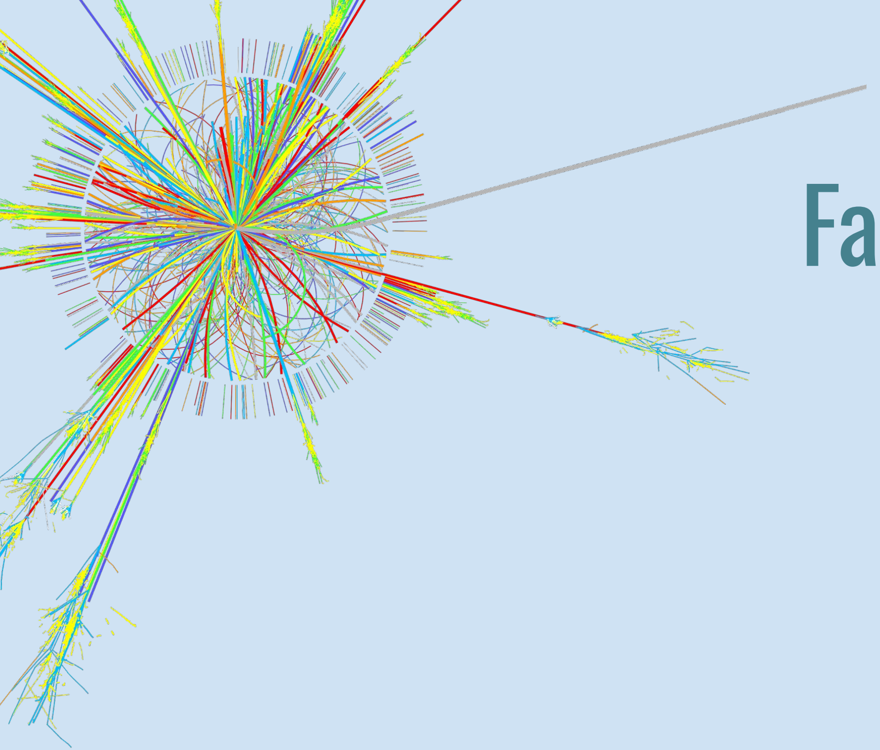
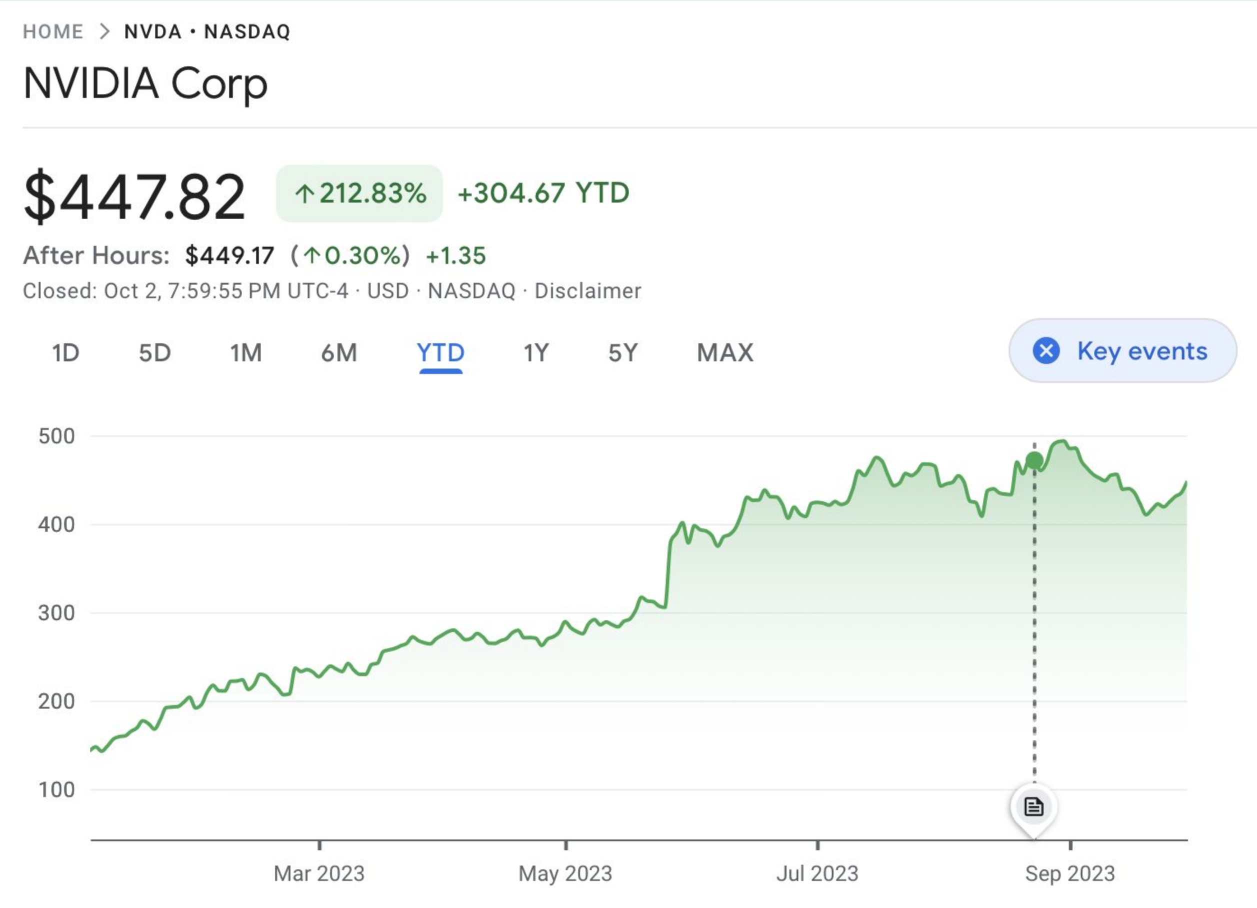


Fast and Furious AI-machines for physics at the LHC

Andrea Coccaro

Institute of Advanced Studies - Geneva

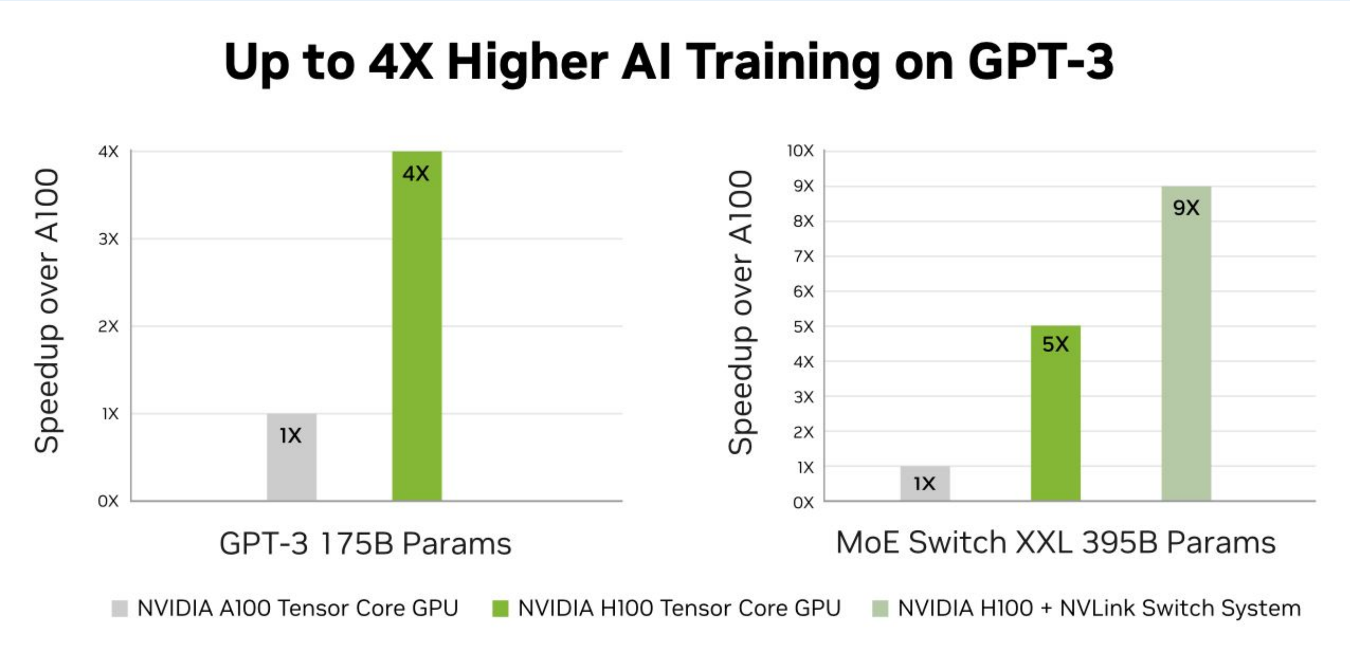


In the YTD timeframe

- Alphabet +50%
- Amazon +50%
- Apple +40%
- Microsoft +30%
- Nvidia +200%

Why?

A driver for training super-large models and behind all the recent hype on generative and LLM models



https://www.nvidia.com/en-us/data-center/hgx/

← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33711

Add your signature

Published
March 22, 2023

Why Pope Francis Is the Star of A.I.-Generated Photos

Francis has become a recurring favorite to show in incongruous situations, such as riding a motorcycle and attending Burning Man, in A.I.-generated images.

Eliezer Yudkowsky
@ESYudkowsky

Possible but hardly inevitable. It becomes moderately more likely as people call it absurd and fail to take precautions against it, like checking for sudden drops in the loss function and suspending training. Mostly, though, this is not a necessary postulate of a doom story.

> **Perry E. Metzger** @perrymetzger · Apr 25
> Eliezer and his acolytes believe it's inevitable AIs will go "foom" without warning, meaning, one day you build an AGI and hours or days later the thing has recursively self improved into godlike intelligence and then eats the world. Is this realistic?

5:44 PM · Apr 25, 2023 · **647K** Views

Ukraine war: Deepfake video of Zelenskyy telling Ukrainians to 'lay down arms' debunked

https://aiindex.stanford.edu/report/

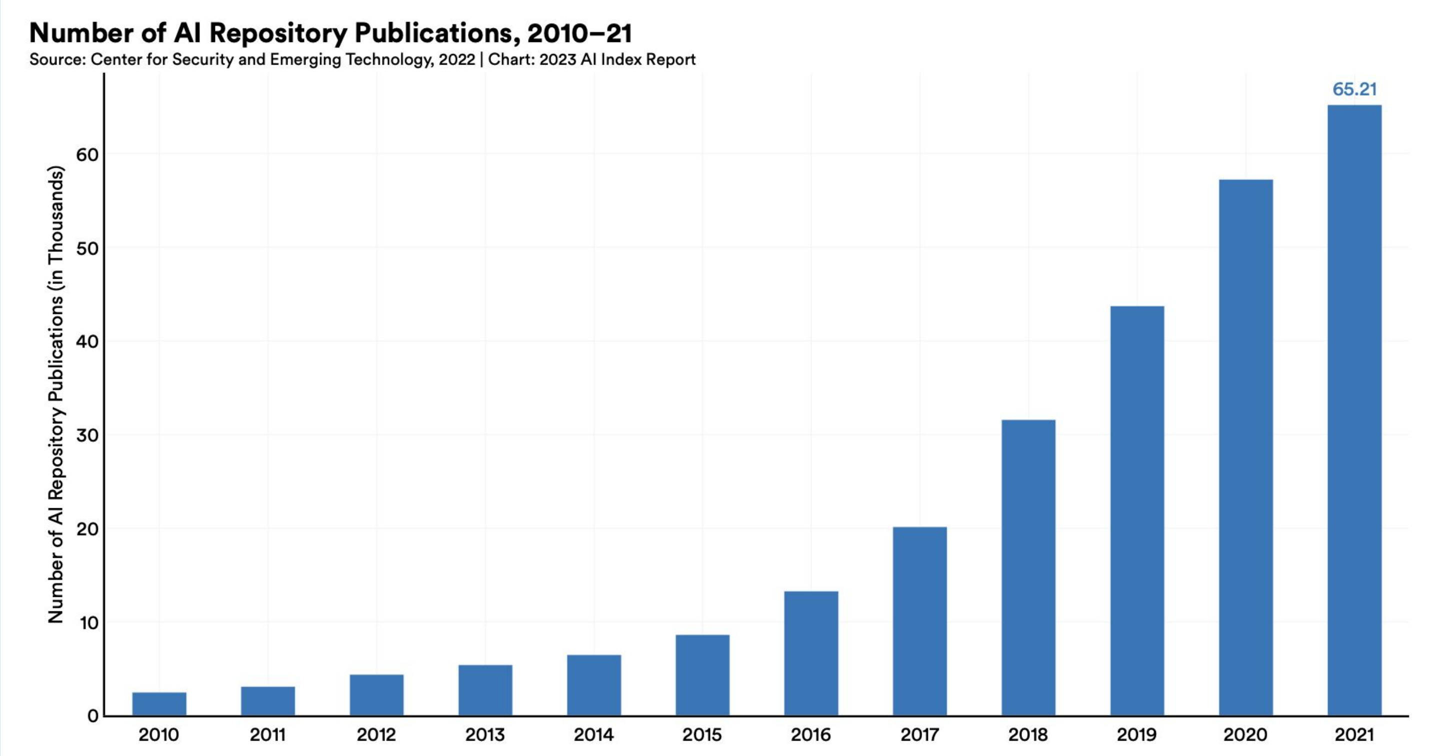
Artificial Intelligence Index Report 2023

Supporting Partners

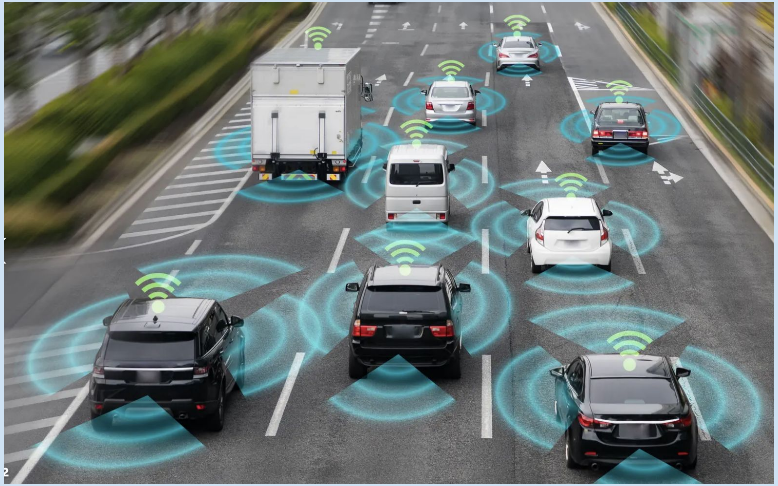
Google
NSF
NETBASE QUID
OpenAI
Open Philanthropy

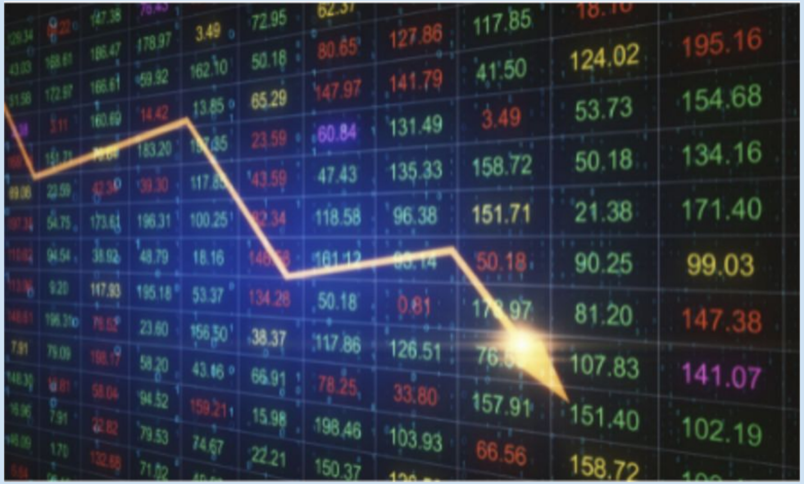
Analytics and Research Partners

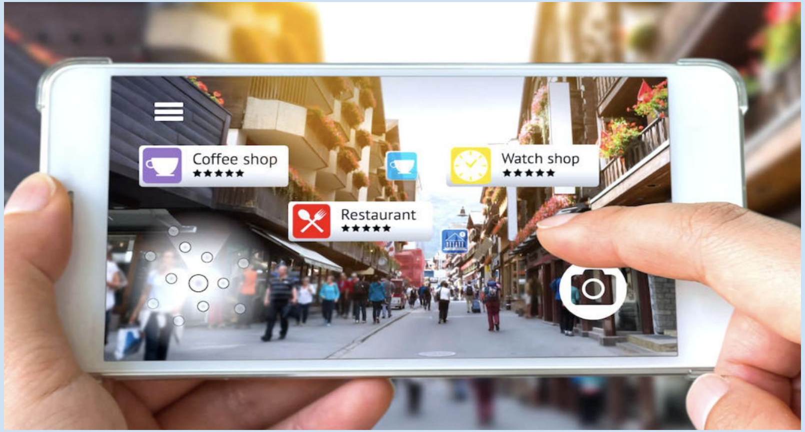
CODE
CRA Computing Research Association
CSET CENTER for SECURITY and EMERGING TECHNOLOGY
GitHub
govini
Lightcast
LinkedIn
McKinsey & Company
OECD.AI Policy Observatory
WiML Women in Machine Learning



Real-time deep learning







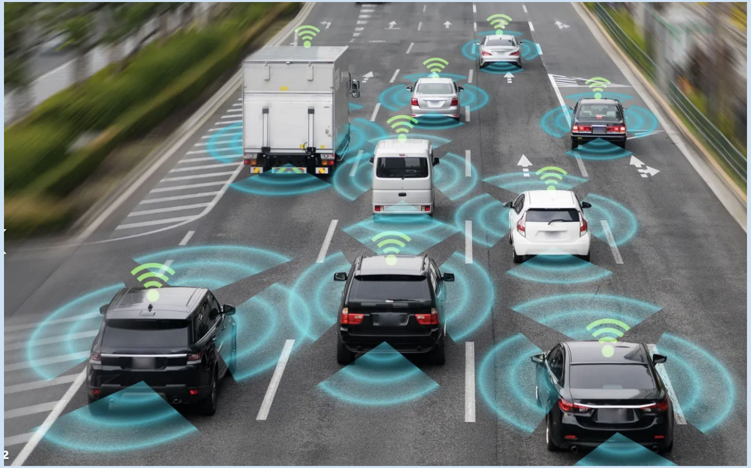
Example of real-time deep learning

Self-driving cars

- Single self-driving car can produce O(10) TB/day
- Number of US circulating cars O(200) millions
- With <1 % autonomous vehicles the generated amount of data is not manageable centrally

How to approach the problem

- Dedicated computing architectures in small dimensions and low-power consumption
- AI programs on-site since latency matters and communication with a central server will always result in a delay



https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html

FPGA vs. GPU for Deep Learning

FPGAs are an excellent choice for deep learning applications that require low latency and flexibility

Artificial intelligence (AI) is evolving rapidly, with new neural network models, techniques, and use cases emerging regularly. While there is no single architecture that works best for all machine and deep learning applications, FPGAs can offer distinct advantages over GPUs and other types of hardware in certain use cases.

https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html

FPGA vs. GPU for Deep Learning

FPGAs are an excellent choice for deep learning applications that require low latency and flexibility

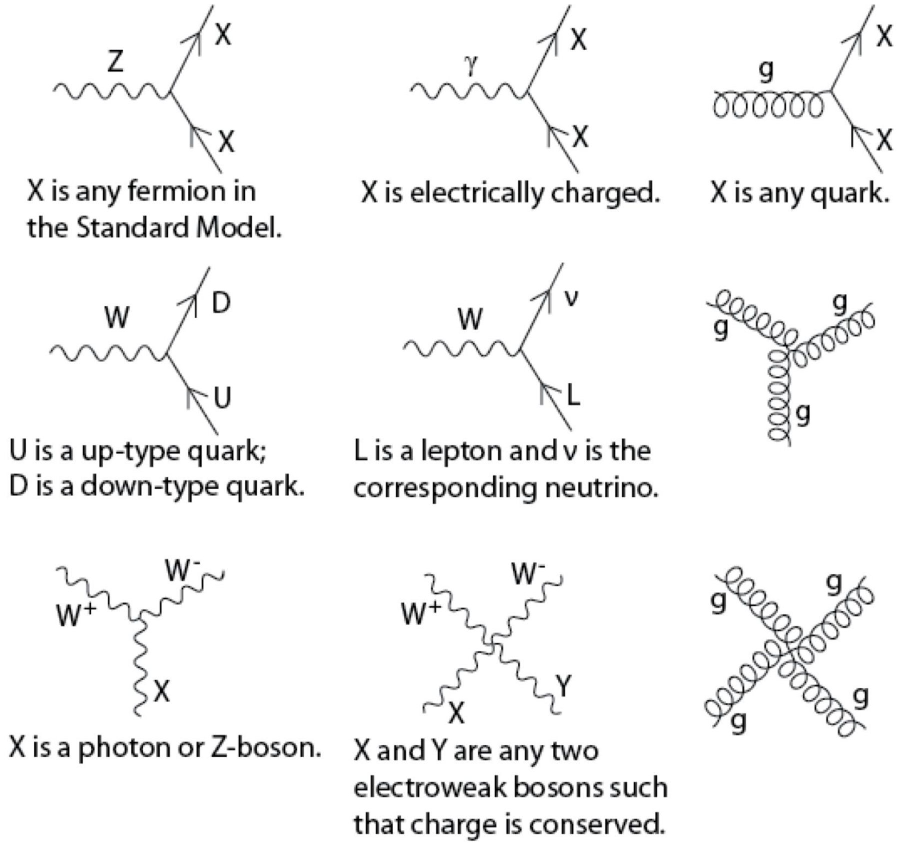
Artificial intelligence (AI) is evolving rapidly, with new neural network models, techniques, and use cases emerging regularly. While there is no single architecture that works best for all machine and deep learning applications, FPGAs can offer distinct advantages over GPUs and other types of hardware in certain use cases.

FPGAs vs GPUs

- Longer lifetime, more compatible with a typical car lifetime
- Lower power dissipation, no need of intense cooling
- Reduced electricity requirements
- Possible higher performance in terms of acceleration and throughput


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC 27 km

Standard Model Interactions
(Forces Mediated by Gauge Bosons)



X is any fermion in the Standard Model.

X is electrically charged.

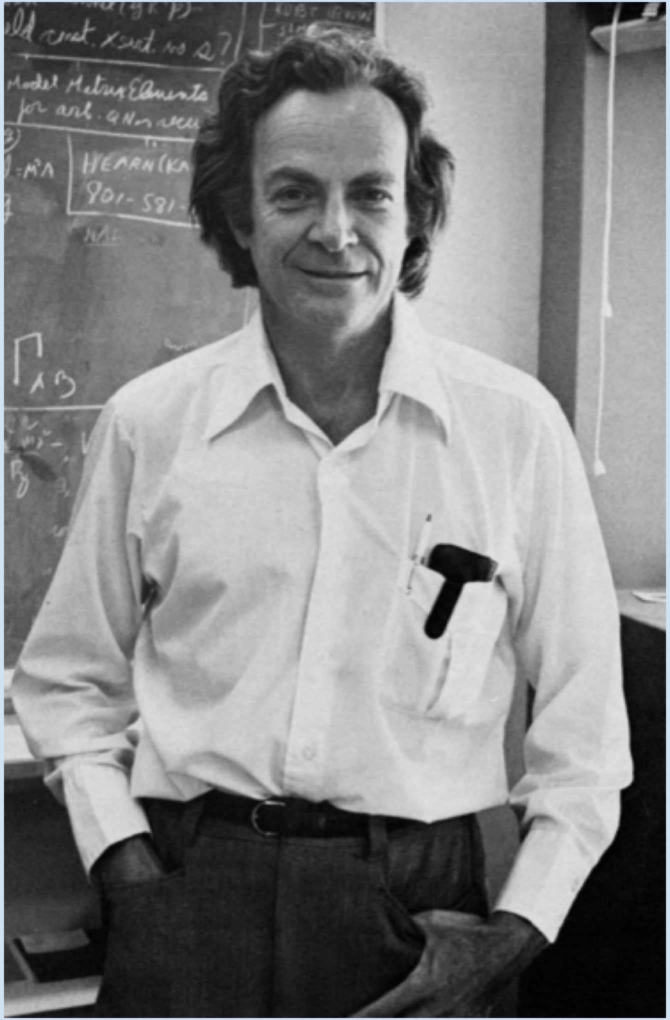
X is any quark.

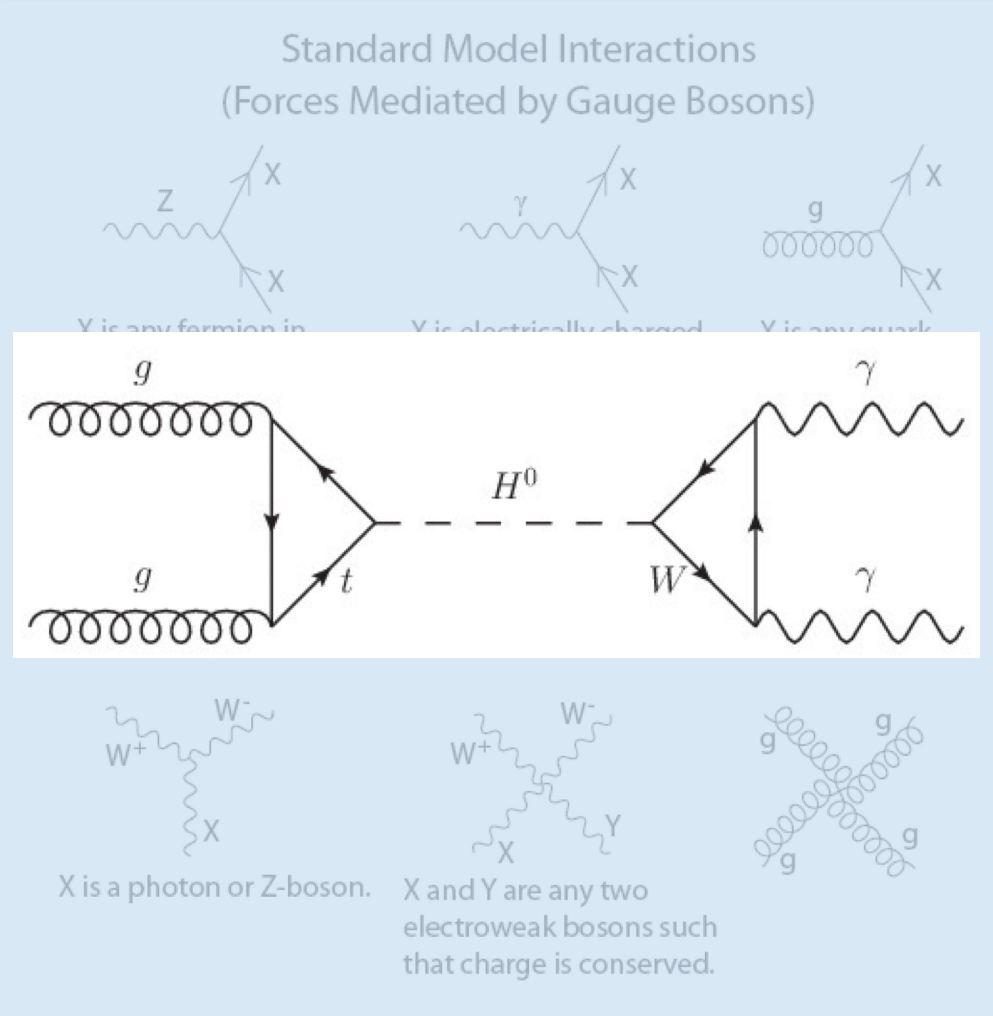
U is a up-type quark; D is a down-type quark.

L is a lepton and ν is the corresponding neutrino.

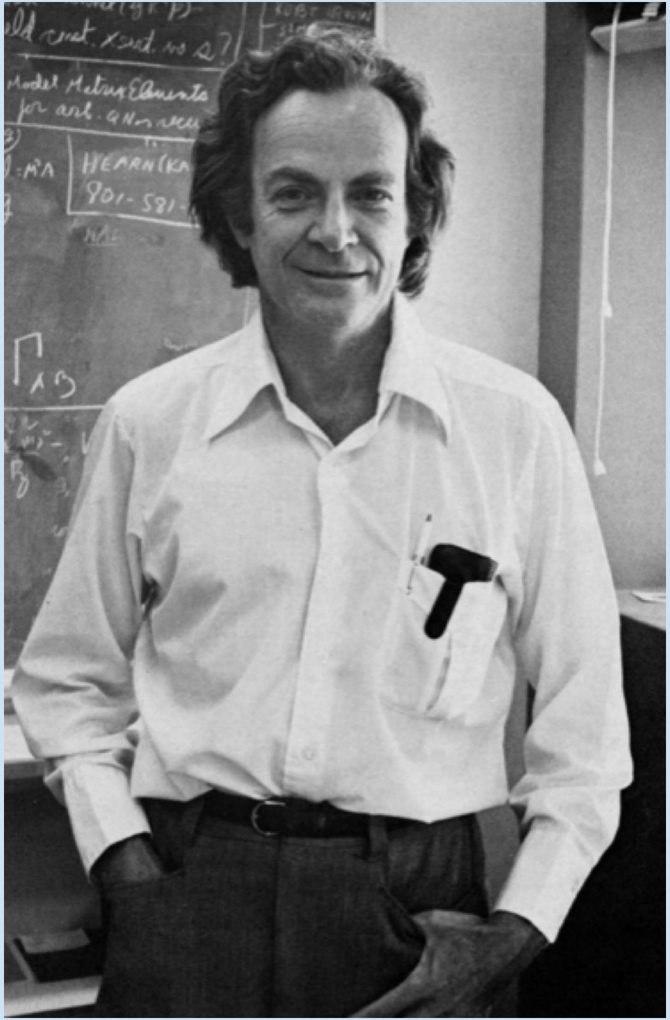
X is a photon or Z-boson.

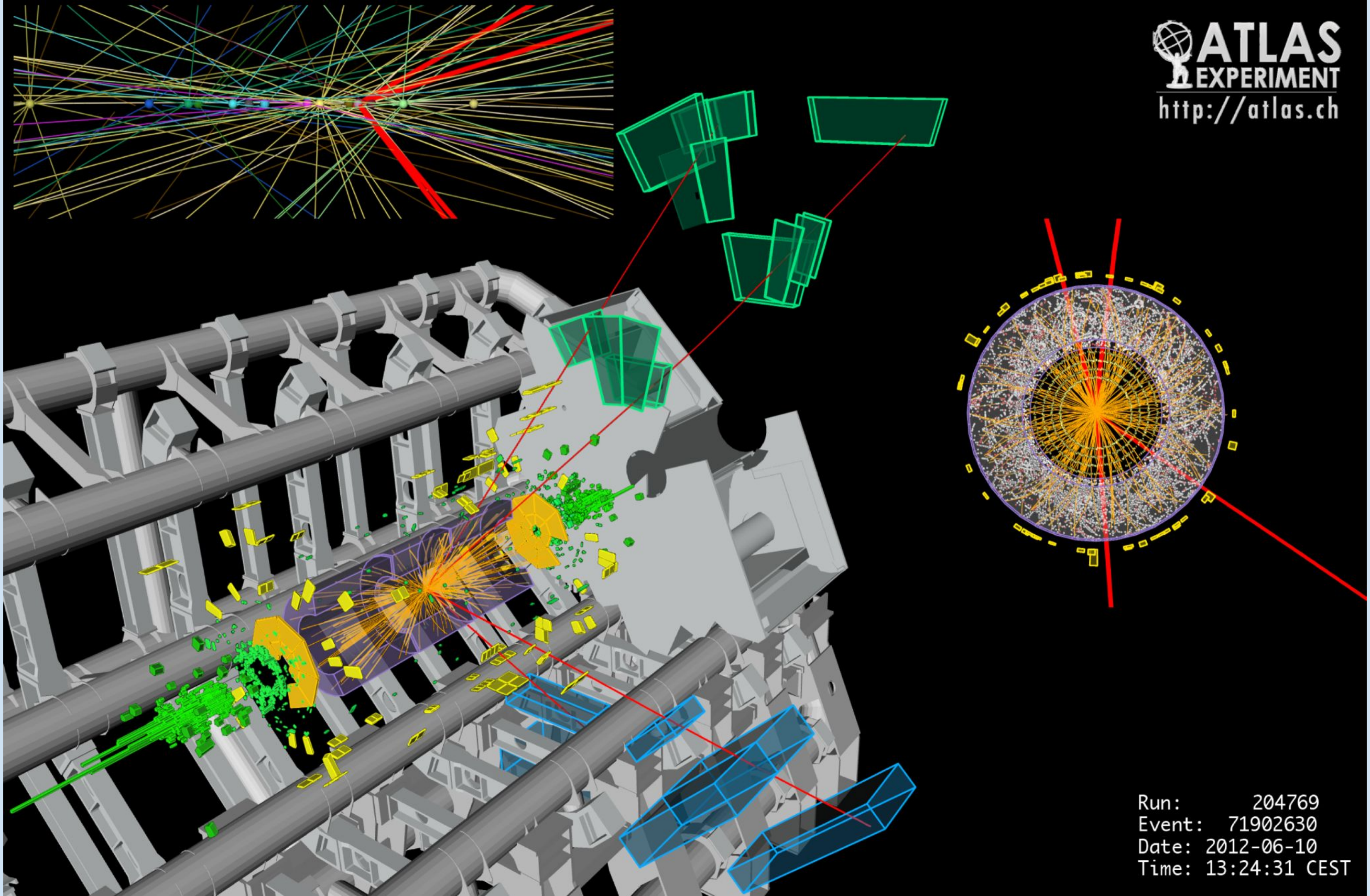
X and Y are any two electroweak bosons such that charge is conserved.



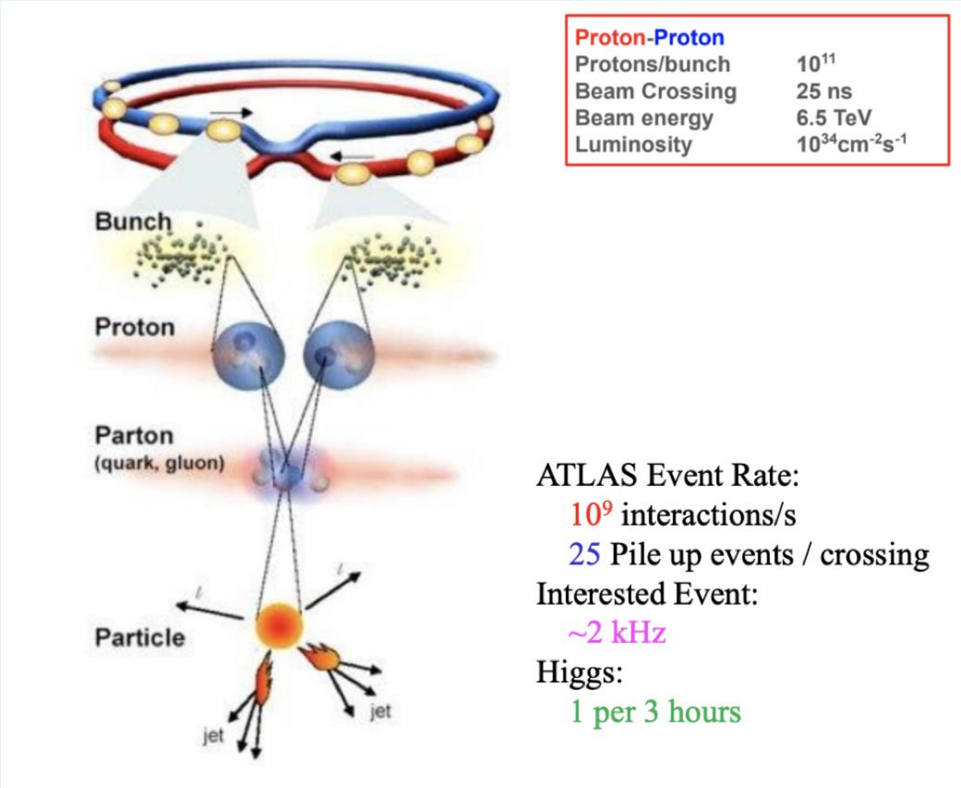


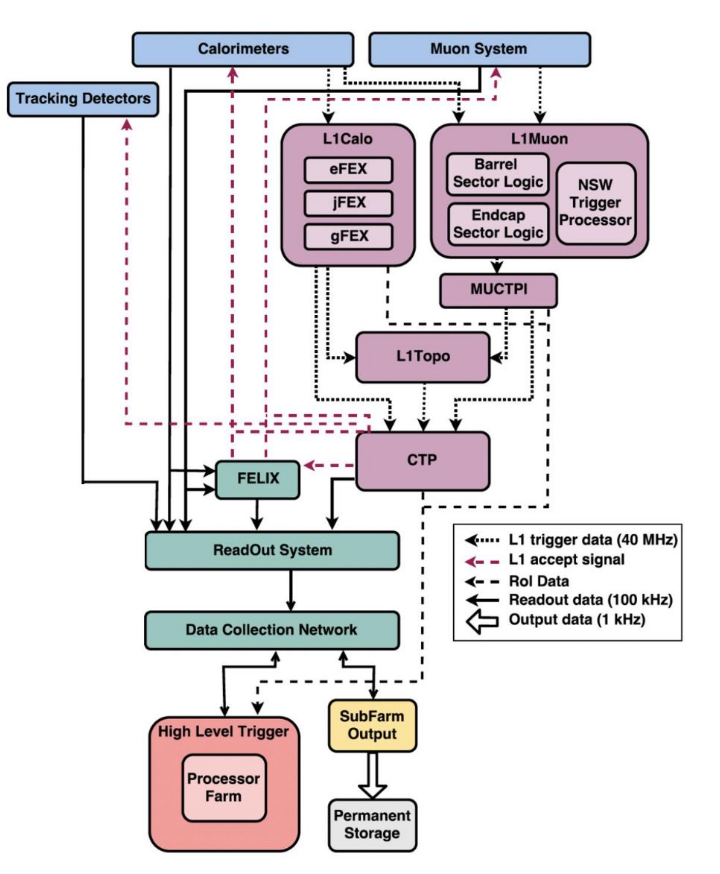
Standard Model Interactions
(Forces Mediated by Gauge Bosons)
Z
X
γ
g
H0
t
W
W+
W-
Y
X is a photon or Z-boson.
X and Y are any two electroweak bosons such that charge is conserved.




ATLAS
EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

Needle in a haystack





Trigger	Typical offline selection	Trigger Selection		L1 Peak Rate [kHz]	HLT Peak Rate [Hz]
		L1 [GeV]	HLT [GeV]	$L=2.0\times10^{34}$ $cm^{-2}s^{-1}$	
Single leptons	Single isolated μ, $p_T > 27$ GeV	20	26 (i)	16	218
	Single isolated tight e, $p_T > 27$ GeV	22 (i)	26 (i)	31	195
	Single μ, $p_T > 52$ GeV	20	50	16	70
	Single e, $p_T > 61$ GeV	22 (i)	60	28	20
	Single τ, $p_T > 170$ GeV	100	160	1.4	42
Two leptons	Two μ, each $p_T > 15$ GeV	2×10	2×14	2.2	30
	Two μ, $p_T > 23, 9$ GeV	20	22, 8	16	47
	Two very loose e, each $p_T > 18$ GeV	2×15 (i)	2×17	2.0	13
	One e & one μ, $p_T > 8, 25$ GeV	20 (μ)	7, 24	16	6
	One loose e & one μ, $p_T > 18, 15$ GeV	15, 10	17, 14	2.6	5
	One e & one μ, $p_T > 27, 9$ GeV	22 (e, i)	26, 8	21	4
	Two τ, $p_T > 40, 30$ GeV	20 (i), 12 (i) (+jets, topo)	35, 25	5.7	93
	One τ & one isolated μ, $p_T > 30, 15$ GeV	12 (i), 10 (+jets)	25, 14 (i)	2.4	17
	One τ & one isolated e, $p_T > 30, 18$ GeV	12 (i), 15 (i) (+jets)	25, 17 (i)	4.6	19
Three leptons	Three very loose e, $p_T > 25, 13, 13$ GeV	20, 2×10	24, 2×12	1.6	0.1
	Three μ, each $p_T > 7$ GeV	3×6	3×6	0.2	7
	Three μ, $p_T > 21, 2 \times 5$ GeV	20	20, 2×4	16	9
	Two μ & one loose e, $p_T > 2 \times 11, 13$ GeV	2×10 (μ)	2×10, 12	2.2	0.5
	Two loose e & one μ, $p_T > 2 \times 13, 11$ GeV	2×8, 10	2×12, 10	2.3	0.1
Signle photon	One loose γ, $p_T > 145$ GeV	24 (i)	140	24	47
Two photons	Two loose γ, each $p_T > 55$ GeV	2×20	2×50	3.0	7
	Two γ, $p_T > 40, 30$ GeV	2×20	35, 25	3.0	21
	Two isolated tight γ, each $p_T > 25$ GeV	2×15 (i)	2×20 (i)	2.0	15
Single jet	Jet ($R = 0.4$), $p_T > 435$ GeV	100	420	3.7	35
	Jet ($R = 1.0$), $p_T > 480$ GeV	111 (topo: $R = 1.0$)	460	2.6	42
	Jet ($R = 1.0$), $p_T > 450$ GeV, $m_{jet} > 45$ GeV	111 (topo: $R = 1.0$)	420, $m_{jet} > 35$	2.6	36
b−jets	One b ($\epsilon = 60\%$), $p_T > 285$ GeV	100	275	3.6	15
	Two b ($\epsilon = 60\%$), $p_T > 185, 70$ GeV	100	175, 60	3.6	11
	One b ($\epsilon = 40\%$) & three jets, each $p_T > 85$ GeV	4×15	4×75	1.5	14
	Two b ($\epsilon = 70\%$) & one jet, $p_T > 65, 65, 160$ GeV	2×30, 85	2×55, 150	1.3	17
	Two b ($\epsilon = 60\%$) & two jets, each $p_T > 65$ GeV	4×15, $\lvert\eta\rvert < 2.5$	4×55	3.2	15
Multijets	Four jets, each $p_T > 125$ GeV	3×50	4×115	0.5	16
	Five jets, each $p_T > 95$ GeV	4×15	5×85	4.8	10
	Six jets, each $p_T > 80$ GeV	4×15	6×70	4.8	4
	Six jets, each $p_T > 60$ GeV, $\lvert\eta\rvert < 2.0$	4×15	6×55, $\lvert\eta\rvert < 2.4$	4.8	15
E_T^{miss}	$E_T^{miss} > 200$ GeV	50	110	5.1	94
B-physics	Two μ, $p_T > 11, 6$ GeV, $0.1 < m(\mu,\mu) < 14$ GeV	11, 6	11, 6 (di-μ)	2.9	55
	Two μ, $p_T > 6, 6$ GeV, $2.5 < m(\mu,\mu) < 4.0$ GeV	2×6 (J/ψ, topo)	2×6 (J/ψ)	1.4	55
	Two μ, $p_T > 6, 6$ GeV, $4.7 < m(\mu,\mu) < 5.9$ GeV	2×6 (B, topo)	2×6 (B)	1.4	6
	Two μ, $p_T > 6, 6$ GeV, $7 < m(\mu,\mu) < 12$ GeV	2×6 (Υ, topo)	2×6 (Υ)	1.2	12
Main Rate				86	1750
B-physics and Light States Rate					200

Real-time event selection at the LHC

With a triggerless acquisition system

- 40 MHz interaction rate with O(1 MB/event)
- 40 TB/s scaling to O(10 EB/year)

Facebook in 2014

- 600 TB/day scaling to O(1 EB/year)
- Clearly a different business model compared to optimising the research output of the largest scientific endeavour

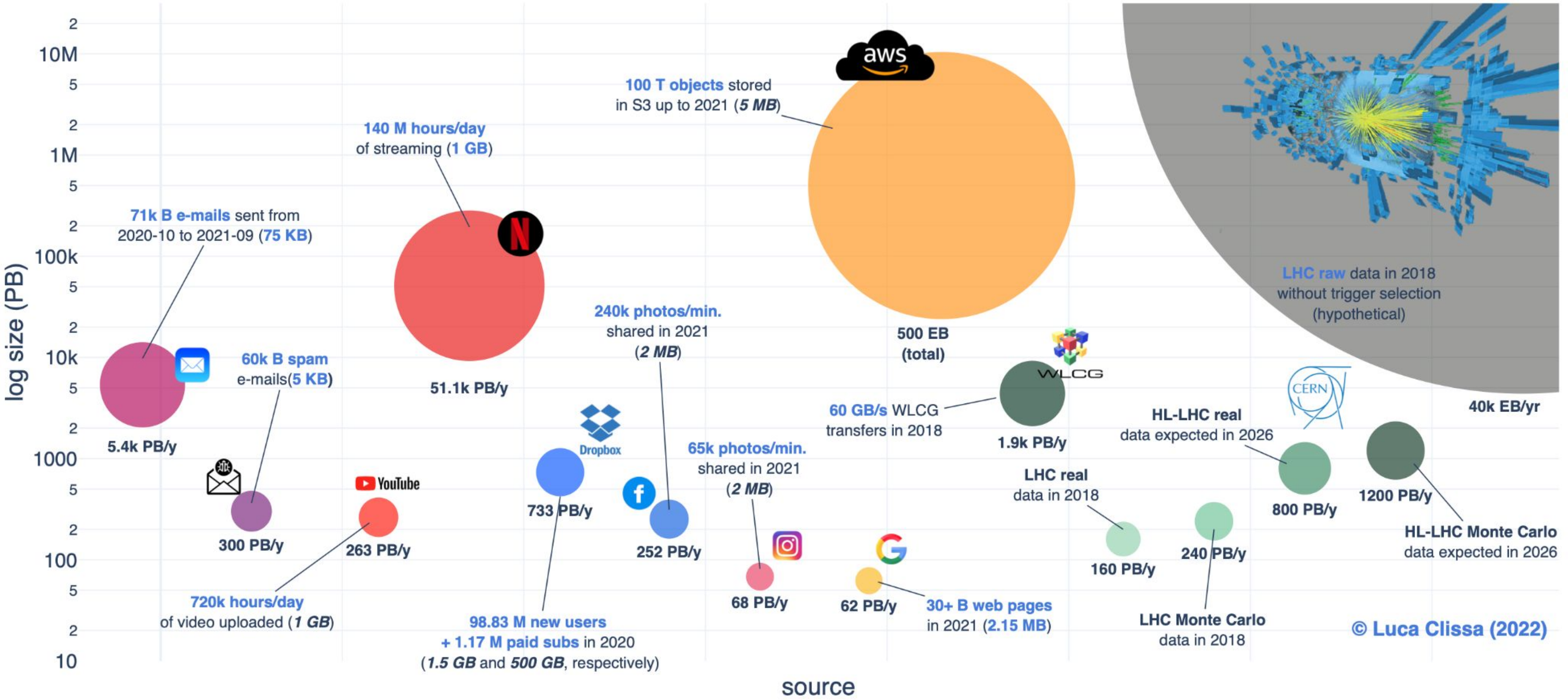
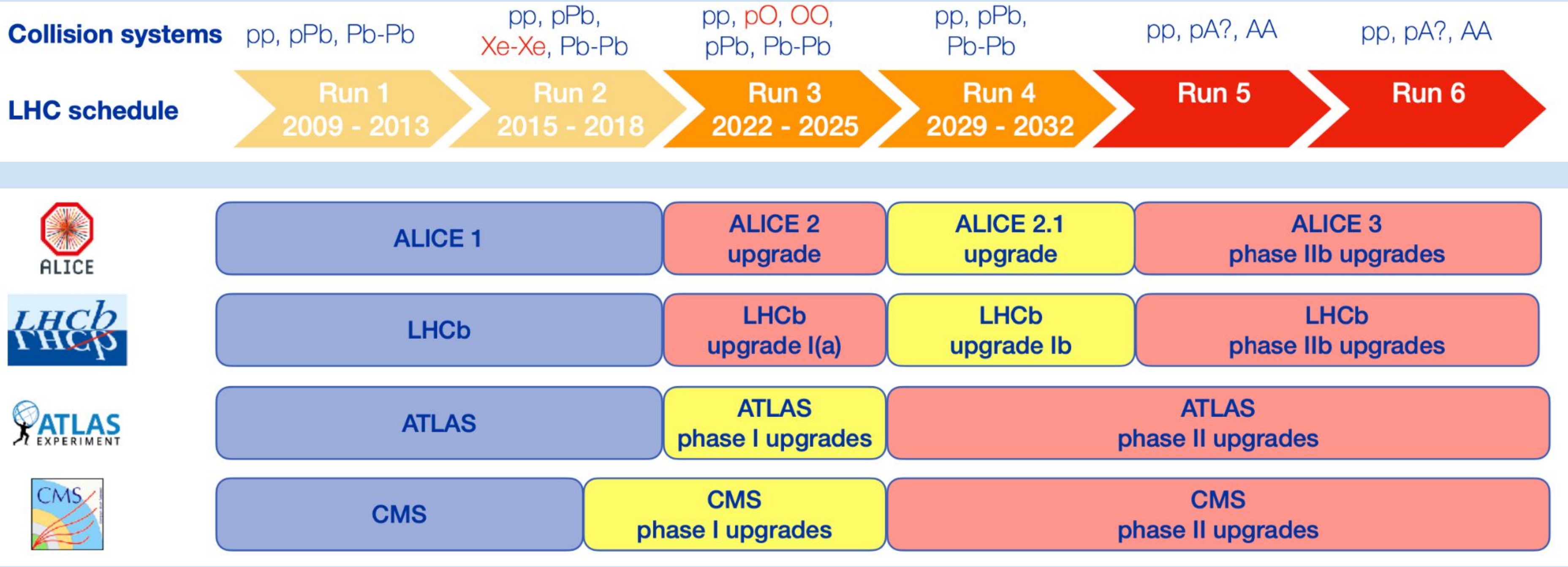


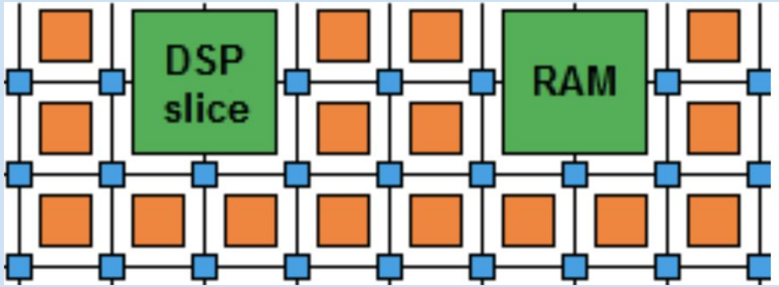
Figure 2.3: **Big Data sizes.** Bubble plot of the orders of magnitude of data produced by important big data players. The balloon areas illustrate the amount of data and the text annotations highlight the key factors considered in the estimates. Average per-unit sizes are reported in parentheses, where italic indicates measures reconstructed based on likely assumptions because no references were found.

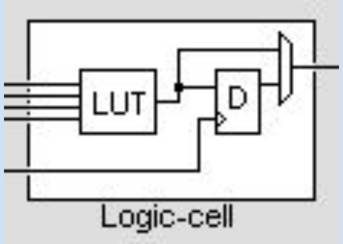
BigData2021



- How will we be triggering events in 2029+?
- Which design choices of the data acquisition system?
- How much AI will help?

As a community we need to provide answers to these questions ~now





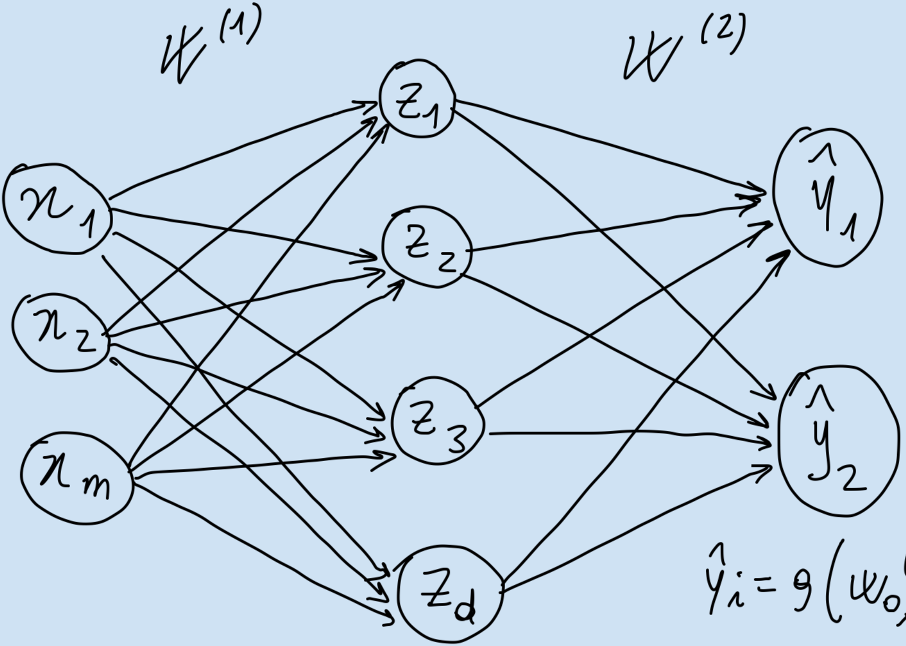
- Logic cell ⇒ A small look-up table with a D flip-flop
- Digital Signal Processors (DSPs) ⇒ logic units for multiplications
- Random-Access memories (RAMs) ⇒ embedded memory elements

Two ways to interact with FPGA programming

- Low-level programming languages to describe electronic circuits (HDL)
- High-level synthesis from C/C++ code (Vivado HLS)

Neural network inference

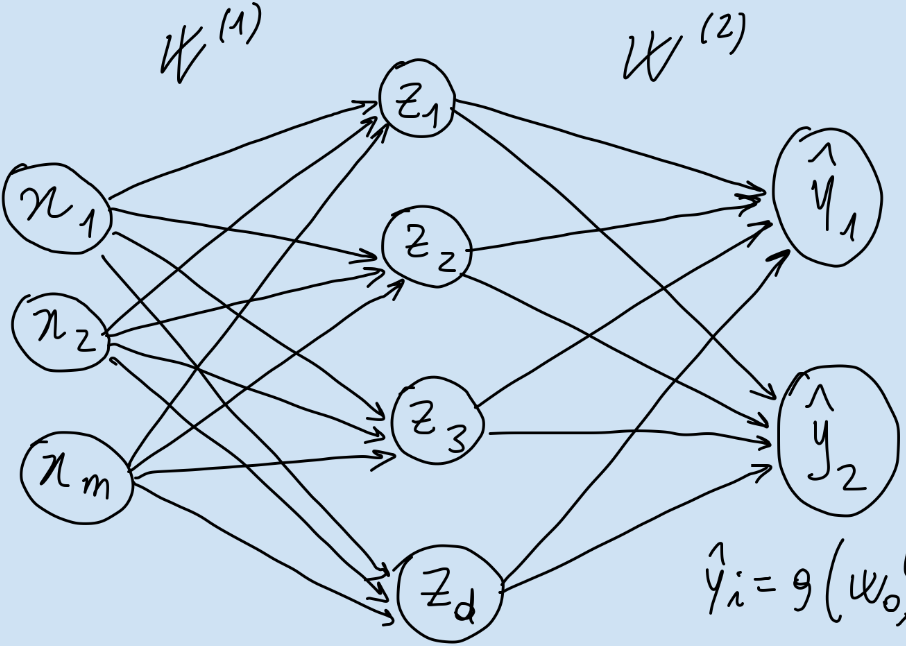
- Addition: logic cells
- Multiplication: DSPs
- Activation function: precomputed in RAMs



$$\hat{y}_i = g\left(w_{0,i}^{(2)} + \sum_{j=1}^{d_1} g(z_j)\, w_{j,i}^{(2)}\right)$$

$$z_i = w_{0,i}^{(1)} + \sum_{j=1}^{m} x_j\, w_{j,i}^{(1)}$$

Neural network inference



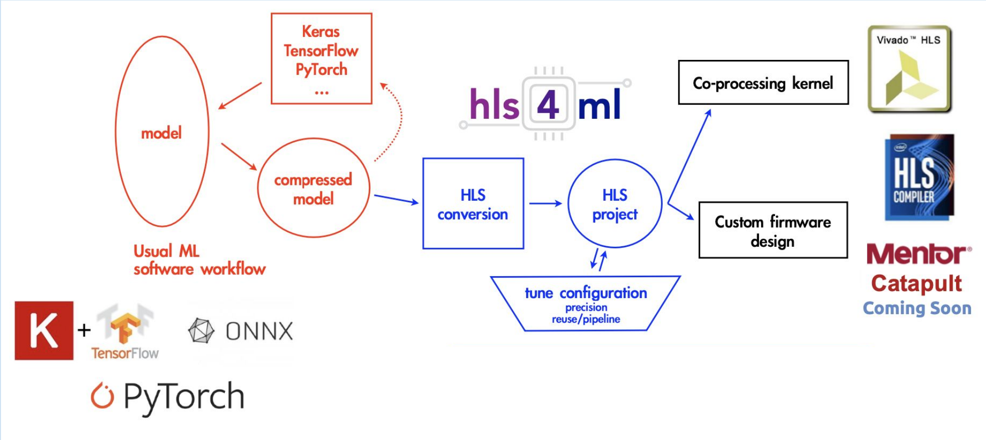
$$\hat{y}_i = g\left(w_{0,i}^{(2)} + \sum_{j=1}^{d_1} g(z_j)\, w_{j,i}^{(2)}\right)$$

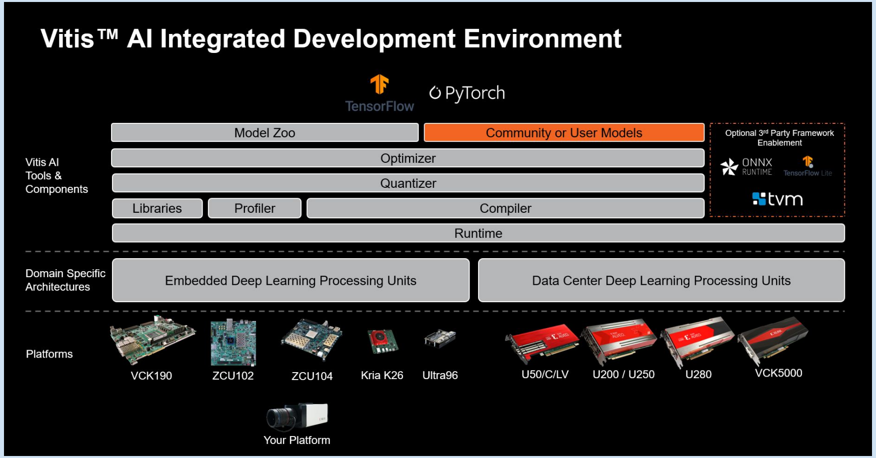
$$z_i = w_{0,i}^{(1)} + \sum_{j=1}^{m} x_j\, w_{j,i}^{(1)}$$

- Addition: logic cells
- Multiplication: DSPs
- Activation function: precomputed in RAMs

A NN can be deployed in a FPGA but not straightforward to understand the needed resources (DSPs, RAMs). Plus transferring rate, latency requirements, etc. Studies are needed

NN deployment into FPGAs



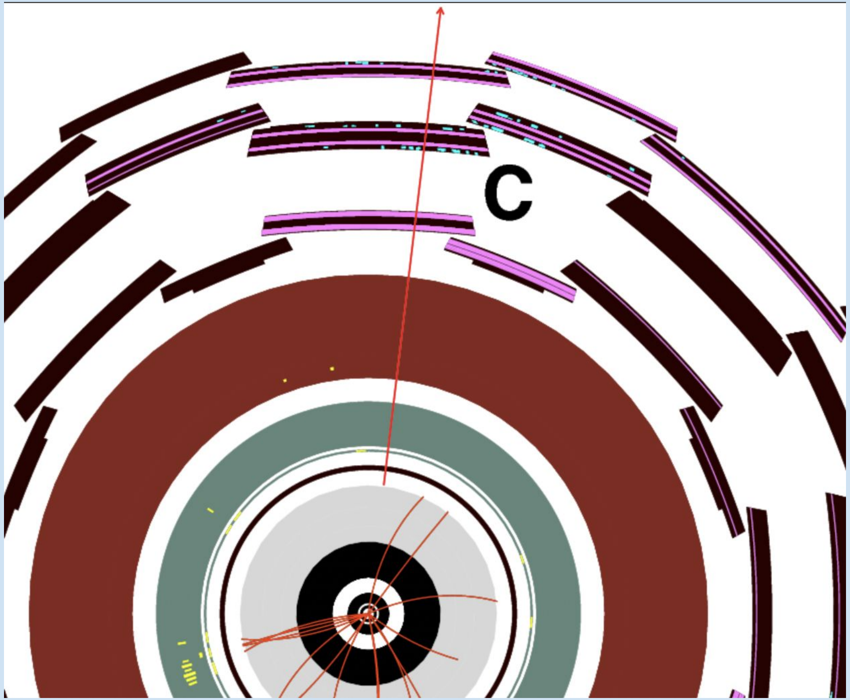


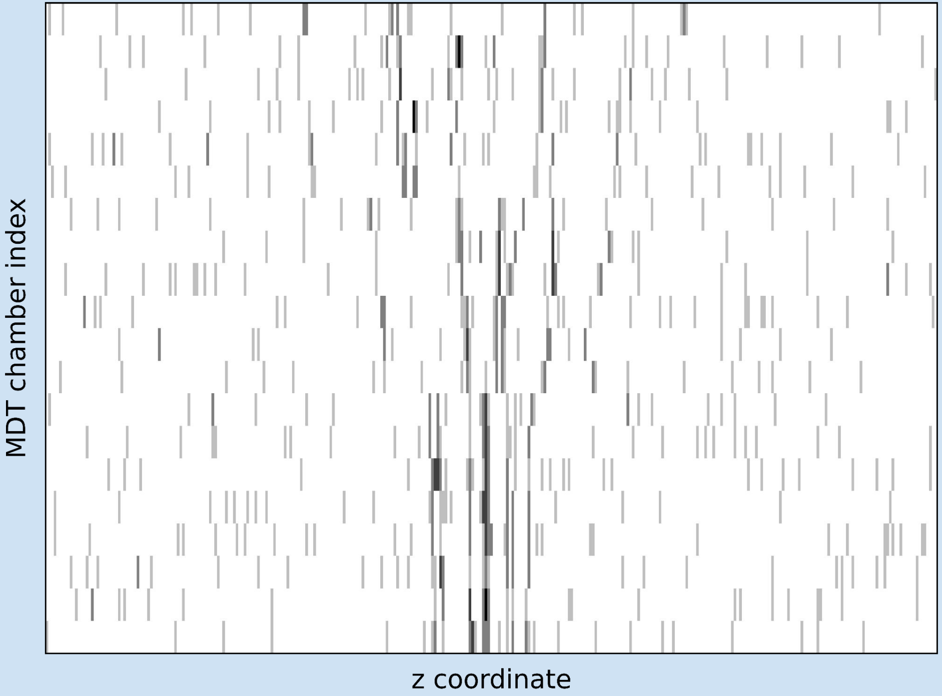
https://fastmachinelearning.org/

Vitis-AI

Physics case - Example #1

Fast inference on FPGAs for triggering on neutral LLPs

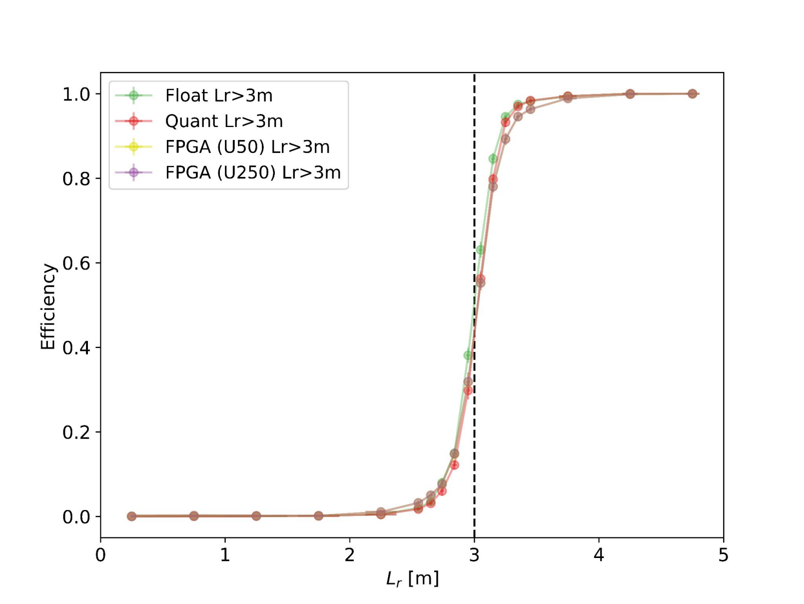




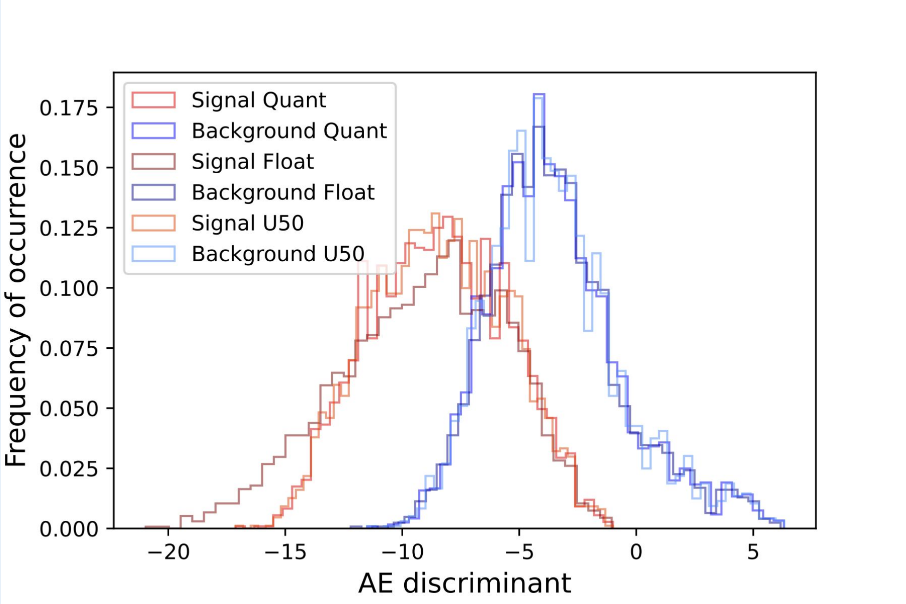
arXiv:2307.05152

Physics case - Example #1

Two models: a CNN to regress the LLP decay length position and an AE to detect anomalies



CNN model

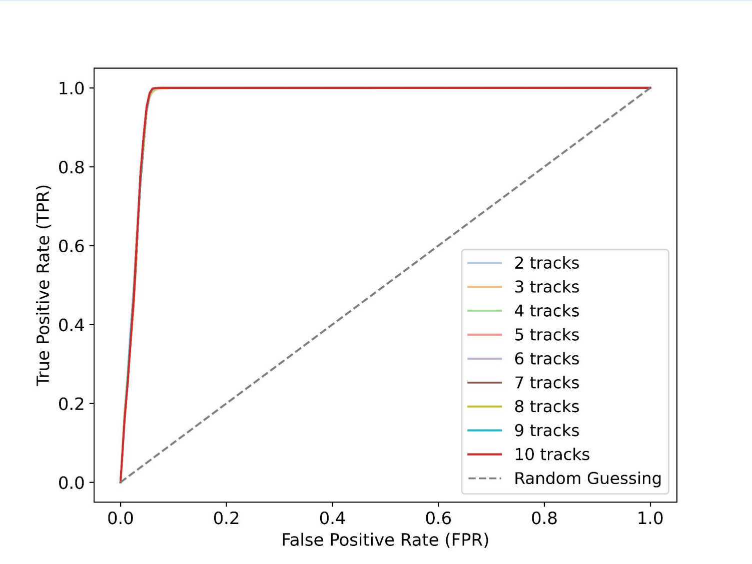


AE model

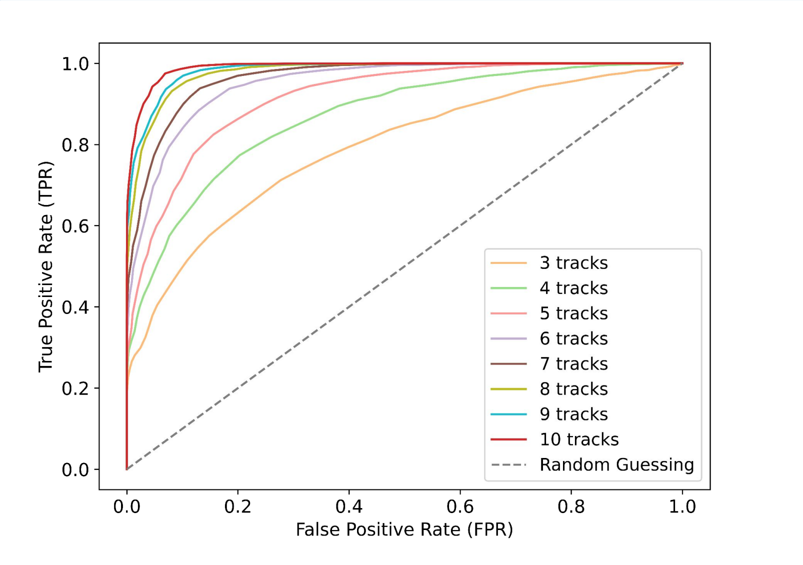
arXiv:2307.05152

Physics case - Example #1

Two models: a CNN to regress the LLP decay length position and an AE to detect anomalies



CNN model



AE model

arXiv:2307.05152

Physics case - Example #1

Two models: a CNN to regress the LLP decay length position and an AE to detect anomalies

	CPU	GPU	U50	U250
Inference time [ms]	5.1 ± 1.1	1.0 ± 0.1	3.7 ± 0.1	3.1 ± 0.4
Throughput [fps]	302 ± 4	9930 ± 187	950 ± 5	553 ± 4

Table 1: Inference time in ms and throughput in frames per second for the CNN model on different target architectures. The results include the actual deployment of the model on the FPGA U50 and U250 accelerator cards.

Real measurement on FPGA devices mounted on dedicated nodes.

arXiv:2307.05152

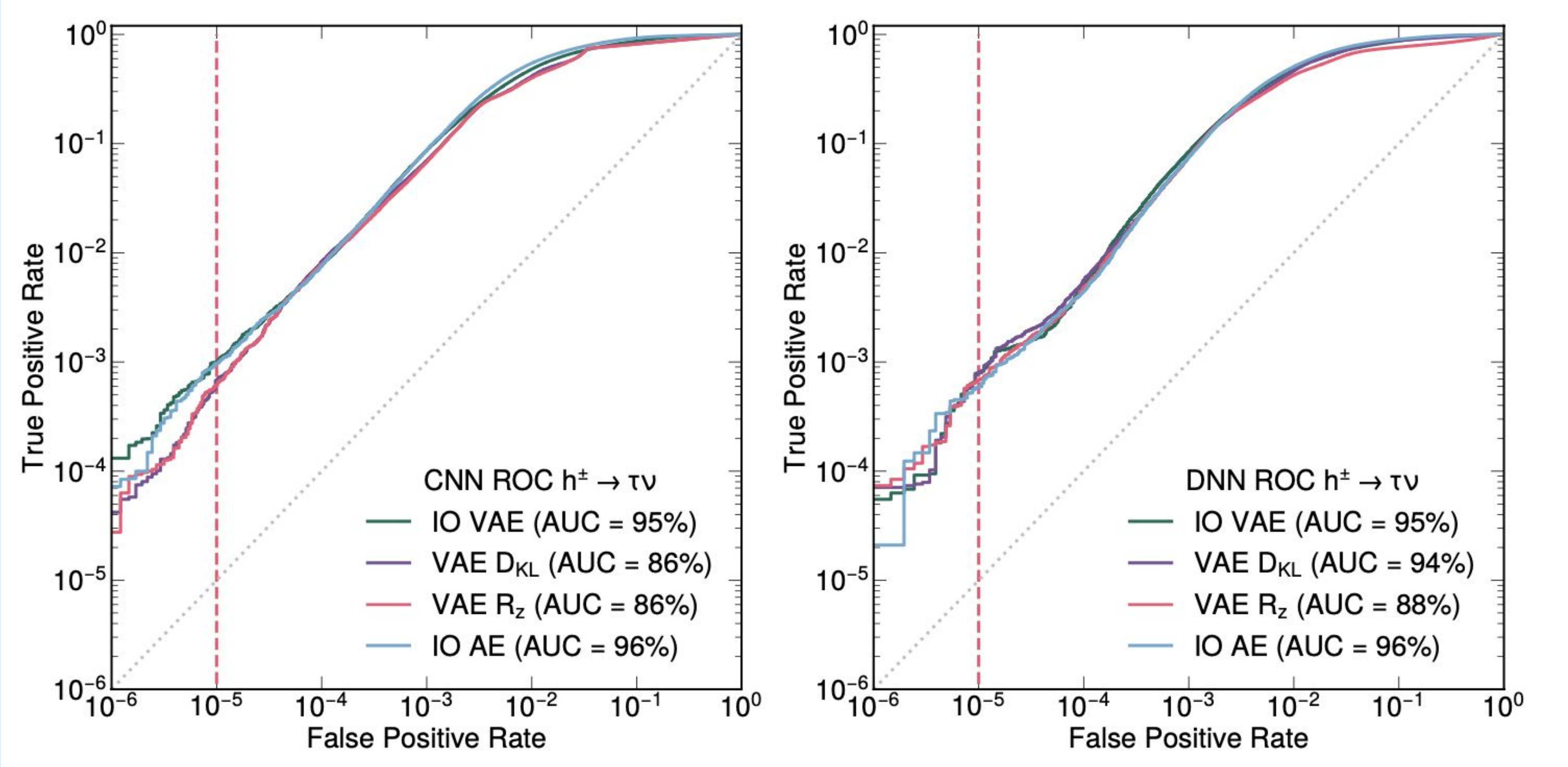
Physics case - Example #2

New physics detection with autoencoders directly at L1 trigger

- (V)AE models based on DNN and CNN trained on kinematics of up to 18 reconstructed physics objects per event
- Quantization-aware training and post-training quantization for reducing resources while maintaining accuracy
- Fully on-chip model implementation to stay within the L1 trigger latency

arXiv:2108.03986

Physics case - Example #2



arXiv:2108.03986

Physics case - Example #2

TABLE III. Resource utilization and latency for the quantized and pruned DNN and CNN (V)AE models. Resources are based on the Vivado estimates from Vivado HLS 2020.1 for a clock period of 5 ns on Xilinx VU9P.

Model	DSP [%]	LUT [%]	FF [%]	BRAM [%]	Latency [ns]	II [ns]
DNN AE QAT 8 bits	2	5	1	0.5	130	5
CNN AE QAT 4 bits	8	47	5	6	1480	895
DNN VAE PTQ 8 bits	1	3	0.5	0.3	80	5
CNN VAE PTQ 8 bits	10	12	4	2	365	115

arXiv:2108.03986

Physics case - Example #3

Low-latency RNN on FPGA for classification

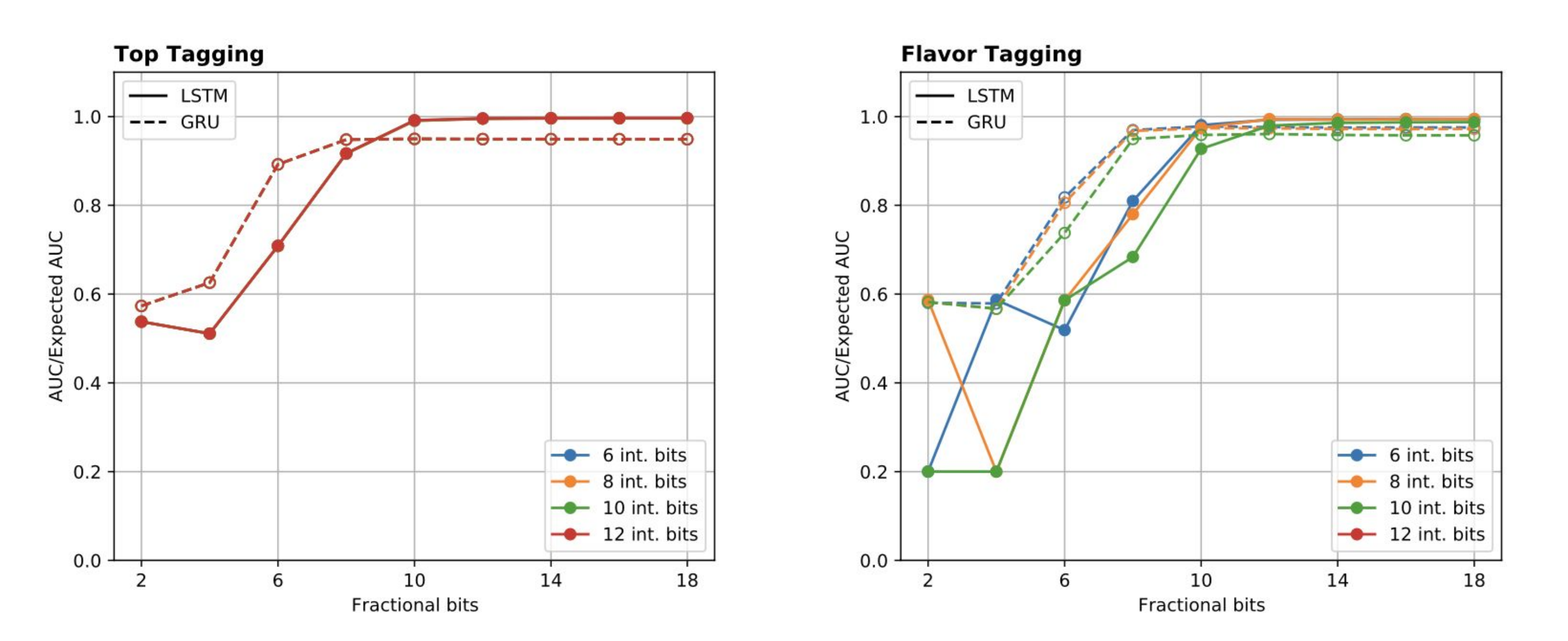
Table 1: Network hyperparameters and total number of trainable parameters for different benchmark models.

Benchmark	Sequence length	Input vector size	Hidden vector size	Dense layer sizes	output vector size	Trainable parameters		
						Non-RNN layers	LSTM	GRU
Top tagging	20	6	20	64	1	1,409	2,160	1,680
Flavor tagging	15	6	120	50/10	3	6,593	60,960	46,080
QuickDraw	100	3	128	256/128	5	66,565	67,584	51,072

- Addition of RNN support within hls4ml
- Successful simulation of FPGA deployment of RNN models with parameters from O(1k) to O(100k) and latencies from O(1 us) to O(100 us)

arXiv:2207.00559

Physics case - Example #3



Ratios of fixed-point and floating-point AUCs for top-tagging and flavour-tagging classification

arXiv:2207.00559

Physics case - More examples

Fast convolutional neural networks on FPGAs with hls4ml

Thea Aarrestad, Vladimir Loncar, Nicolò Ghielmetti, Maurizio Pierini, Sioni Summers, Jennifer Ngadiuba, Christoffer Petersson, Hampus Linander, Yutaro Iiyama, Giuseppe Di Guglielmo, Javier Duarte, Philip Harris, Dylan Rankin, Sergo Jindariani, Kevin Pedro, Nhan Tran, Mia Liu, Edward Kreinar, Zhenbin Wu, Duc Hoang

FPGA-accelerated machine learning inference as a service for particle physics computing

Javier Duarte · Philip Harris · Scott Hauck · Burt Holzman · Shih-Chieh Hsu · Sergo Jindariani · Suffian Khan · Benjamin Kreis · Brian Lee · Mia Liu · Vladimir Lončar · Jennifer Ngadiuba · Kevin Pedro · Brandon Perez · Maurizio Pierini · Dylan Rankin · Nhan Tran · Matthew Trahms · Aristeidis Tsaris · Colin Versteeg · Ted W. Way · Dustin Werran · Zhenbin Wu

Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP

Simone Francescato[2], Stefano Giagu[1], Federica Riti[3], Graziella Russo[1], Luigi Sabetta[1,a], Federico Tortonesi[1]

Conclusions

Smarter triggers are needed for the high-luminosity program at the LHC

AI is with us, and will remain

Low-latency is the key, and FPGA-based acceleration is an interesting area of active R&D

Various interesting projects and studies already in the literature, summarised a few here

A rigorous comparison of farm designs and the impact on physics is still lacking