Soft-error Sensitivity of Distributed Deep Learning

14th International SuperComputing Camp 2023

Esteban Meneses, PhD

Advanced Computing Laboratory
Costa Rica National High Technology Center
emeneses@cenat.ac.cr

Where do I come from?

Costa Rica, Central America

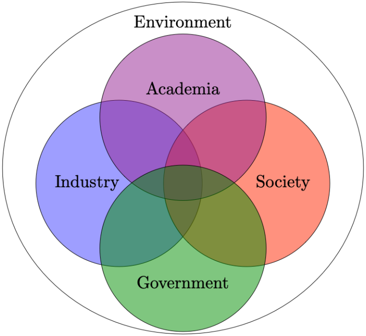
is not Puerto Rico
has no standing army since 1949
hosts 6% of world's biodiversity
produces 99% of its electricity from renewable sources

CeNAT

CeNAT-CONARE Campus, San José, Costa Rica

Costa Rica National High Technology Center

CeNAT



Development through Knowledge

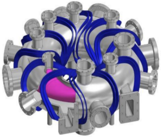
Advanced Computing Laboratory

Powered by a computing-centered diverse team

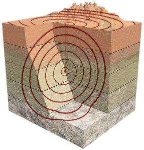
Collaborative Research Projects

Accelerating scientific discovery

Energy

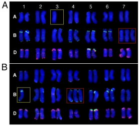


Seismology



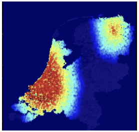
Biodiversity

Bioinformatics

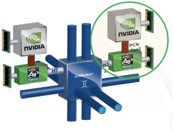


Climate

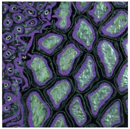
Epidemics

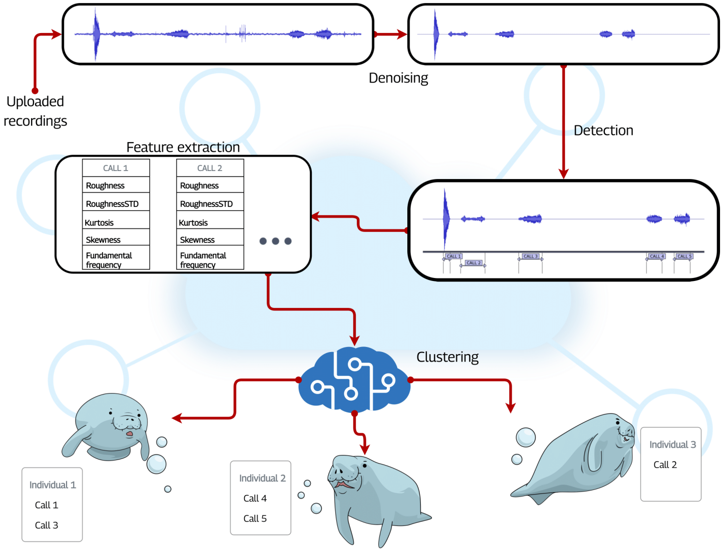


HPC

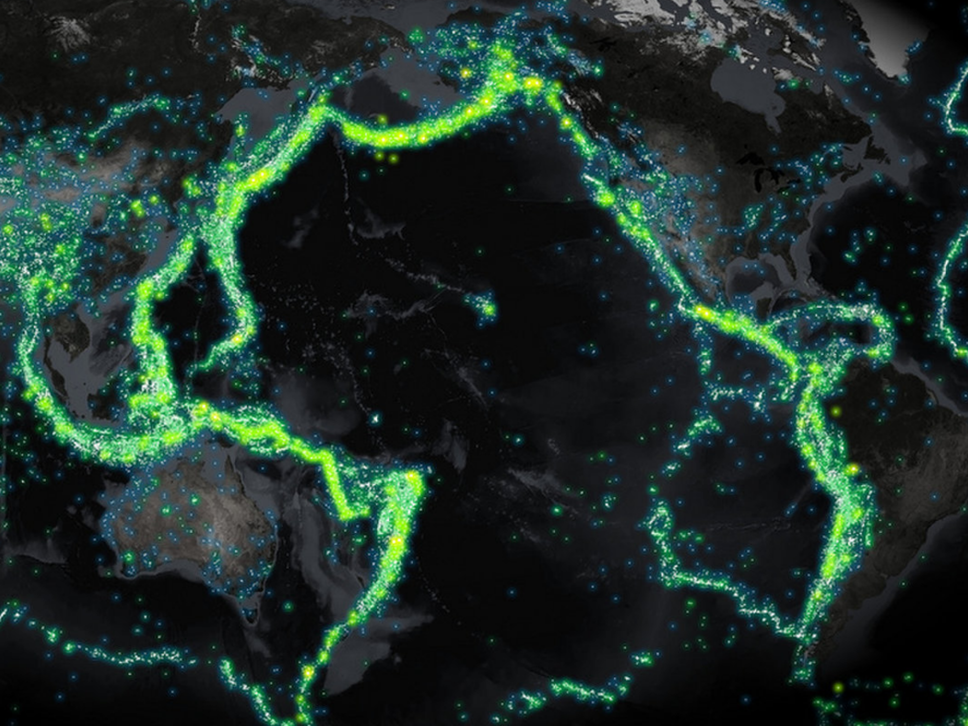


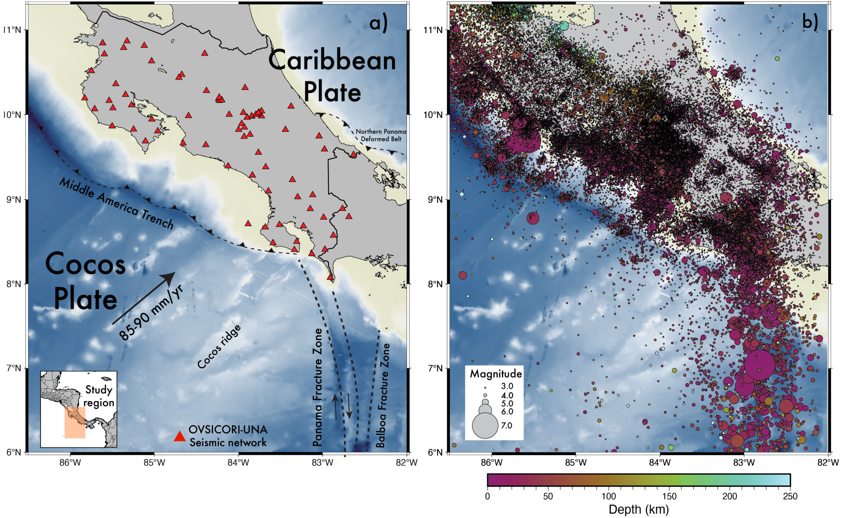
Health





Uploaded recordings
Denoising
Detection
Feature extraction
CALL 1
Roughness
RoughnessSTD
Kurtosis
Skewness
Fundamental frequency
CALL 2
Roughness
RoughnessSTD
Kurtosis
Skewness
Fundamental frequency
CALL 1
CALL 2
CALL 3
CALL 4
CALL 5
Clustering
Individual 1
Call 1
Call 3
Individual 2
Call 4
Call 5
Individual 3
Call 2




a)
Caribbean Plate
Northern Panama Deformed Belt
Middle America Trench
Cocos Plate
85-90 mm/yr
Cocos ridge
Panama Fracture Zone
Balboa Fracture Zone
Study region
OVSICORI-UNA Seismic network
11°N
10°N
9°N
8°N
7°N
6°N
86°W
85°W
84°W
83°W
82°W
b)
Magnitude
3.0
4.0
5.0
6.0
7.0
0
50
100
150
200
250
Depth (km)

Supercomputing Infrastructure

Simulation + Data Science + Artificial Intelligence + Bioinformatics

KABRÉ

SUPERCOMPUTER

Advanced Computing Training

Empowering collaborators

Programming

Scientific Computing

Machine Learning

System Tutorial

Programming

Data Visualization

Statistical Analysis

Bioinformatics

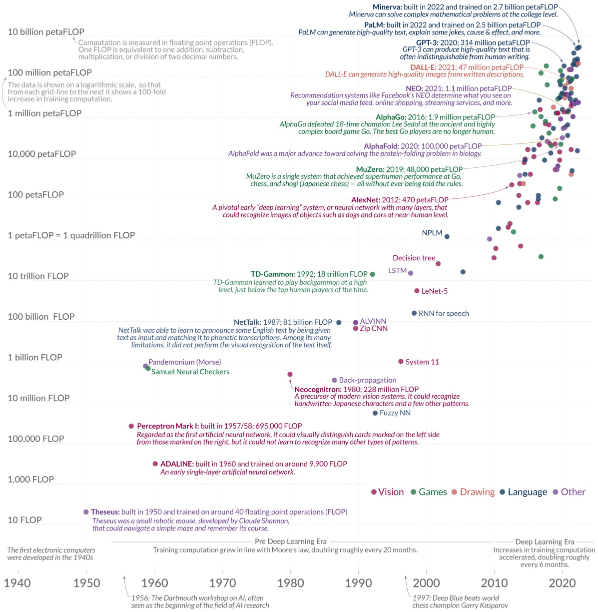
COSTA RICA Big Data SCHOOL

HPC SCHOOL COSTA RICA

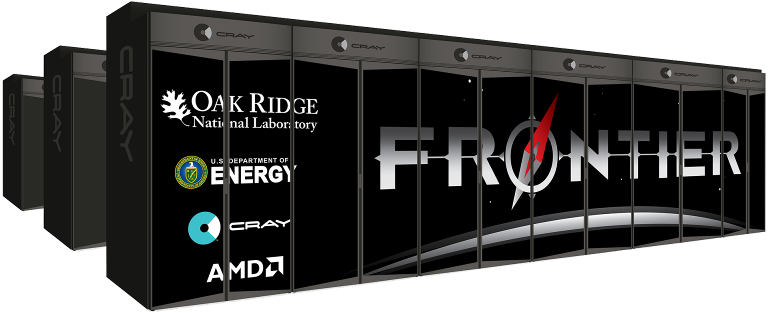
Costa Rica HPC School, January, 2023

Outline

August Kamp and DALL-E, from *Girl with a Pearl Earring* by Johannes Vermeer



Source: https://ourworldindata.org/



CRAY
OAK RIDGE
National Laboratory
U.S. DEPARTMENT OF
ENERGY
CRAY
AMD
FRONTIER

Component (system-wide)	Failure Rate (over one year)	Failure Location (on/off node)
compute unit (soft error)	53.95%	on node
card	14.47%	off node
cable	8.55%	off node
link module	6.58%	off node
process/daemon	5.26%	off node
coolant monitor	4.61%	off node
other	6.58%	N/A

Source: *Failure Analysis and Quantification for Contemporary and Future Supercomputers* by Tan and DeBardeleben

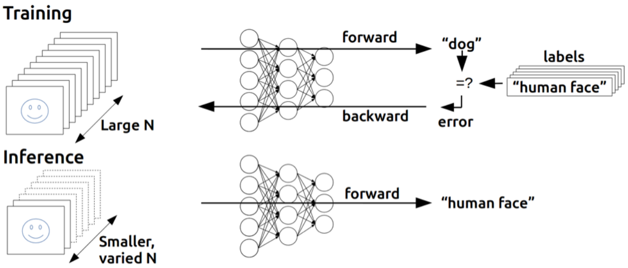
Soft errors are prevalent due to:

- Component count
- Feature size vulnerability
- Energy savings through sub-threshold voltage
- Cost of detection logic

The Astronomer, Johannes Vermeer, 1668

Deep Learning

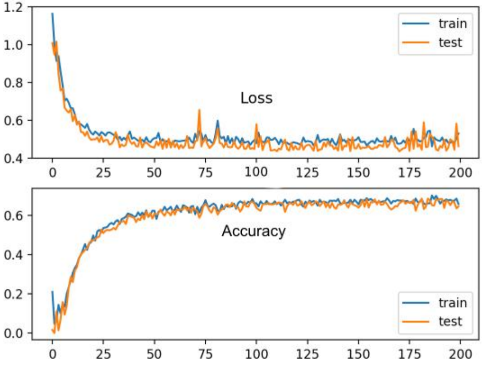
Artificial Neural Networks



Source: https://developer.nvidia.com/

Deep Learning Performance

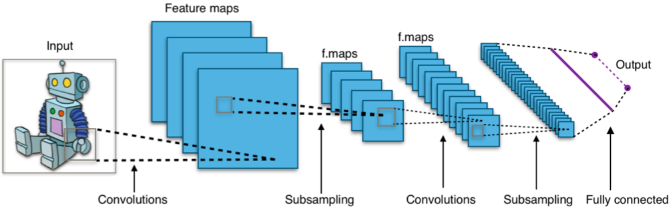
Accuracy and loss metrics



CeNAT

Deep Learning Architecture

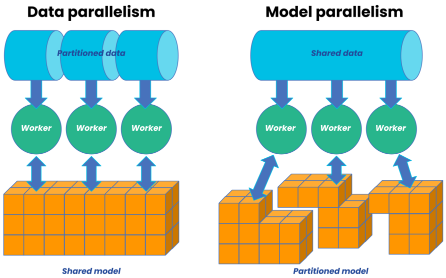
Many layers in specialized blocks



Typical structure of a convolutional neural network

Types of Distributed Deep Learning

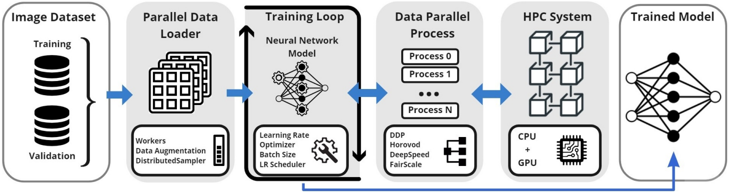
Data parallel or model parallel



Source: https://www.anyscale.com/

Distributed Deep Learning Training

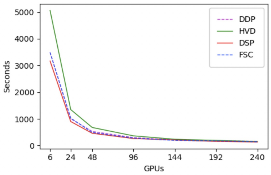
Mechanisms and tools



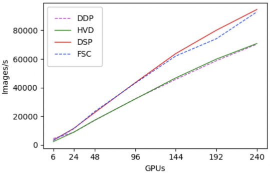
Elvis Rojas, Fabricio Quirós-Corella, Terry Jones, and Esteban Meneses. **Large-Scale Distributed Deep Learning: A Study of Mechanisms and Trade-Offs with PyTorch**. Latin America High Performance Computing Conference (CARLA), 2021, Guadalajara, Mexico.

Tradeoffs in Distributed Learning

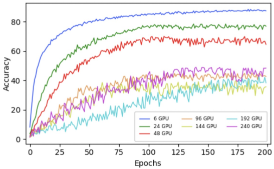
Faster execution and lower accuracy on Summit supercomputer



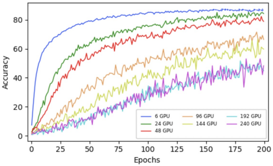
(a) Training time performance.



(b) Training throughput.



(a) DDP ResNet50.



(b) DDP ResNet101.

Hyperparameter Tuning

Crucial in bringing up performance

Accuracy of DT mechanisms and DL models scaling on GPUs

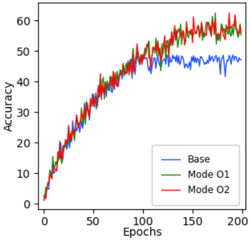
	R50	LR	R101	LR	V16	LR	V19	LR	R50	LR	R101	LR	V16	LR	V19	LR
GPU	**DDP**								**HVD**							
6	87/66	α	86/68	β	75/64	α	69/58	β	73/68	α	69/63	α	71/60	α	65/54	α
24	78/63	β	79/65	β	74/61	α	64/55	β	70/65	α	67/61	α	63/55	α	53/46	α
48	69/59	β	74/60	β	61/55	β	56/50	β	68/63	α	63/59	α	45/44	α	30/31	α
96	44/40	β	40/37	β	48/45	γ	52/45	γ	63/59	α	65/61	α	48/47	α	36/32	α
144	35/34	β	47/40	γ	46/45	γ	47/37	γ	61/56	α	68/61	α	42/40	α	40/36	α
192	38/37	β	55/50	γ	37/33	γ	27/26	γ	57/55	α	60/58	α	36/30	α	26/25	α
240	48/40	γ	48/40	γ	46/40	γ	35/27	γ	57/53	α	73/64	α	32/27	α	28/20	α
	DSP								**FSC**							
6	71/59	β	72/63	β	74/62	β	70/64	β	86/75	β	79/67	β	77/5 9	β	73/60	β
24	72/63	β	68/60	β	70/63	β	68/57	β	81/76	β	77/74	β	75/63	β	71/64	β
48	66/54	β	61/57	β	52/45	β	51/43	β	77/55	β	75/68	β	71/61	β	70/65	β
96	65/55	β	60/51	β	44/40	β	42/37	β	52/45	β	51/50	β	48/42	β	55/65	β
144	58/51	β	63/56	β	45/41	β	40/36	β	50/44	β	47/42	β	54/44	β	55/50	β
192	60/50	β	64/60	β	41/38	γ	39/37	γ	37/33	α	40/37	α	42/39	β	40/35	β
240	54/48	γ	52/45	γ	40/38	γ	33/31	γ	45/42	α	45/40	α	43/40	β	38/33	β

LR: $\alpha = 0.01$, $\beta = 0.001$, $\gamma = 0.0001$

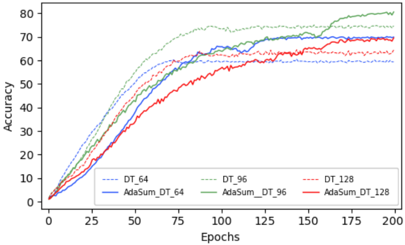
Parameters: learning rate, patience, batch size

Critical Optimizations

Significant impact on performance



Using mixed precision

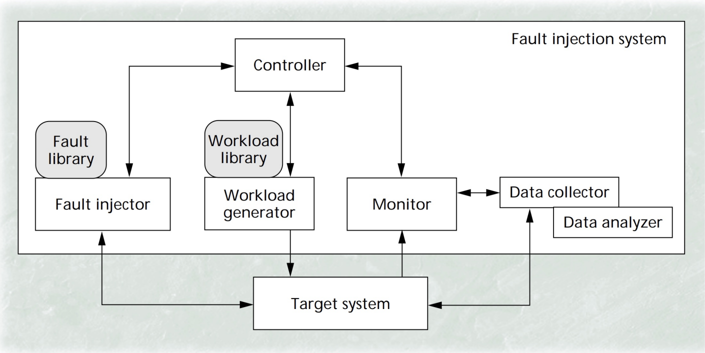


Using optimizer

The Milkmaid, Johannes Vermeer, 1657

Fault Injection

A first categorization

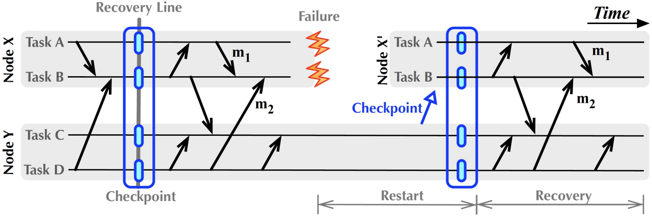


Source: *Fault Injection: Techniques and Tools* (Hsueh et al)

Types: dynamic (runtime) and static (compile time)

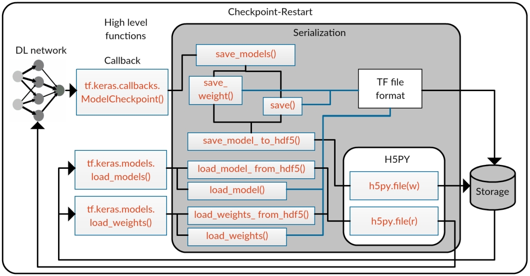
Checkpointing

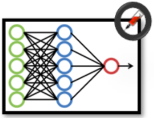
Dumping state of application



Checkpoint Alteration

Using checkpoints to inject failures





- Disadvantage: does not capture entire spectrum of faults
- Advantages: controllable and portable

Elvis Rojas, Diego Pérez, Jon C. Calhoun, Leonardo Bautista Gomez, Terry Jones, Esteban Meneses. **Understanding Soft Error Sensitivity of Deep Learning Models and Frameworks through Checkpoint Alteration**. IEEE International Conference on Cluster Computing (CLUSTER), 2021, Portland, Oregon, United States.

CeNAT

Failure Injector

Highly configurable

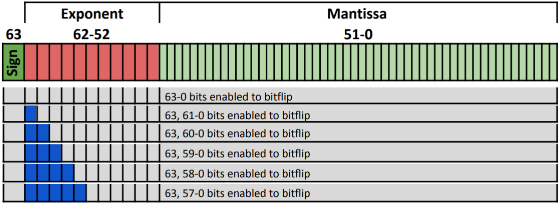
Setting	**Description**
hdf5_file	The path of the HDF5 file to corrupt
injection_probability	The probability that each injection attempt is successful
injection_type	Either “count” or “percentage”; the former allows an integer number of injection attempts, the latter allows a percentage of the file that can be corrupted
injection_attempts	The value for the *injection_type*
float_precision	16, 32, or 64 bit precision to use for each corruption of a floating-point number
corruption_mode	• *bit_mask*, A pattern of bits to flip (e.g., 101101), the first bit to apply the mask in each value is randomly selected from [0 to *float_precision* - length(*bit_mask*)], zeros are padded to both sides of the mask to match *float_precision*, then we XOR the mask against the floating-point value • *bit_range*, [first_bit, last_bit] a range of corruptible bits from 0 to *float_precision* -1 • *scaling_factor*, a scale factor to multiply each value
allow_NaN_values	If false, the corrupter does not transform a value to a NaN or INF
locations_to_corrupt	The list of locations to corrupt; all sublocations inside a location will be corrupted
use_random_locations	If true, it will ignore *locations_to_corrupt* and pick a random location every time

Works on HDF5 checkpoint files (Chainer, PyTorch, TensorFlow)

CeNAT

IEEE-754 Sensitivity

Impact of bit flip on floating-point numbers



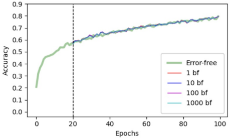
Incidence of NaN and extreme values (N-EV).

		Chainer						PyTorch						TensorFlow					
		ResNet50		VGG16		AlexNet		ResNet50		VGG16		AlexNet		ResNet50		VGG16		AlexNet	
Bit-flips	Trainings	N-EV	%	N-EV	%	N-EV	%	N-EV	%	N-EV	%	N-EV	%	N-EV	%	N-EV	%	N-EV	%
1	250	1	0.4	0	0	0	0	1	0.4	1	0.4	0	0	1	0.4	0	0	1	0.4
10	250	18	7.2	7	2.8	15	6	22	8.8	17	6.8	12	4.8	17	6.8	7	2.8	7	2.8
100	250	122	48.8	32	12.8	96	38.4	142	56.8	163	65.2	119	47.6	167	66.8	83	33.2	106	42.4
1000	250	249	99.6	188	75.2	241	96.4	249	99.6	248	99.2	249	99.6	246	98.4	227	90.8	234	93.6

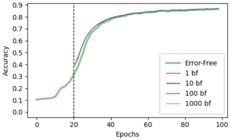
Critical bit: most significant bit of the exponent

Impact of Bitflips

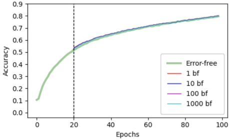
Deep neural networks are quite resilient



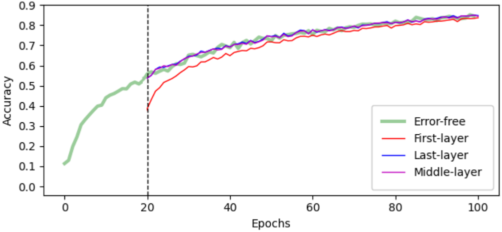
(a) Chainer with ResNet50.



(b) PyTorch with VGG16.



(c) TensorFlow with AlexNet.



Impact of Other Variables

Bit masks and floating-point precision

Multi-bit mask applied to DL framework training.

		Chainer		PyTorch		TensorFlow	
Bits	Mask	AvgI-Acc	N-EV	AvgI-Acc	N-EV	AvgI-Acc	N-EV
0	00000000	57.6		30.01		39.2	
3	10001010	57.3	1	29.9	1	36.8	0
4	01101010	57.1	3	29.9	0	36.6	0
4	10110010	57.4	0	29.1	1	36.7	1
5	11110001	53	0	27.2	0	36.5	3
6	11101101	57.4	1	29.9	2	36.8	3

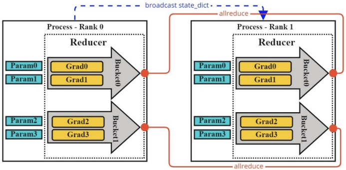
Incidence of NaN and extreme values in 16 bit and 32 bit precision.

		16 bits			32 bits		
Bit-flips	DL Train	ResNet(%)	VGG(%)	AlexNet(%)	ResNet(%)	VGG(%)	AlexNet(%)
1	250	0.4	0	0.4	1.2	2.4	2.8
10	250	10.4	11.6	7.2	15.6	17.2	13.2
100	250	59.2	69.2	60	76.8	72.4	68
1,000	250	96	77.2	86	98	78	91.6

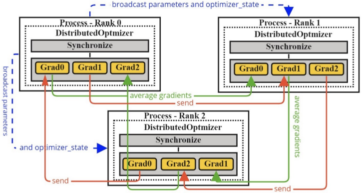
View of Delft, Johannes Vermeer, 1661

Distributed Deep Learning

Implementations



Distributed training process in DDP.



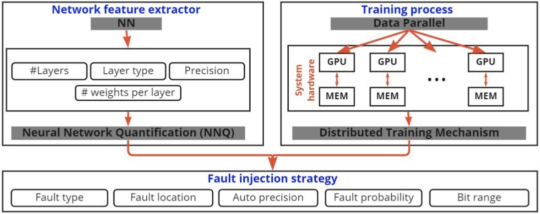
Distributed training process in HVD.

Elvis Rojas, Diego Pérez, Esteban Meneses. **Exploring the Effects of Silent Data Corruption in Distributed Deep Learning Training**. International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), 2022, Bordeaux, France.

CeNAT

Failure Injection

Adapted to a distributed environment



Failure injection strategy

Effects of a single bit-flip on distributed training

CNN	DT Mech.	Structure	#BF (Bit-Af)	# Processes	# Trainings	RWC	RWC %
ResNet	DDP	Optimizer	1 (Any)	1/8	180	85	47.2
		Model	1 (Any)	1/8	180	109	60.5
	HVD	Optimizer	1 (Any)	1/8	180	76	42.2
		Model	1 (Any)	1/8	180	103	57.2
VGG	DDP	Optimizer	1 (Any)	1/8	180	82	45.5
		Model	1 (Any)	1/8	180	101	56.1
	HVD	Optimizer	1 (Any)	1/8	180	80	44.4
		Model	1 (Any)	1/8	180	127	70.5

Injection Impact

Measured on model and optimizer

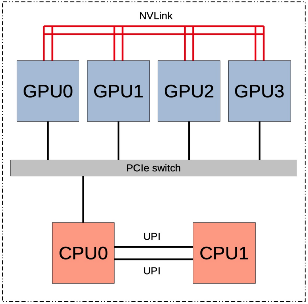
Effects of multiple bit-flips on distributed training

		Model							
		DDP				HVD			
		1	2	3	4	5	6	7	8
#BF (Bit-Af)	# Proc/TProc	ACC	STD	CV	NaN	ACC	STD	CV	NaN
1 (Any)	4/8	71.102	0.299	0.004	1	69.814	0.123	0.002	1
1 (Any)	8/8	71.073	0.391	0.006	1	69.871	0.099	0.001	0
10 (Any)	1/8	71.128	0.457	0.006	4	69.746	0.077	0.001	4
10 (Any)	4/8	71.202	0.321	0.005	8	69.813	0.139	0.002	3
10 (Any)	8/8	71.272	0.385	0.005	8	69.749	0.14	0.002	2
100 (Any)	1/8	70.992	0.453	0.006	16	69.774	0.223	0.003	16
100 (Any)	4/8	-	-	-	18	69.766	0.183	0.003	14
100 (Any)	8/8	-	-	-	18	69.771	0.104	0.001	12

		Optimizer											
		DDP						HVD					
		9	10	11	12	13	14	15	16	17	18	19	20
#BF (Bit-Af)	# Proc/TProc	ACC	STD	CV	NaN	AccR	TF	ACC	STD	CV	NaN	AccR	TF
1 (Any)	4/8	71.217	0.335	0.005	-	-	-	69.772	0.111	0.002	-	-	-
1 (Any)	8/8	71.044	0.586	0.008	-	-	-	69.768	0.089	0.001	-	-	-
10 (Any)	1/8	70.705	1.596	0.023	1	1	1	69.757	0.273	0.004	-	-	2
10 (Any)	4/8	70.596	1.913	0.027	2	-	-	69.768	0.152	0.002	-	-	-
10 (Any)	8/8	69.043	2.125	0.031	4	3	-	69.82	0.12	0.002	1	-	-
100 (Any)	1/8	66.776	5.596	0.084	5	5	3	69.048	1.256	0.018	5	2	3
100 (Any)	4/8	58.17	-	-	16	1	1	69.774	0.079	0.001	7	-	5
100 (Any)	8/8	-	-	-	17	-	1	68.728	1.792	0.026	6	3	3

Performance Analysis

Studying sensitivity to noise



Supercomputing node

(source: https://hpcf.umbc.edu/)

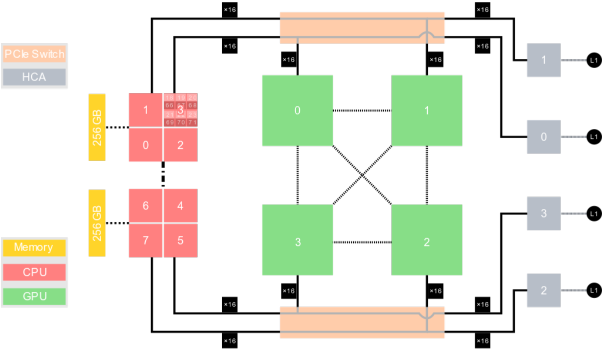
Juwels Supercomputer

Elvis Rojas, Michael Knobloch, Nour Daoud, Esteban Meneses, Bernd Mohr. **Early Experiences of Noise-Sensitivity Performance Analysis of a Distributed Deep Learning Framework**. HPC for International Collaboration between Europe and Latin America Workshop, IEEE Cluster, 2022, Heidelberg, Germany.

CeNAT

Performance Analysis Tools

Experimental Setup



Architecture of computational node in Juwels-Booster

Sensitivity and Coexistence

Experimental results

PERFORMANCE OF DL TRAINING (EXECUTION TIME IN SECONDS) IN SINGLE NODE SHARED AND EXCLUSIVE NUMA ENVIRONMENTS.

	Exclusive NUMA Domains		**Shared NUMA Domains**	
Repetition	No Noise	With Noise	No Noise	With Noise
MIN	134.22	134.83	143.31	190.37
Average time	135.77	138.20	145.62	256.27
Standard Dev.	1.377	2.97	1.45	39.26

SUMMARY OF THE PERFORMANCE OF DL TRAINING (EXECUTION TIME IN SECONDS) IN TWO NODE SHARED AND EXCLUSIVE NUMA ENVIRONMENTS.

	Exclusive NUMA Domain		**Shared NUMA Domain**	
Repetition	No Noise	With Noise	No Noise	With Noise
MIN	81.13	77.92	83.44	213.86
Average time	81.81	81.05	85.58	231.21
Standard Dev.	1.54	1.45	1.36	23.50

Acknowledgements

- Prof. Carlos J. Barrios for invitation
- Elvis Rojas, former doctoral student at Costa Rica Institute of Technology
- Diego Pérez, former master's student at Costa Rica Institute of Technology
- Leonardo Bautista at Barcelona Supercomputing Center
- Bernd Mohr at Jülich Supercomputing Center
- Jon Calhoun at Clemson University
- Terry Jones at Oak Ridge National Laboratory
- Members of the Advanced Computing Laboratory at the Costa Rica High Technology Center

Concluding Remarks

- Strong convergence of HPC, AI, and data science
- Checkpoint alteration is a controllable, portable and scalable fault injection approach
- Deep learning resilient to soft errors, albeit there is a critical bit
- Distributed deep learning can coexist with other complementary executions

Thank you!
emeneses@cenat.ac.cr

A supercomputer painted by Johannes Vermeer

(Daniel Amador and DALL-E)

CeNAT
Centro Nacional de Alta Tecnología