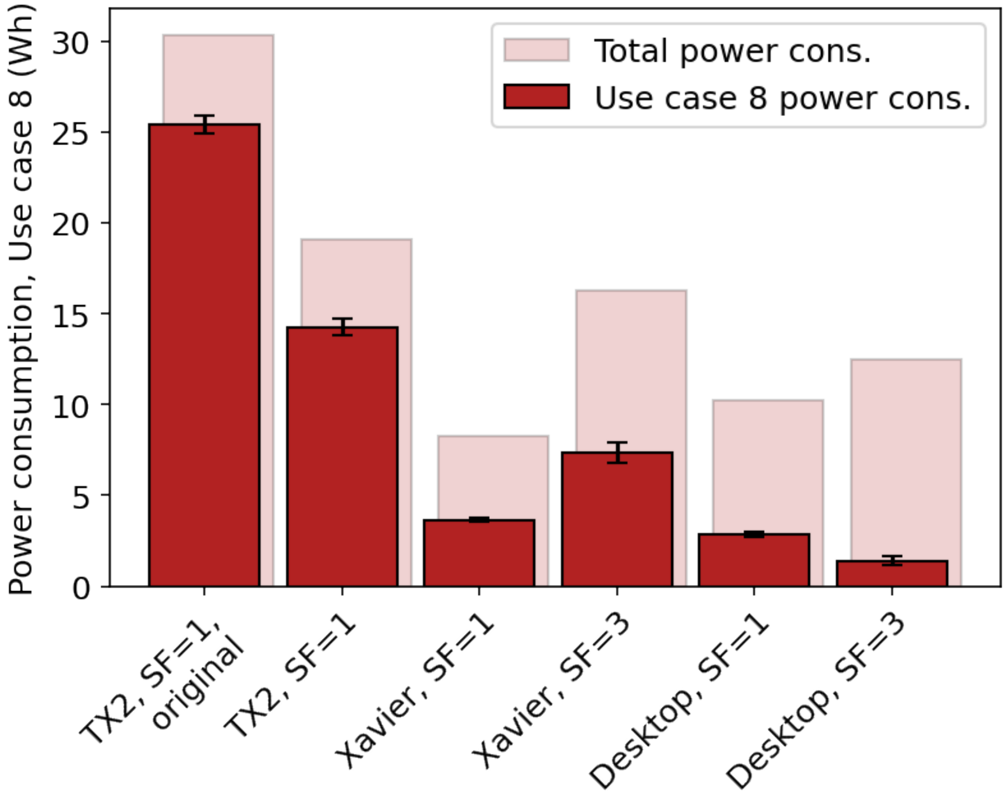
Mind the y-axis



Benchmark provides good variety and stresses different parts of the pipeline

Use case 8





➔ Most of the benchmark uses single thread to do data preprocessing

➔ Time to finish use case 8 reduced by 70-84% by parallelizing single statement (fixed to 3 threads)

Conclusion

Machine learning on Jetsons:

➔**Low memory can be limiting factor**

➔**Desktop's powerful CPU compensates for this**

➔**Xavier is very energy-efficient**

➔**Comparing Xavier and Desktop is hard**

TPCxAI for edge:

➔**Even the lowest scaling factor too high for TX2**

➔**Edge workloads are focused and closely coupled to hardware**