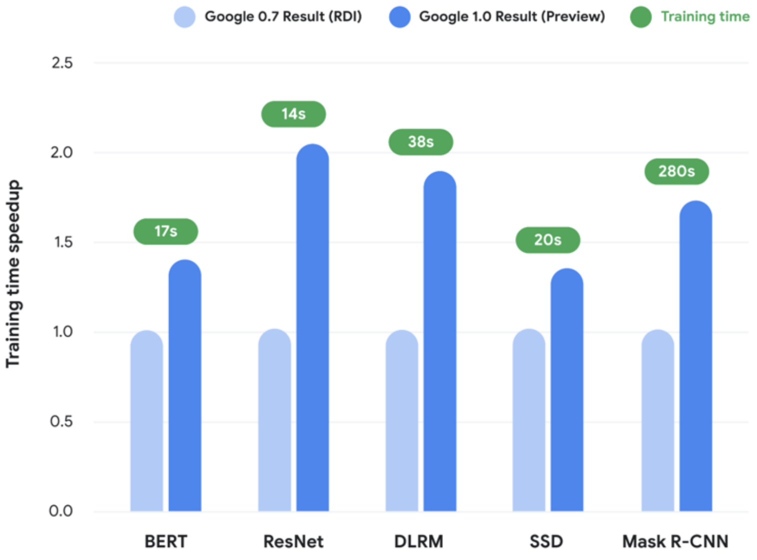
ElastiQuant: Elastic Quantization Strategy for Communication Efficient Distributed Machine Learning in IoT

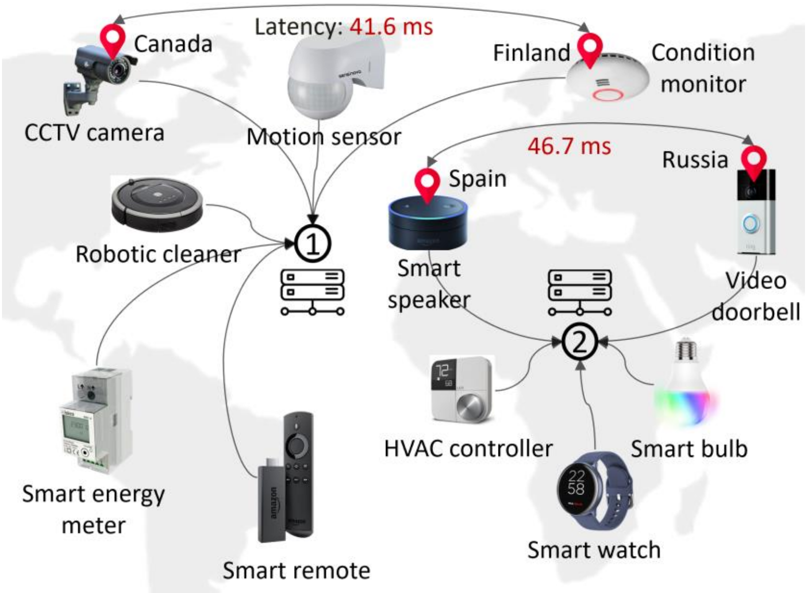
Bharath Sudharsan

Confirm Smart Manufacturing
A World Leading SFI Research Centre
Science Foundation Ireland For what's next
University of Limerick, Tyndall National Institute, UCC University College Cork, Ireland, NUI Maynooth, OÉ Gaillimh NUI Galway, Cork Institute of Technology, AIT, LIT

Overview



Training ML models on Google TPU Pod v4 - MLPerf



Training ML models on IoT devices - Globe2Train

- **Distributed ML in Data Centers/TPU Pods/GPU Clusters** - Train SOTA models faster, at greater scale, and at lower cost
- **Distributed ML in IoT** - On-device training with privacy preservation. Distribute neural workloads on 1000s of idle IoT devices

Related Work

Confirm
Smart Manufacturing

Table 1: ElastiQuant comparison with related papers. Not Applicable (NA), Not Investigated (NI), No Guarantee/assurance (NG).

Paper	Test Setup	Gradients; Momentum	Variance	Com bits	Scalability	Worst-case	Highlight
ATOMO [30]	AWS EC2 cloud	Atomic sparsification; NI	NG	Bounds	Within cluster nodes	NI	Sparsification to minimize variance
Terngrad [32]	128-node each with 4 Nvidia P100	Quantize to ternary levels; Yes	NI	NI	Within cluster nodes	NI	Need only three levels to aggressively reduce communication time
Globe2Train [20]	MCUs, CPUs	NA	NI	NI	Global IoT devices	NI	Latency, congestion tolerance
AD-PSGD [13]	32-node each with 4 Nvidia P100	NI	Bounds	NI	Within cluster nodes	NI	Robust to heterogeneity
QSGD [1]	AWS EC2 cloud	Lossy compression; NI	Bounds	Bounds	Within cluster nodes	NI	Good practical performance
NUQSGD [15]	8 NVIDIA 2080 Ti GPUs	Nonuniform quantization; Yes	Assurance, bounds	Bounds	Within cluster nodes	Yes	Stronger guarantees, higher empirical performance
D2 [27]	16 workers	NI	Assurance, bounds	NI	Within cluster nodes	NI	Much improve convergence rate, robust to data variance
EF-SignSGD [11]	Multiple workers	Error-feedback; Yes	NI	NI	Within cluster nodes	NI	Simply add EF to recover performance
DGC [14]	64-node each with 4 Nvidia Titan XP	Deep compression; Yes	NI	NI	Within cluster nodes, mobiles	NI	270 x - 600 x gradient compression ratio without losing accuracy
PowerSGD [29]	8-node each with 2 Nvidia Titan X	Low-rank compression; Yes	NI	NI	Within cluster nodes	NI	Consistent wall-clock speedups, test performance on par with SGD
ElastiQuant	18 IoT boards, edge GPUs	Elastic quantization; Yes	Assurance, bounds	Bounds	IoT boards, mobiles, edge GPUs	Yes	Higher solution quality, scalability - assurance with results

ElastiQuant Design

1. ElastiQuant Quantization Strategy
2. ElastiQuant Efficient Encoding
3. ElastiQuant Theoretical Assurance
   - Variance Bound
   - Bound on Number of Commn Bits
4. ElastiQuant with Momentum
   - For Convex Optimization
   - For Smooth Nonconvex Optimization
5. ElastiQuant for Decentralized Training
6. Worst-case Variance Analysis
   - Generally Spaced Levels
   - Exponentially Spaced Levels

ElastiQuant Design

Confirm
Smart Manufacturing

1

ElastiQuant Quantization Strategy

- Existing schemes that compress gradients does not take into consideration the properties of gradient vectors
  - ✓ Leads to slowing overall convergence as the gradient variance increase
  - ✓ To optimize overall performance, the communication savings should be balanced with variance
- ElastiQuant elastically distributes quantization levels in the unit interval
  - ✓ Uses a custom parameterized generalization to **control** communication cost and gradient variance
  - ✓ Reduces quantization error and variance as it can **match** the properties of gradient vectors
  - ✓ It increases the number of quantization levels near zero to obtain a **stronger** variance bound

ElastiQuant Design

Confirm
Smart Manufacturing

2

ElastiQuant Efficient Encoding

3

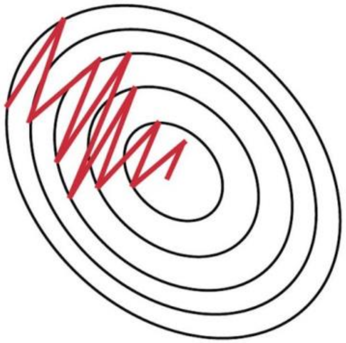
ElastiQuant Theoretical Assurance

- To provide tighter bounds on the code-length:
  - ✓ ***Encode*** function encodes quantized gradients before transmission
  - ✓ Uses b bits floating point
  - ✓ b set to 4 produces 4-bit-ElastiQuant
  - ✓ In rounds, the ***Decode*** function reads b bits, uses ERC^{-1} to reconstruct gradients

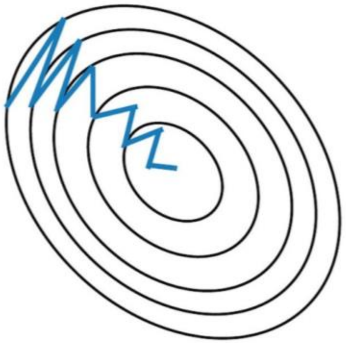
- Bound on Number of Communication Bits:
  - ✓ Variance bound + code-length bound = total commn costs bound
  - ✓ Suboptimal commn **savings** as compared to other schemes
- ElastiQuant Variance Bound - models with high variance fail to generalize
  - ✓ ElastiQuant bound is **tighter** than related schemes
- Both bounds demonstrated by training ResNets on CIFAR & ImageNet

ElastiQuant Design

Confirm
Smart Manufacturing



Stochastic Gradient Descent **withhout** Momentum



Stochastic Gradient Descent **with** Momentum

- Momentum helps accelerate gradients
  - ✓ Consistently in the right directions
  - ✓ Also dampens oscillations
  - ✓ Leads to faster convergence

4

ElastiQuant
with
Momentum

- Momentum is added to ElastiQuant training algorithm
- For nonconvex optimization, there can exist multiple locally optimal points - requires extra computation
- For convex optimization, there can be only one optimal solution
- We establish convergence assurance for momentum ElastiQuant for both optimizations:
  - ✓ ElastiQuant with Momentum for Convex Optimization
  - ✓ ElastiQuant with Momentum for Smooth Nonconvex Optimization

ElastiQuant Design

Confirm Smart Manufacturing

5

ElastiQuant for Decentralized Training

- For gradient communication, low latency and high bandwidth cannot be guaranteed to all devices
  - ✓ ElastiQuant can **integrate** with communication-efficient variants of decentralized parallel SGD
  - ✓ Provides a solution for distributed training of deep networks in **constrained** networked systems

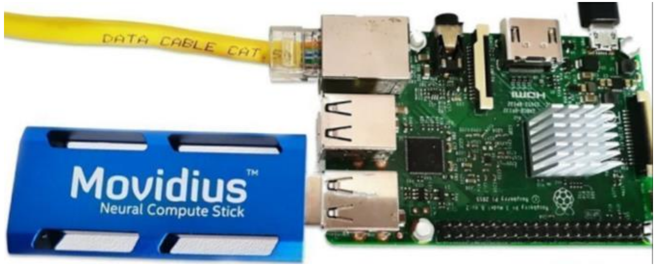
6

Worst-case Variance Analysis

- Incorporated into any solution selection process for a robust system design
- Performed to gain insights on the behavior of the variance upper bound
  - ✓ An upper bound for **tight worst-case** variance is established
  - ✓ Analysis shows ElastiQuant is **nearly optimal** in the worst-case
- Extend elastic quantization to arbitrary sequence of levels
  - ✓ Generally Spaced Levels
  - ✓ Exponentially Spaced Levels

Experimental Evaluation

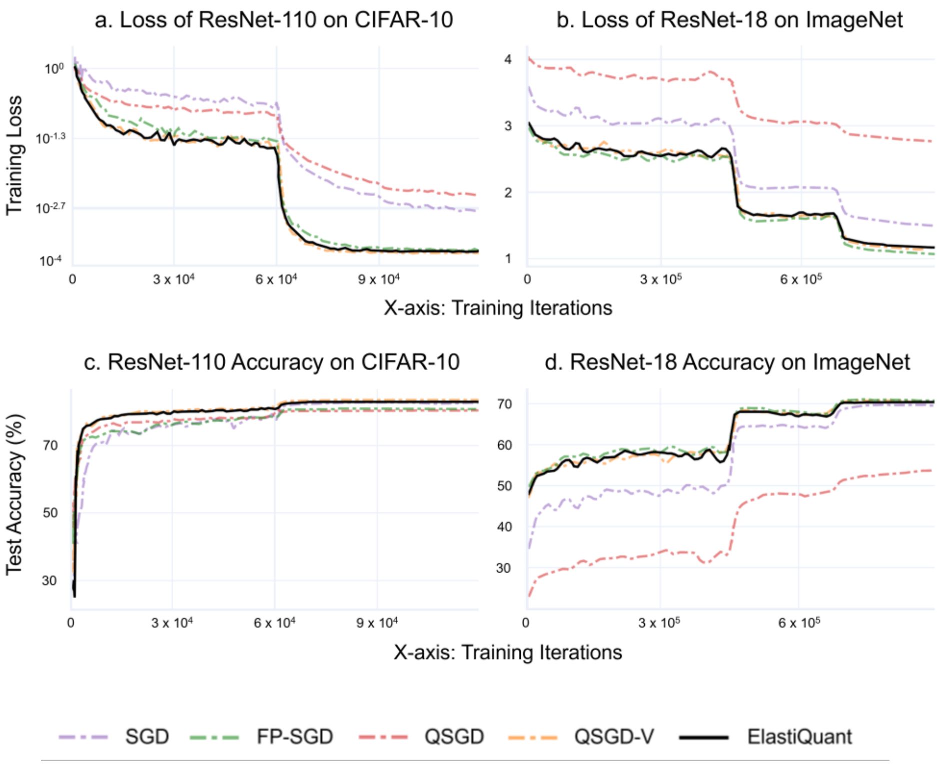


- **Setup.** 18 development boards wirelessly set up to replicate real-world heterogeneous IoT:
  - ✓ 7 Jetson Xaviers
  - ✓ 4 Jetson Nanos inserted on a Jetson Mate carrier board
  - ✓ 3 accelerated Google Coral boards
  - ✓ 4 Intel Movidius NCS accelerated Raspberry Pi
- **Data.** Portions of ImageNet, CIFAR-10, CIFAR-100 datasets are supplied to these boards for distributed training
- **Network.** ResNet family - ResNet-18, ResNet-20, ResNet-34, ResNet-50, Resnet-110 are trained
- **Implementation.** ElastiQuant in TensorFlow

Solution Quality

Confirm Smart Manufacturing



- **Training Loss -** Distributed training on 18 GPUs
  - ✓ ImageNet - ElastiQuant, QSGD-V has lower loss compared to QSGD
  - ✓ CIFAR10 - Significant gap in training loss which grows as training progresses
- **Test Accuracy**
  - ✓ Unlike ElastiQuant, QSGD does not achieve the accuracy of FP-SGD
  - ✓ For both datasets, ElastiQuant and QSGD-V outperform QSGD
  - ✓ The gap between ElastiQuant and QSGD is significant on ImageNet
  - ✓ ElastiQuant and QSGD-V show lower variance in comparison to QSGD

Scalability and Speedup Behavior

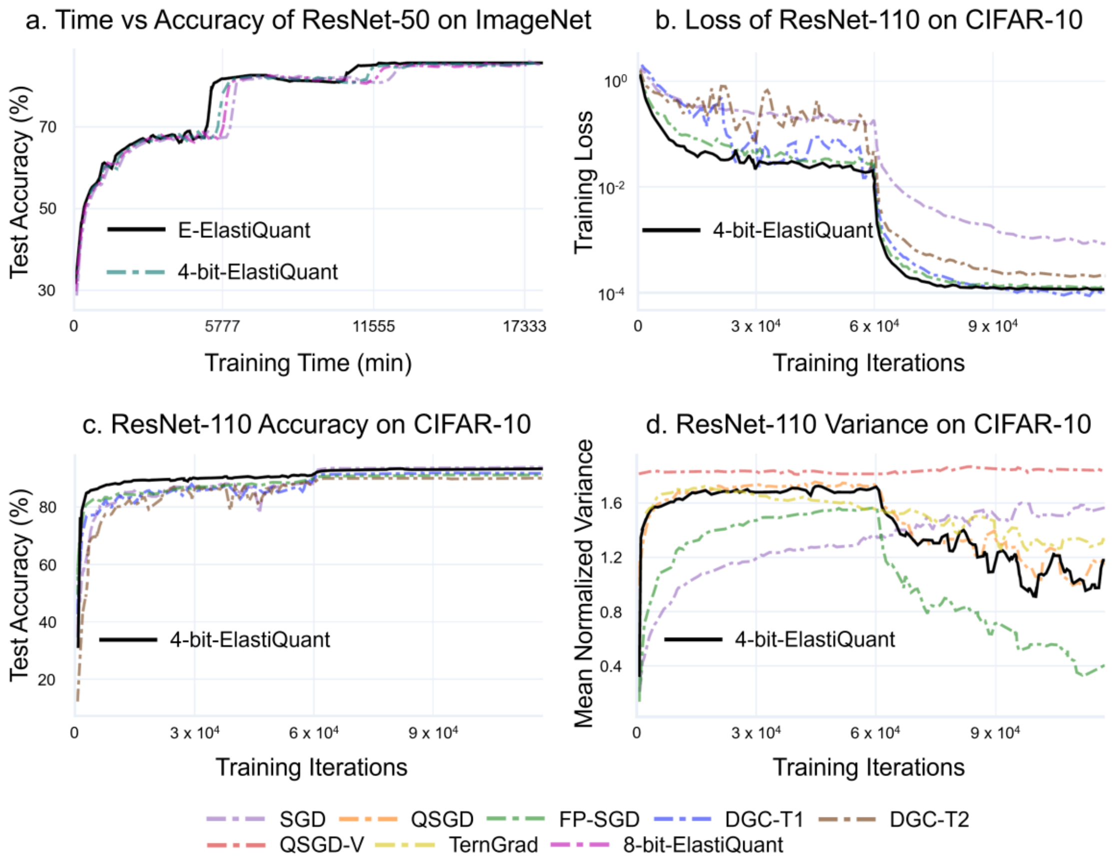
Table 3: Evaluating scalability performance of ElastiQuant by distributed training on 2 to 7 devices, and comparison with SGD: Calculating speedup (*Sp*) and total time (*T*) per epoch in minutes - sum of computation (*Cp*), encoding (*En*), transmission (*Tx*).

Network, Dataset	Scheme	2 Edge GPUs				4 Edge GPUs					7 Edge GPUs				
		Cp	*En*	*Tx*	*T*	*Cp*	*En*	*Tx*	*T*	*Sp*	*Cp*	*En*	*Tx*	*T*	*Sp*
ResNet-34, ImageNet	SGD	16.06	NA	17.93	33.99	10.47	NA	20.62	31.09	-2.9 ↑	15.34	NA	26.84	42.18	11.09 ↓
	8-bit-ElastiQuant	14.61	4.15	16.16	34.92	7.98	4.04	11.92	23.94	-10.98 ↑	10.05	4.04	10.26	25.35	1.41 ↓
	4-bit-ElastiQuant	15.75	0.93	15.55	32.23	11.19	1.25	9.32	21.76	-10.47 ↑	9.64	1.34	7.47	18.45	-3.31 ↑
	E-ElastiQuant	14.72	1.34	15.24	31.3	10.57	1.55	8.29	20.41	-10.89 ↑	8.91	1.35	6.01	16.27	-4.14 ↑
ResNet-50, ImageNet	SGD	177.23	NA	222.53	399.76	167.9	NA	265.17	433.07	33.31 ↓	85.28	NA	465.06	550.34	117.27 ↓
	8-bit-ElastiQuant	179.21	38.65	190.55	409.09	145.25	39.97	145.25	330.49	-78.62 ↑	99.94	38.64	183.89	322.47	-8.02 ↑
	4-bit-ElastiQuant	179.89	9.33	183.89	373.11	142.58	8.12	118.59	269.17	-103.94 ↑	110.6	6.66	126.59	243.85	-25.32 ↑
	E-ElastiQuant	169.23	21.32	171.9	362.45	126.59	18.66	103.93	249.18	-113.27 ↑	119.93	19.99	78.62	218.54	-30.64 ↑

- SGD **negative** scalability - *T* increases from 31.09 to 42.18 min for ResNet-34, and 433.07 to 550.34 min for ResNet-50
- 4-bit-ElastiQuant attains **positive** scalability - 1.48 times speedup for ResNet-34 when GPUs scaled from 2 to 4
- 8-bit-ElastiQuant faces a scalability **stall** when GPUs scaled from 4 to 7 - due to elevated encoding & communication costs
- E-ElastiQuant shows **top-of-the-class** scalability and communication compression

Results Comparison

Confirm Smart Manufacturing



- Time vs accuracy results:
    - ✓ All ElastiQuant variants match the non-quantized ResNet-50 accuracy, with speedups over SGD (baseline)
    - ✓ QSGD-V not plotted as its performance overlaps 4-bit-ElastiQuant
- EF-SignSGD needed intensive tuning to make it converge plus to bring up its accuracy
    - ✓ Tuning makes EF-SignSGD send additional scaling data - reduced efficiency
- E-ElastiQuant offers competitive performance
    - ✓ Provides convergence assurances
    - ✓ Additional comn bandwidth savings from its gradients encoding feature

Results Comparison

Table 4: Accuracy comparison of ElastiQuant trained ResNets with models trained using DGC, compression ratio (CR) tuned DGC, Atomo, TernGrad, others. FP-SGD is baseline.

Network, Dataset	Scheme	Tune	Edge GPUs	Test Accuracy (%)
ResNet-20, CIFAR-10	Atomo	Default	2	87.6
	TernGrad			No convergence
	FP-SGD			90.5
	4-bit-ElastiQuant			86.3
ResNet-18, CIFAR-10	DGC	0.02 CR	2	91.25
			6	88.87
		0.12 CR	2	90.08
			6	87.36
	FP-SGD	Default	6	92.72
	4-bit-ElastiQuant		6	91.96
ResNet-18, CIFAR-100	DGC	0.02 CR	2	74.41
			6	72.69
	FP-SGD	Default	6	74.33
	4-bit-ElastiQuant			73.63
ResNet-110, CIFAR-10	SGD	Default	6	89.76
	QSGD			89.22
	QSGD-V			90.10
	FP-SGD			92.03
	TernGrad			91.33
	4-bit-ElastiQuant			90.80

- ElastiQuant vs Deep Gradient Compression (DGC)
  - ✓ ResNet-18 on CIFAR-10, CIFAR-100 - unlike ElastiQuant, DGC **accuracy degrades** when GPUs scaled from 2 to 4
  - ✓ So even if ElastiQuant could save lesser commn bandwidth than DGC, ElastiQuant is **practical** due to its better scalability
- ElastiQuant vs ATOMO and TernGrad
  - ✓ For ResNet-20, although ATOMO shows slightly higher accuracy than ElastiQuant, ATOMO has **high train time** plus GPU strain
  - ✓ TernGrad convergence was **under par** for standard parameters - tuning to bring performance close to ATOMO & ElastiQuant
  - ✓ For ResNet-110, TernGrad shows the closest performance to FPSGD and slightly **outperforms** 4-bit-ElastiQuant

Conclusion

- ElastiQuant improves communication efficiency during distributed learning in IoT. It consistently demonstrated:
  - ✓ Improved solution quality as the resultant ResNet models achieved **lower loss and better accuracy**
  - ✓ Higher **training scalability and speedup** due to reduced communication volume
  - ✓ Reduced quantization induced **variance** due to its elastic quantization approach
  - ✓ On-the-fly **communication efficiency** as ElastiQuant can re-use parameters of full-precision schemes with slight tuning

Confirm
Smart Manufacturing

Confirm
Smart Manufacturing

Contact: Bharath Sudharsan
Email: bharath.sudharsan@insight-centre.org

www.confirm.ie

Science Foundation Ireland
For what's next