Towards a Better 16-Bit Number Representation for Training Neural Networks

Himeshi De Silva, Hongshi Tan, Nhut-Minh Ho, John L. Gustafson, Weng-Fai Wong

Scientist I, I2R

A*STAR

02nd March 2023

CREATING GROWTH, ENHANCING LIVES

Outline

Mixed-precision Training

FP16

Comparison Results

Accuracy Analysis

Discussion and Conclusion

CREATING GROWTH, ENHANCING LIVES

Mixed-precision Training

Institute for
Infocomm Research

I²R

Training Neural Networks

- Reducing precision/optimizations in inference has received a lot of attention
- Training complex networks with large datasets can take time
- Bandwidth and Memory limitations
    - Reduced precision
    - Requires representations to efficiently reduce numerical error

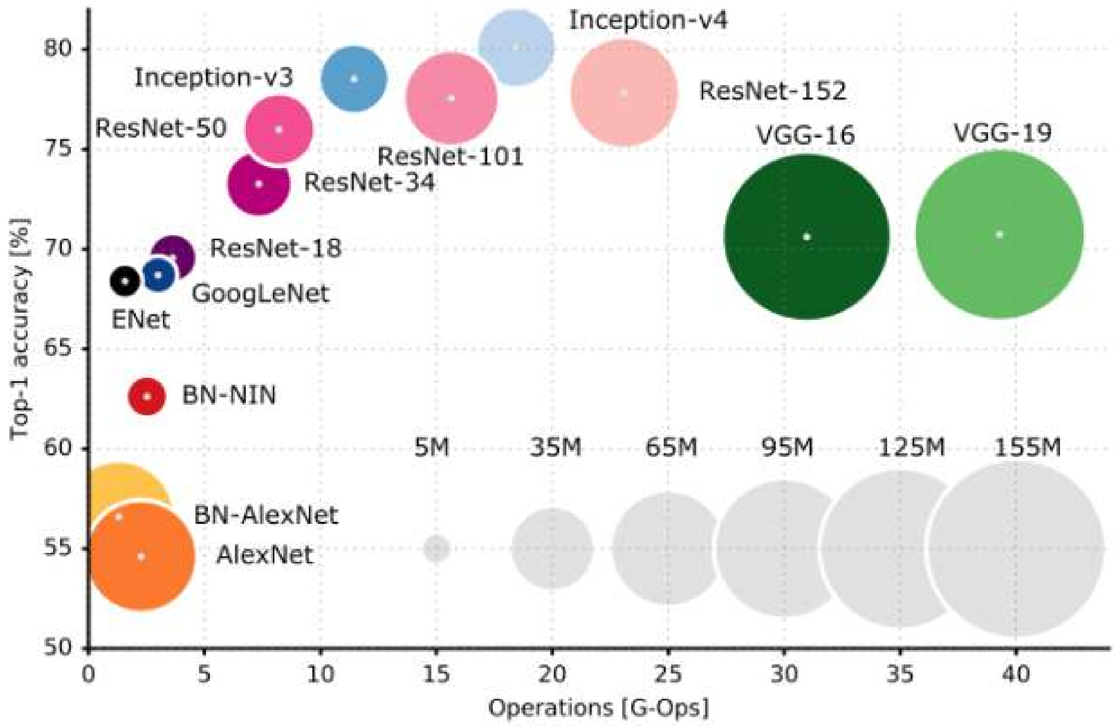


Image Source: https://www.topbots.com/a-brief-history-of-neural-network-architectures/

41

Mixed-precision training

- 16-bit floating-point formats cannot work on their own
- Unlike inference, training has higher dynamic range and precision requirements
- Additional techniques
    - Scaling, master copy of weights, fused operations, maintaining some computations at IEEE32

- FP8 training – complex scaling
- With each format using different techniques, how do they compare against each other under *identical conditions*?

- IEEE16 (half-precision), bfloat16, DLfloat, Posit

FP16

bfloat16

- Developed by Google
- Follows IEEE 754 rules, *w* – 8 bits, *t* – 7 bits, *emax* – 127, *bias* - 127

$$v = \begin{cases} \text{NaN} & \text{if } E = 255, T \neq 0 \\ (-1)^S \times (+\infty) & \text{if } E = 255, T = 0 \\ (-1)^S \times 2^{-126} \times (2^{-7} \times T) & \text{if } E = 0, T \neq 0 \\ 0 & \text{if } E = 0, T = 0 \\ (-1)^S \times 2^{E-127} \times (1 + 2^{-7} \times T) & \text{otherwise} \end{cases}$$

0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0

= 0.15625

Sign
(+ve)

Exponent
(2^{-3})

Significand
$1 + 2^5/2^7$

DLFloat

- Developed by IBM

$$v = \begin{cases} \infty/\mathrm{NaN} & \text{if } E = 63, T = 512 \\ 0 & \text{if } E = 0, T = 0 \\ (-1)^S \times 2^{E-31} \times (1 + 2^{-9} \times T) & \text{otherwise} \end{cases}$$

- 6-bit exponent field
- No sub-normal values, one representation for NaN and infinity

0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0

= 0.15625

Sign (+ve)

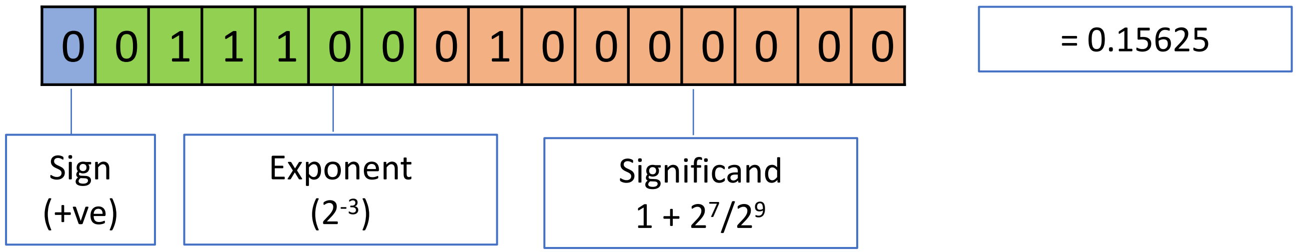
Exponent (2^{-3})

Significand $1 + 2^7/2^9$

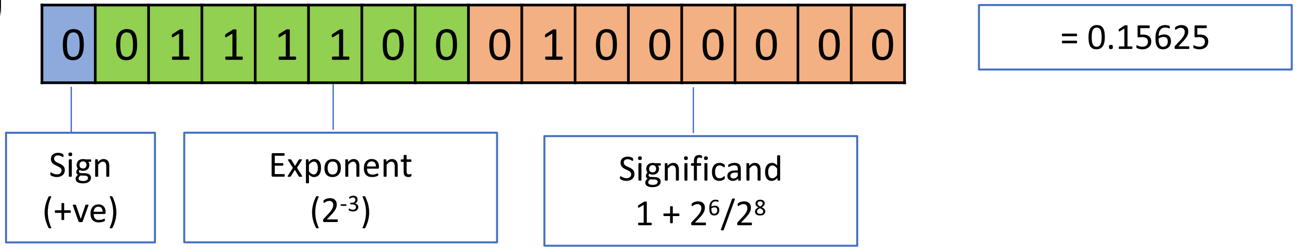
IEEE16_6 and IEEE16_7

- IEEE 754 compliant 16-bit floating-point representations with 6 and 7 bits of exponent

$$v = \begin{cases} \text{NaN} & \text{if } E = 63, T \neq 0 \\ (-1)^S \times (+\infty) & \text{if } E = 63, T = 0 \\ (-1)^S \times 2^{-30} \times (2^{-9} \times T) & \text{if } E = 0, T \neq 0 \\ 0 & \text{if } E = 0, T = 0 \\ (-1)^S \times 2^{E-31} \times (1 + 2^{-9} \times T) & \text{in all other cases} \end{cases}$$



$$v = \begin{cases} \text{NaN} & \text{if } E = 127, T \neq 0 \\ (-1)^S \times (+\infty) & \text{if } E = 127, T = 0 \\ (-1)^S \times 2^{-62} \times (2^{-8} \times T) & \text{if } E = 0, T \neq 0 \\ 0 & \text{if } E = 0, T = 0 \\ (-1)^S \times 2^{E-63} \times (1 + 2^{-8} \times T) & \text{in all other cases} \end{cases}$$



Posits

- Posits
- *nbits* (16) and *es* set the environment
  (standard defines *es* = 2)
- Sign S, regime R, exponent E, fraction F
- *maxpos* - largest real value expressible as a posit,*minpos* - smallest nonzero value expressible as a posit

$$v = \begin{cases} 0 & \text{if } p = 00....0 \\ \text{NaR} & \text{if } p = 10....0 \\ (1 - 3S + F) \times 2^{(-1)^S (R \times 2^{es} + E + S)} & \text{otherwise} \end{cases}$$

sign	regime				exponent			fraction							
0	0	0	0	1	1	0	1	1	1	0	1	1	1	0	1

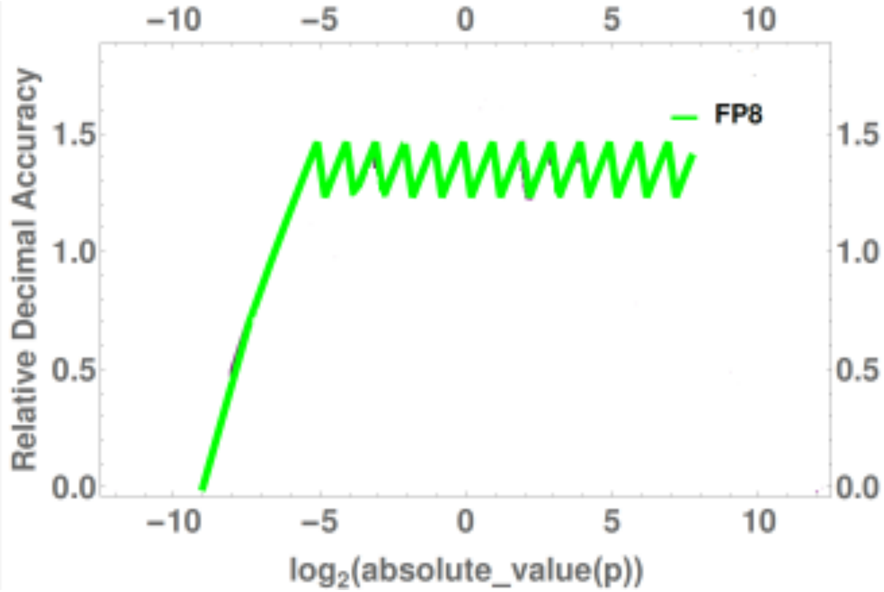
$+ \quad 256^{-3} \quad \times \quad 2^5 \times (1 + \quad 221/256 \quad)$

3.55393 × 10−6

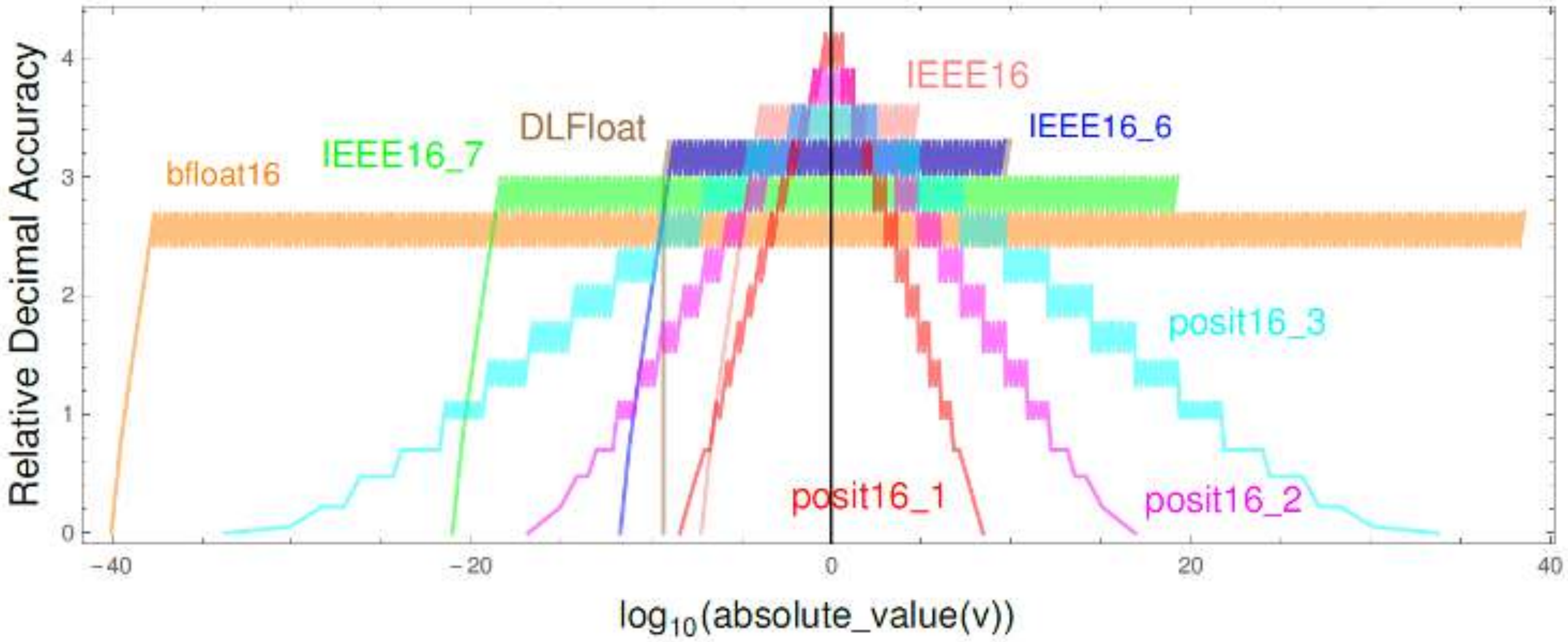
Relative Decimal Accuracy

- Relative Decimal Accuracy (RDA) between an exact value x and its approximated value $\hat{x}$,

$$RDA(x, \hat{x}) = log10\ (|x|/|x - \hat{x}|)$$

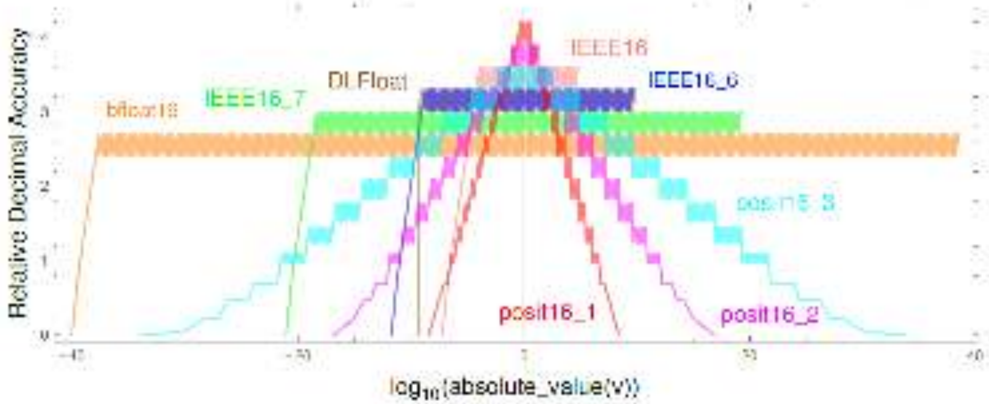


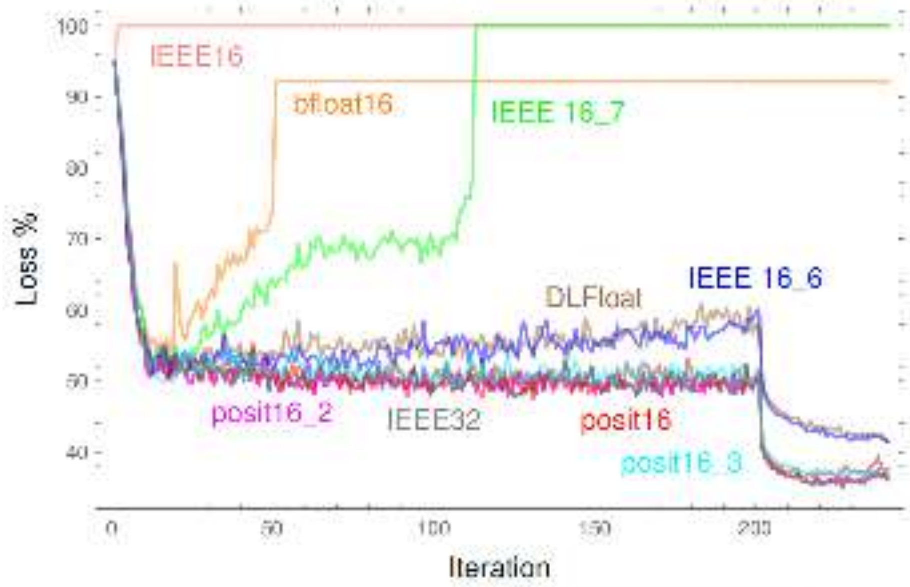
Relative Decimal Accuracy



Optimal Format?

Format	Min. Exponent	Max. Exponent	Precision	Min. Value	Max. Value
IEEE32	-149 (-126)	127	23 bits	1.47e-45 (1.75e-38)	3.48e+38
bfloat16	-133 (-126)	127	7 bits	9.18e-41 (1.75e-38)	3.38e+38
DLFloat	-31	32	9 bits	2.33e-10	8.58e+9
IEEE16	-24 (-14)	15	10 bits	5.96e-8 (6.10e-5)	6.55e+4
IEEE16_6	-39 (30)	31	9 bits	1.82e-12 (9.32e-10)	4.29e+10
IEEE16_7	-70 (-62)	63	8 bits	8.47e-22 (2.17e-19)	1.84e19
posit16	-28	28	0-12 bits	3.73e-9	2.68e+8
posit16_2	-56	56	0-11 bits	1.38e-17	7.20e+16
posit16_3	-112	112	0-10 bits	1.93e-34	5.19e+33





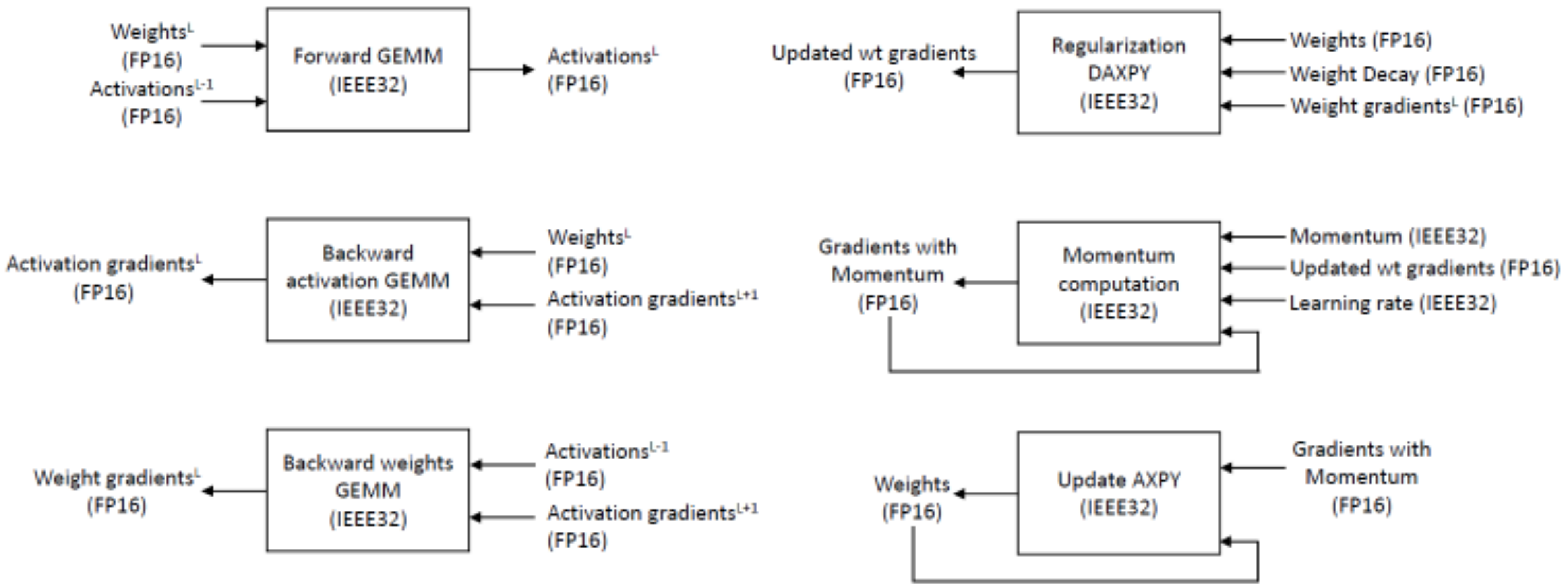
CREATING GROWTH, ENHANCING LIVES

Comparison Results

Institute for
Infocomm Research
I²R

Evaluation

- Using Caffe and Pytorch



Comparison Results

Model	LeNet	LeNet	convnet	NIN	Squeeze-Net	AlexNet	ResNet 18	Trans. Base.	Trans. Base.
Dataset	MNIST	FMNIST	CIFAR10	CIFAR10	ImageNet	ImageNet	ImageNet	30K	IWSLT14
IEEE32	98.70%	89.10%	78.70%	56.06%	56.40%	57.04%	67.88%	35.42	23.54
bfloat16	98.24%	89.08%	76.02%	0.96%	0.32%	52.40%	61.88%	35.18	21.68
DLfloat	98.66%	89.38%	77.96%	45.48%	54.24%	46.56%	69.12%	35.49	9.43
IEEE16	98.70%	89.22%	73.02%	NaN	0.00%	53.08%	NaN	0	Error
IEEE16_6	98.72%	89.60%	78.56%	46.28%	54.72%	46.84%	68.00%	35.59	12.6
IEEE16_7	98.46%	89.54%	78.74%	NaN	0.24%	9.96%	67.52%	35.16	9.46
posit16_1	98.72%	89.38%	9.76%	54.92%	50.80%	50.68%	0.00%	34.31	9.45
posit16_2	98.78%	89.36%	77.74%	53.92%	56.80%	53.60%	67.64%	35.18	24.97
posit16_3	98.66%	89.30%	79.72%	53.74%	56.48%	53.16%	67.60%	35.06	24.32

Observations

- posit16_1's dynamic range is smaller than all the other formats except for IEEE16
- Formats with greater dynamic range such as bfloat16 do not perform as well
- IEEE16 shows better performance than other float types in the case of AlexNet/Imagenet

Performance on Hardware

- Xilinx U250 Alveo Data Center Accelerator Card with synthesis done in Vivado 2018.2
- Conversion to and from IEEE32 for FP16
- Measure LUT, LUT memory, Registers, Depth

Performance on Hardware contd.

FP16	Conversion	Depth	LUT	LUTMem	Registers
bfloat16	F	1	0	0	0
	B	1	3	0	0
DLFloat	F	2	27	0	18
	B	3	63	0	43
IEEE16	F	18	223	1	469
	B	12	325	0	278
IEEE16_6	F	18	217	1	462
	B	12	331	0	277
IEEE16_7	F	18	216	1	457
	B	12	341	0	278
posit16_1	F	3	115	0	84
	B	5	703	0	318
posit16_2	F	3	112	0	83
	B	5	723	0	316
posit16_3	F	3	118	0	84
	B	5	596	0	314

*F – IEEE32 to FP16
B – FP16 to IEEE32

Accuracy Analysis

Institute for
Infocomm Research

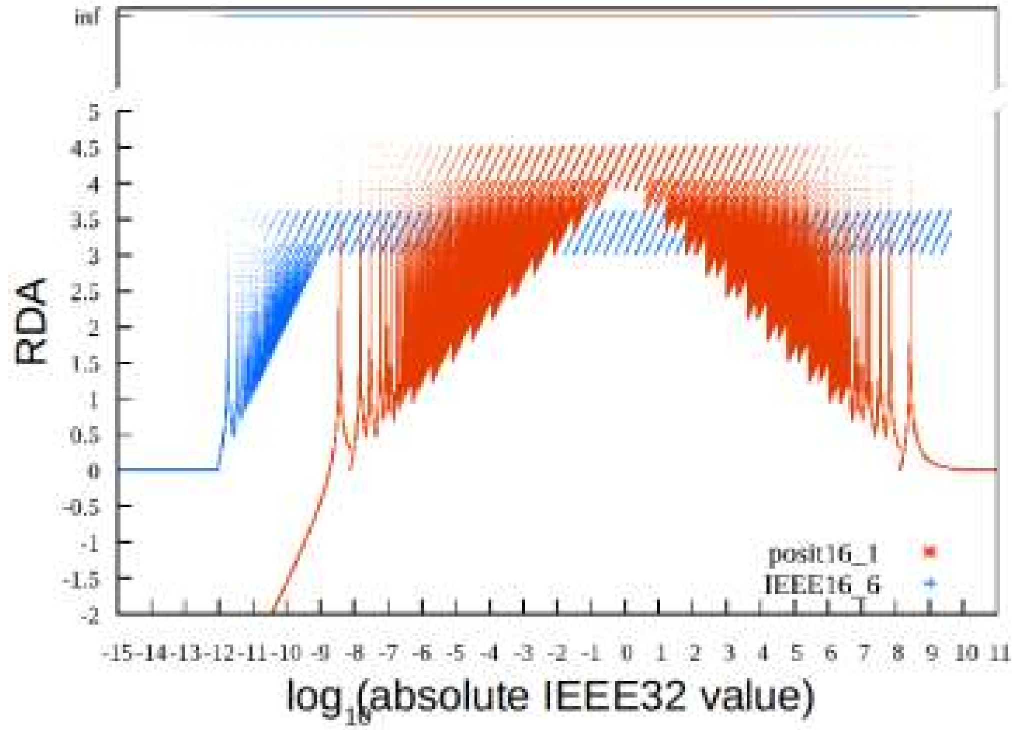
I²R

Analyzing Accuracy Behavior

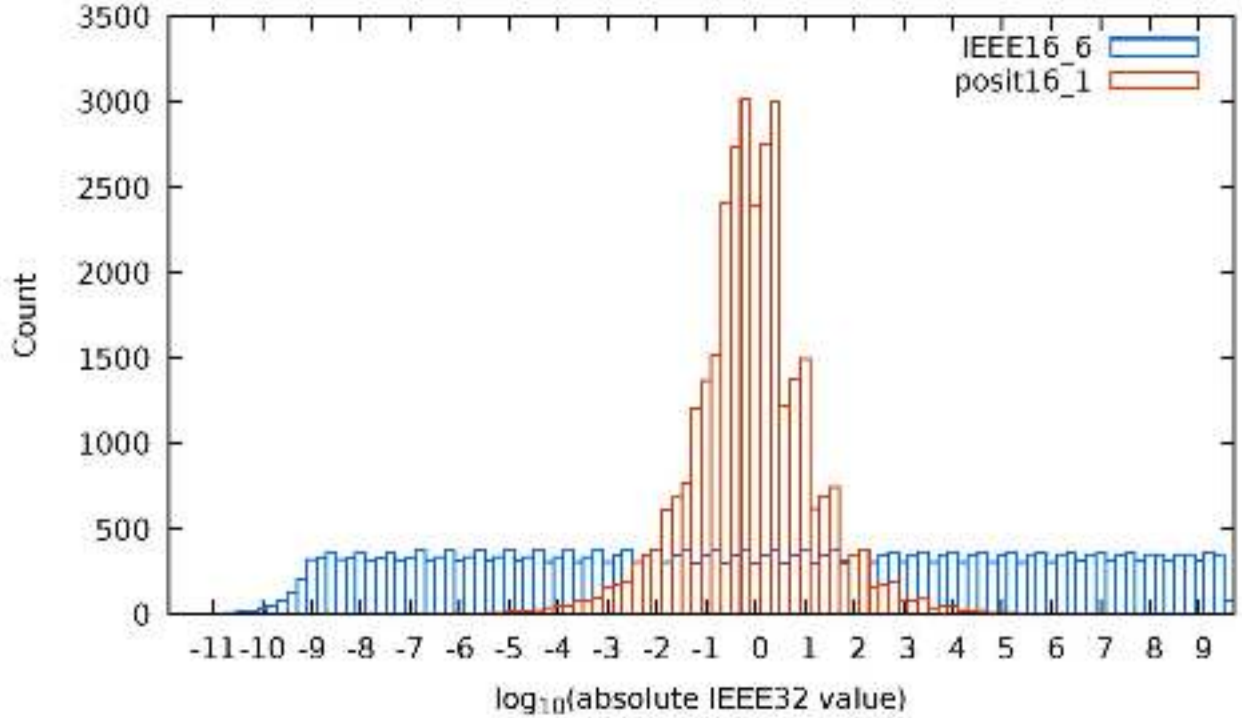
- Understand and explain the behavior of number representations in training
- NIN/CIFAR100
- 120K iterations
- posit16_1 vs IEEE16_6

Model	NIN
Dataset	**CIFAR100**
IEEE32	56.06%
DLFloat	45.48%
IEEE16_6	46.28%
posit16_1	54.92%

Accuracy differences between posit16_1 and IEEE16_6



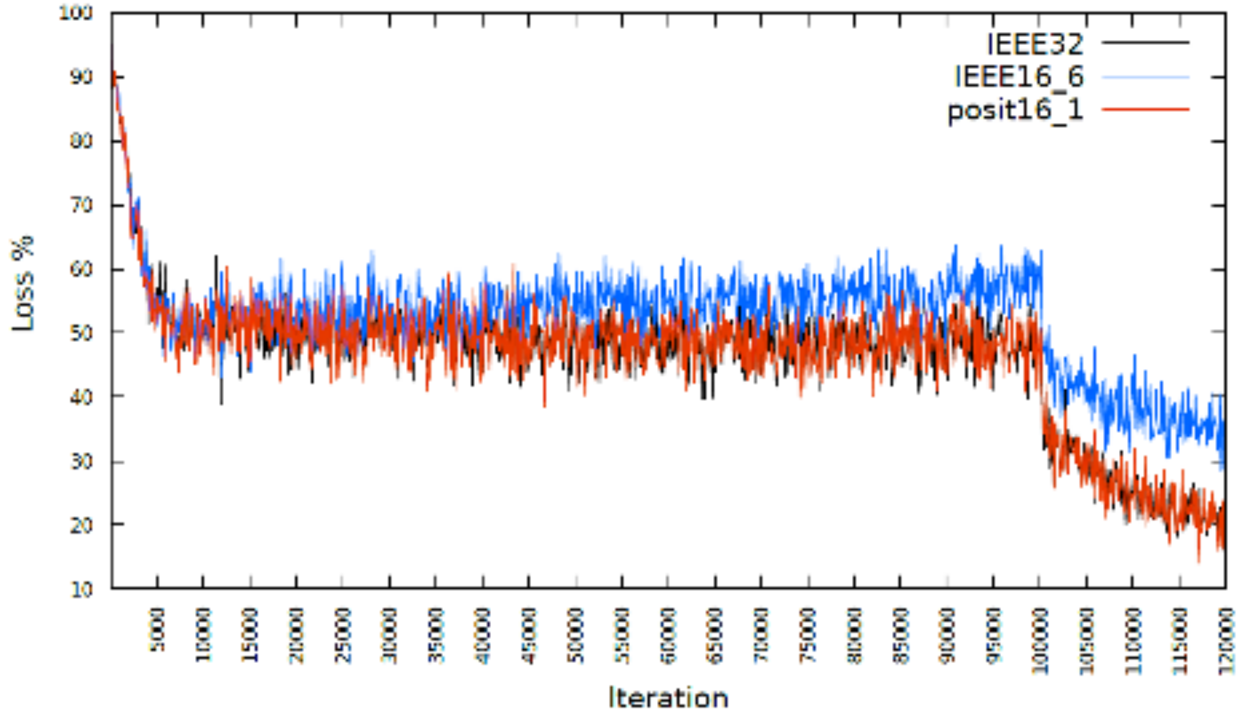
RDA w.r.t IEEE32 of IEEE16_6 vs posit16_1



Distribution of Infinitely Accurate Numbers of IEEE16_6 vs posit16_1

Loss behavior

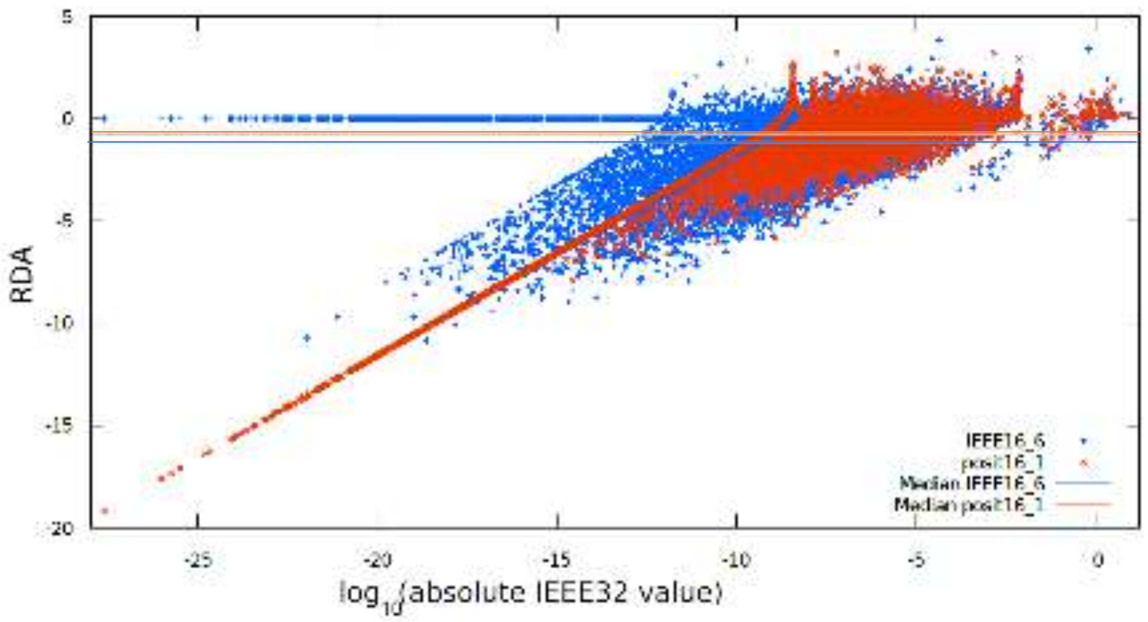
- Begin at one end of the network



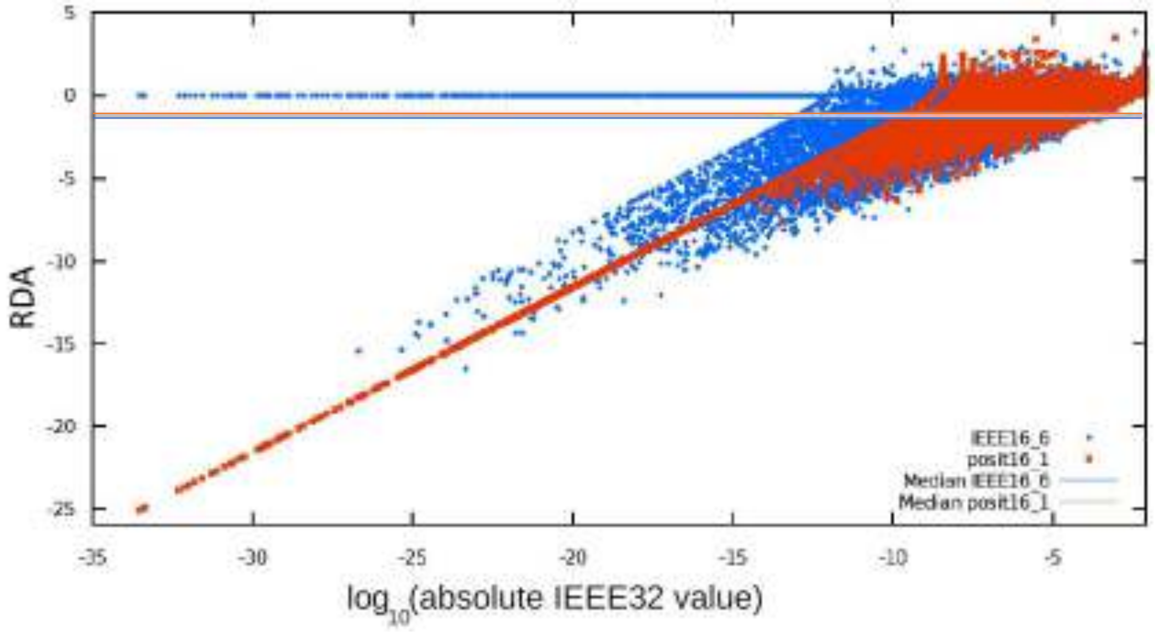
NIN/CIFAR100 Training Loss

Loss behavior contd.

- End of training



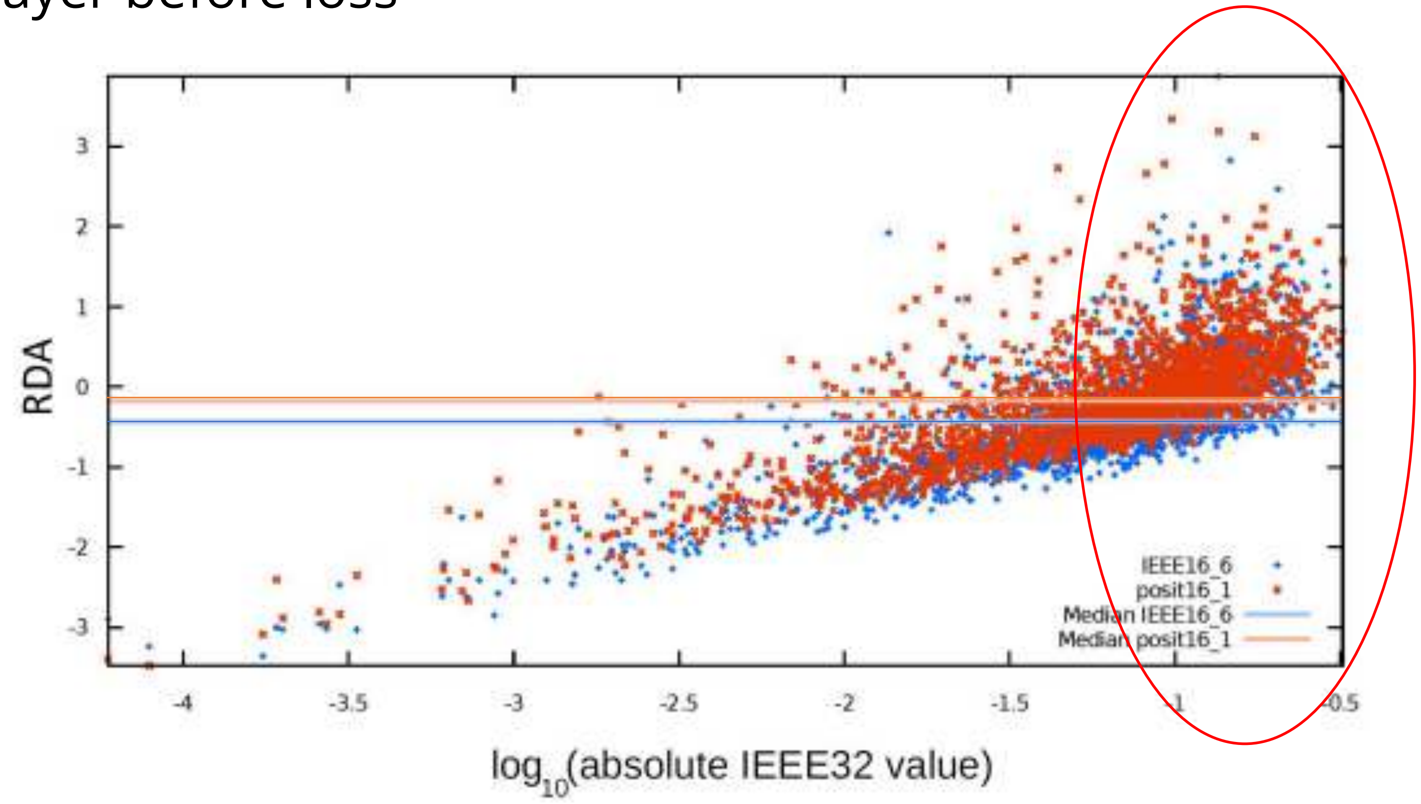
Accuracy of Loss Layer Activations



Accuracy of Loss Layer Gradients

Effect on Weights

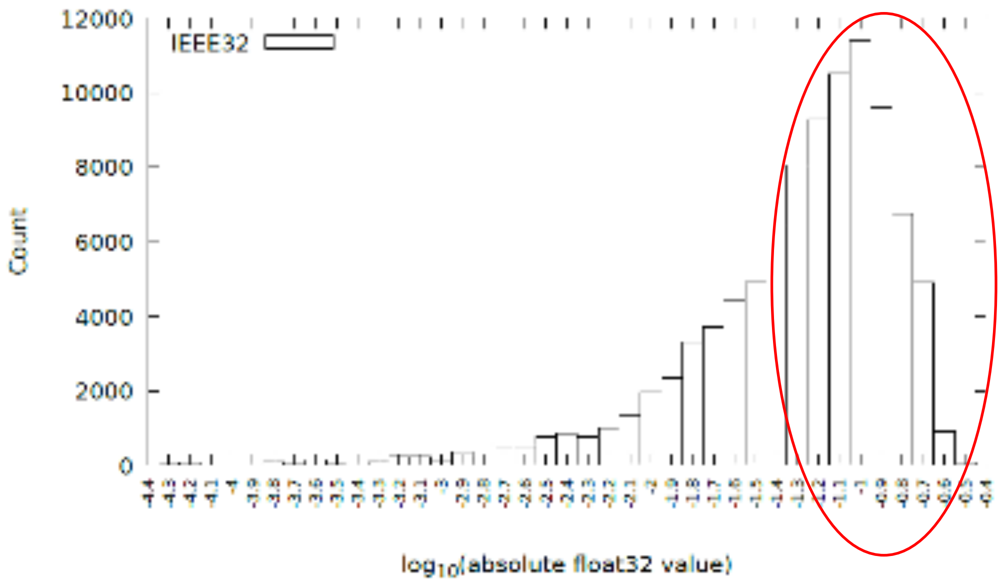
- Weight layer before loss



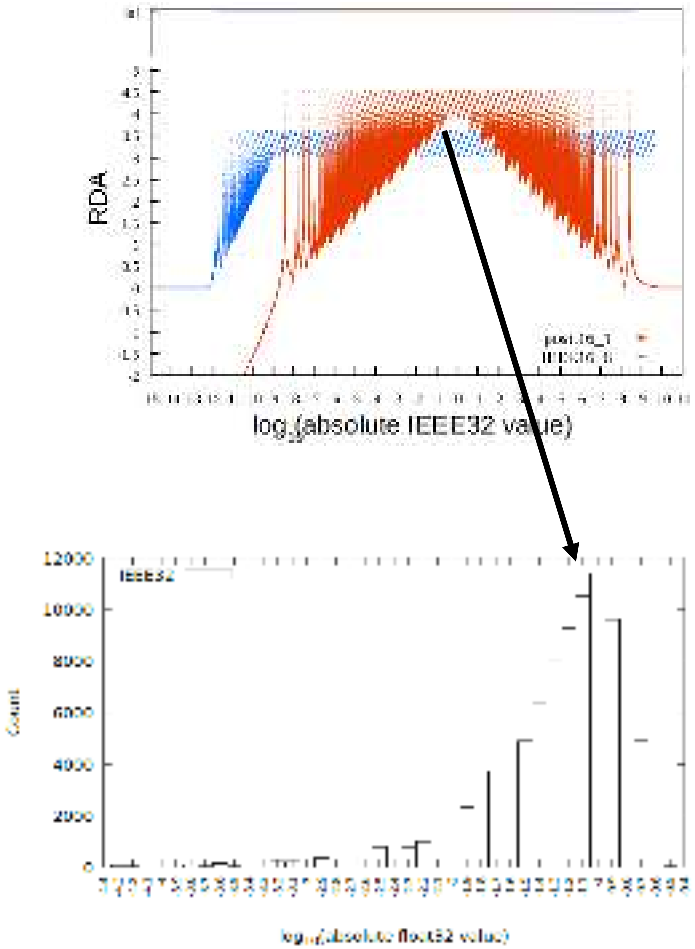
Accuracy of Weights

Posit Accuracy in Training

- Accuracy of learned weights has a significant impact on the training process
- Higher posit accuracy for weights transcends to other values such as gradients
- The weight values for which posits achieve superior accuracy is larger in magnitude
- Range of the weight values stabilizes early in the training
- This results in improved overall training accuracy for posits
- *Larger weight values which also occur more frequently inside the optimal accuracy range of posits, contributes to posits' superior accuracy result of this benchmark.*



Shifting the Accuracy Peak



- The unique accuracy distribution of posits allow us to customize the accuracy for a distribution ***without*** requiring more bits
- Shift to achieve scaling factor a power of 2
- Achieve IEEE32 performance

Discussion and Conclusion

Institute for
Infocomm Research
I²R

Discussion and Conclusion

- Traditional FP16 formats studied so far for CNN training all have uniform accuracy distributions and differ mostly in their bit configuration
- The IEEE 754 standard 16-bit format is inferior for out-of-the-box training of neural networks compared to the other float types
- Non-uniform accuracy formats such as posits provide broader versatility for neural network training
- Analyzing the dynamic range and precision as they relate to the distribution of the weights is a useful indicator for selecting the FP16 format to use
- Shifting the accuracy peak of posits leads to better training results

CREATING GROWTH, ENHANCING LIVES

Institute for
Infocomm Research

I²R

THANK YOU

www.a-star.edu.sg