NOTE:
The following slides are for reference by HotChips 2023 registered attendees only.

IBM NorthPole Neural Inference Machine

Dr. Dharmendra S. Modha
dmodha@us.ibm.com
IBM Research – Almaden, San Jose, CA

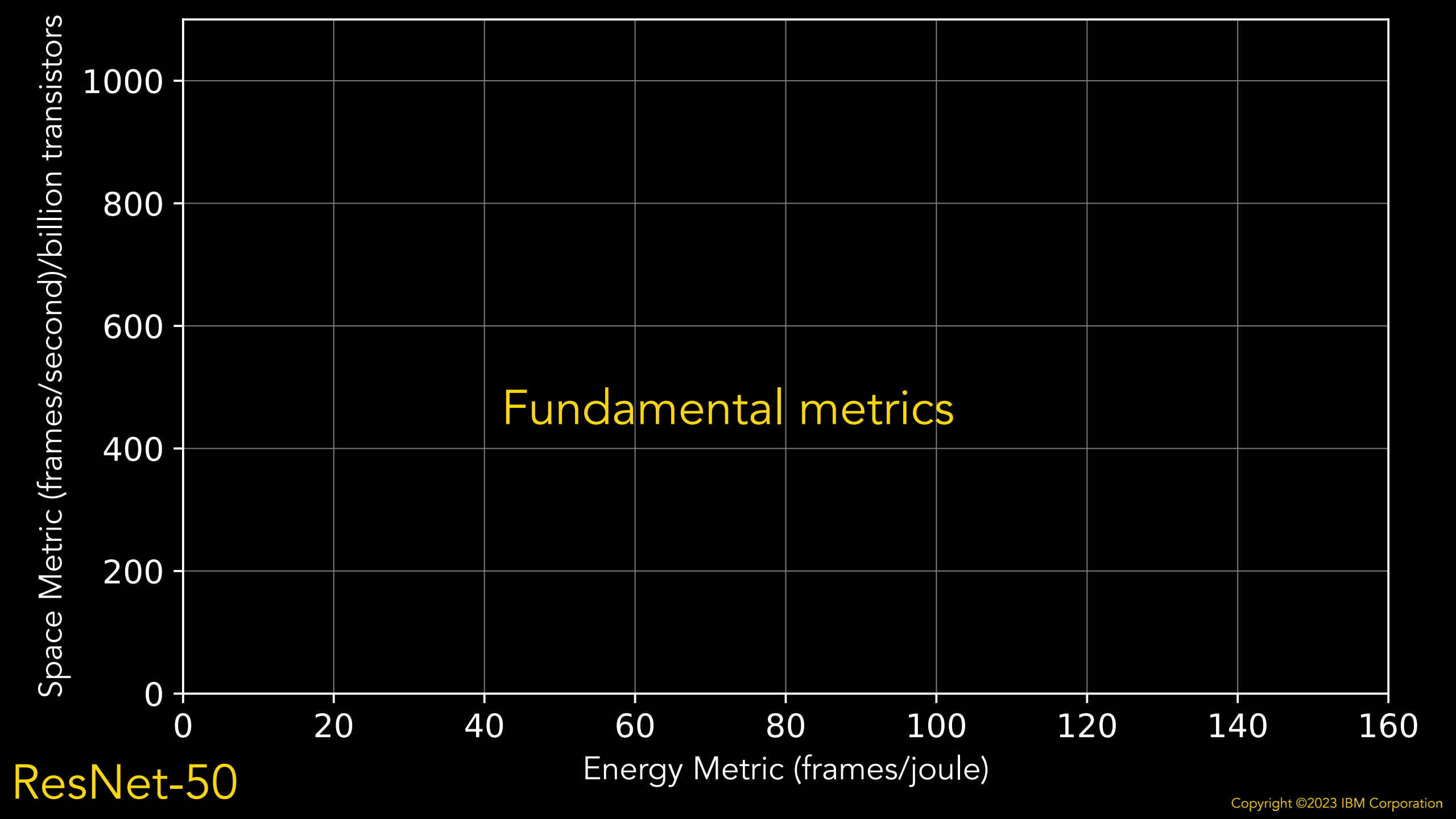
August 29, 2023

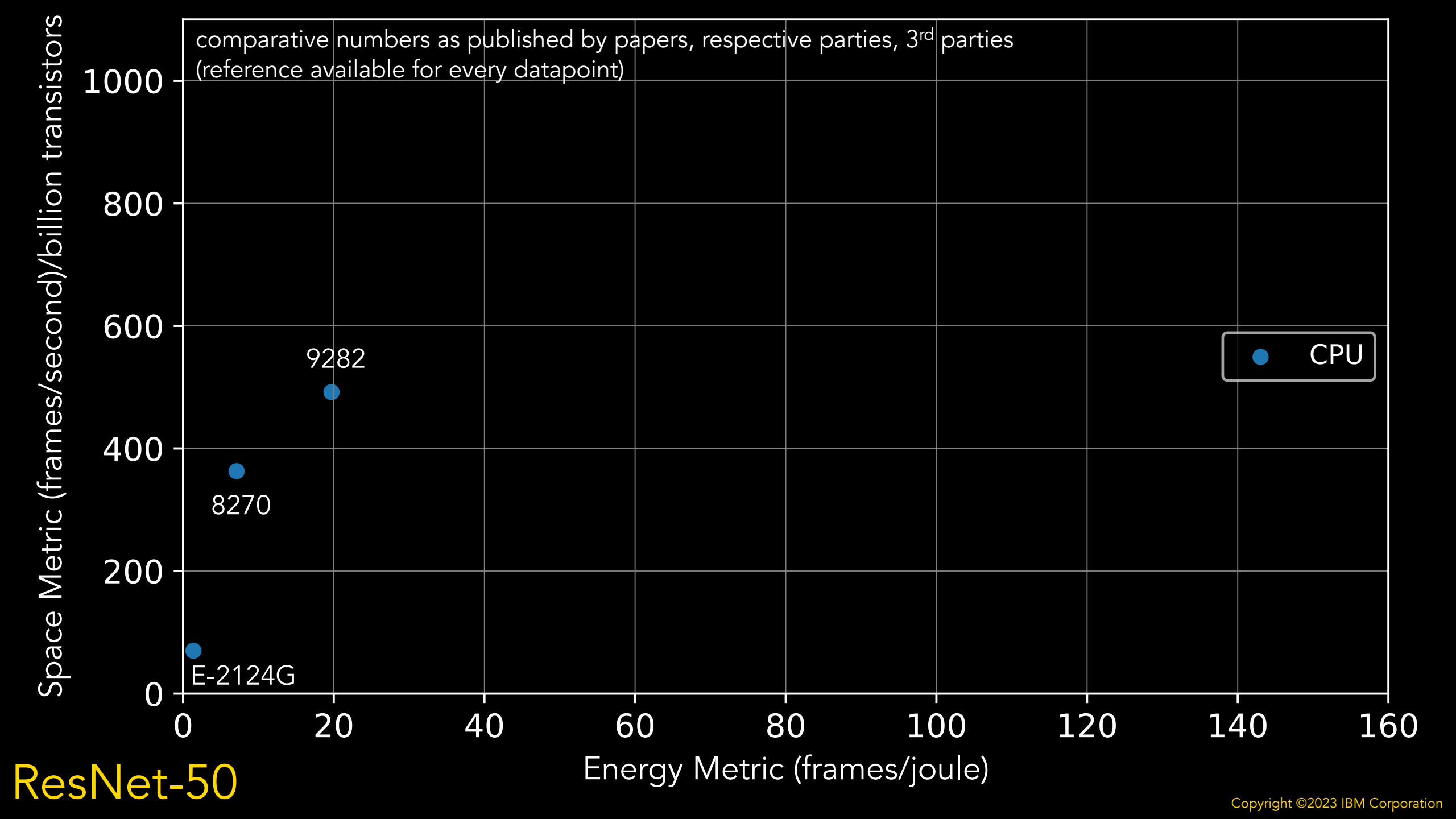
Dharmendra S. Modha*, Filipp Akopyan†,
Alexander Andreopoulos†,
Rathinakumar Appuswamy†, John V. Arthur†,
Andrew S. Cassidy†, Pallab Datta†,
Michael V. DeBole†, Steven K. Esser†,
Carlos Ortega Otero†, Jun Sawada†, Brian Taba†,
Arnon Amir, Deepika Bablani, Peter J. Carlson,
Myron D. Flickner, Rajamohan Gandhasri,
Guillaume J. Garreau, Megumi Ito,
Jennifer L. Klamo, Jeffrey A. Kusnitz,
Nathaniel J. McClatchey, Jeffrey L. McKinstry,
Yutaka Nakamura, Tapan K. Nayak,
William P. Risk, Kai Schleupen, Ben Shaw,
Jay Sivagnaname, Daniel F. Smith,
Ignacio Terrizzano, Takanori Ueda

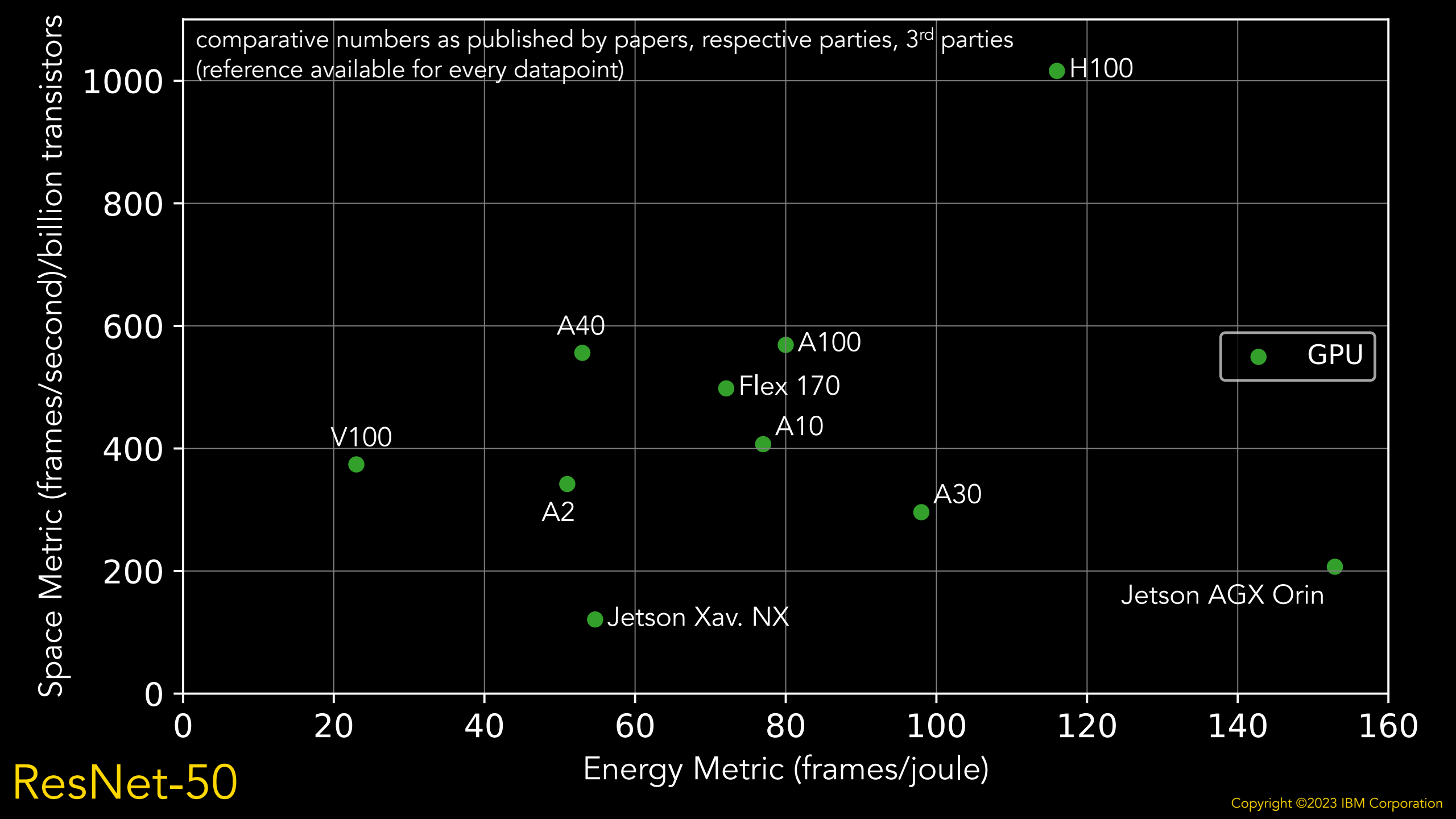
IBM Research
*Corresponding author. dmodha@us.ibm.com
†These authors contributed equally to this work.

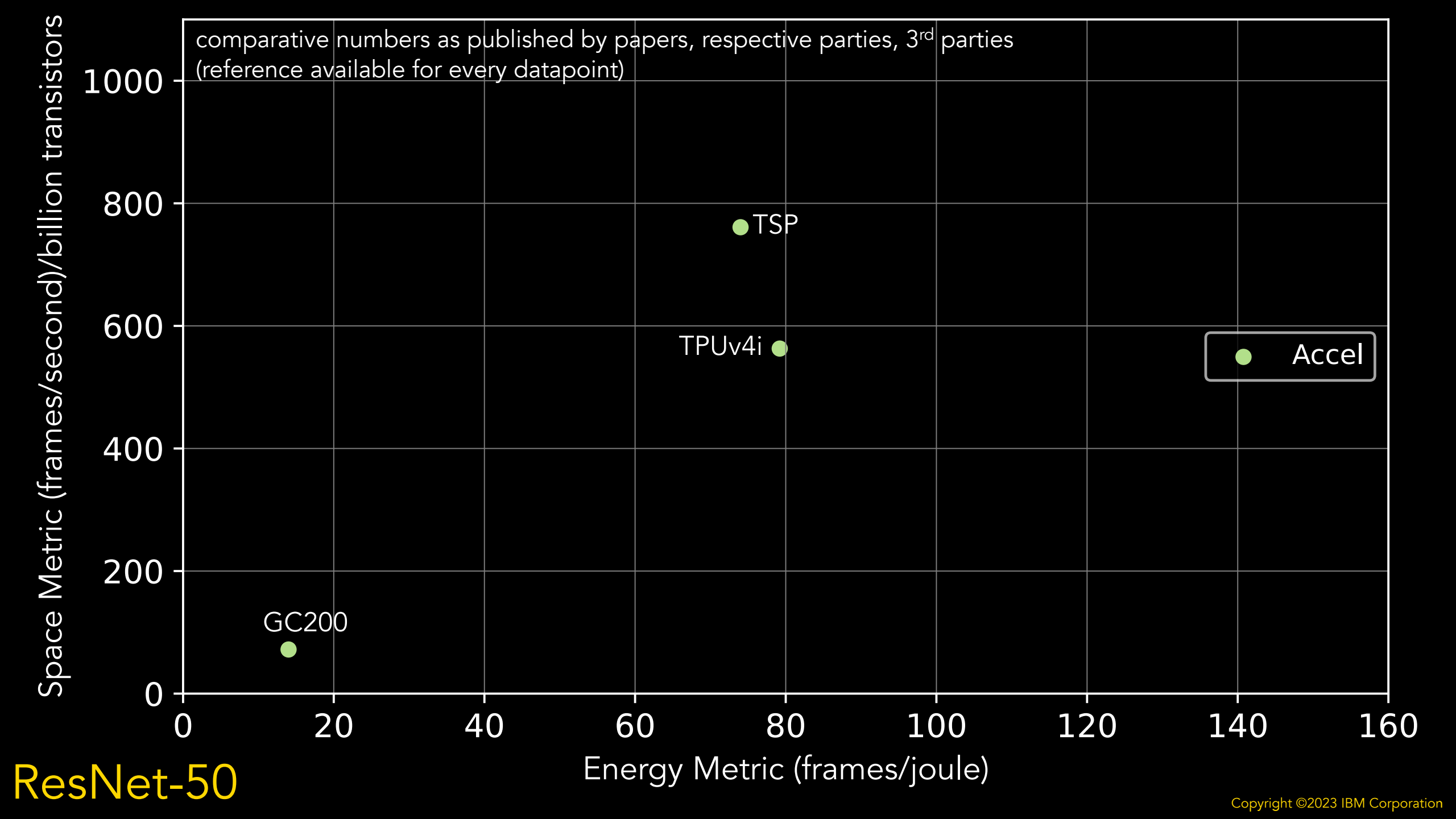
This material is based upon work supported by the United States Air Force under Contract No. FA8750-19-C-1518. Support from OUSD(R&E) is gratefully acknowledged.

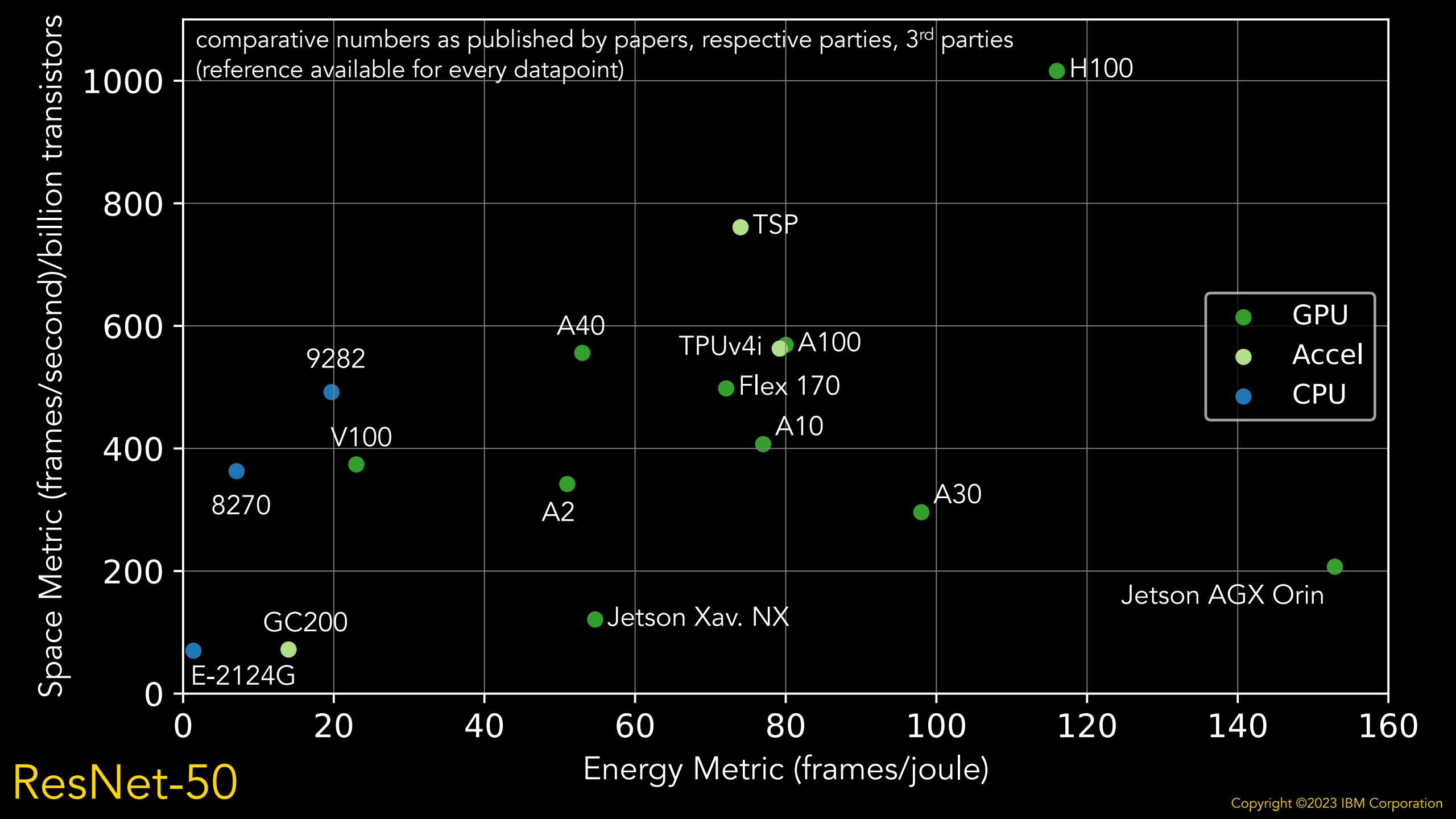
Why NorthPole?
Energy- and Space-efficiency

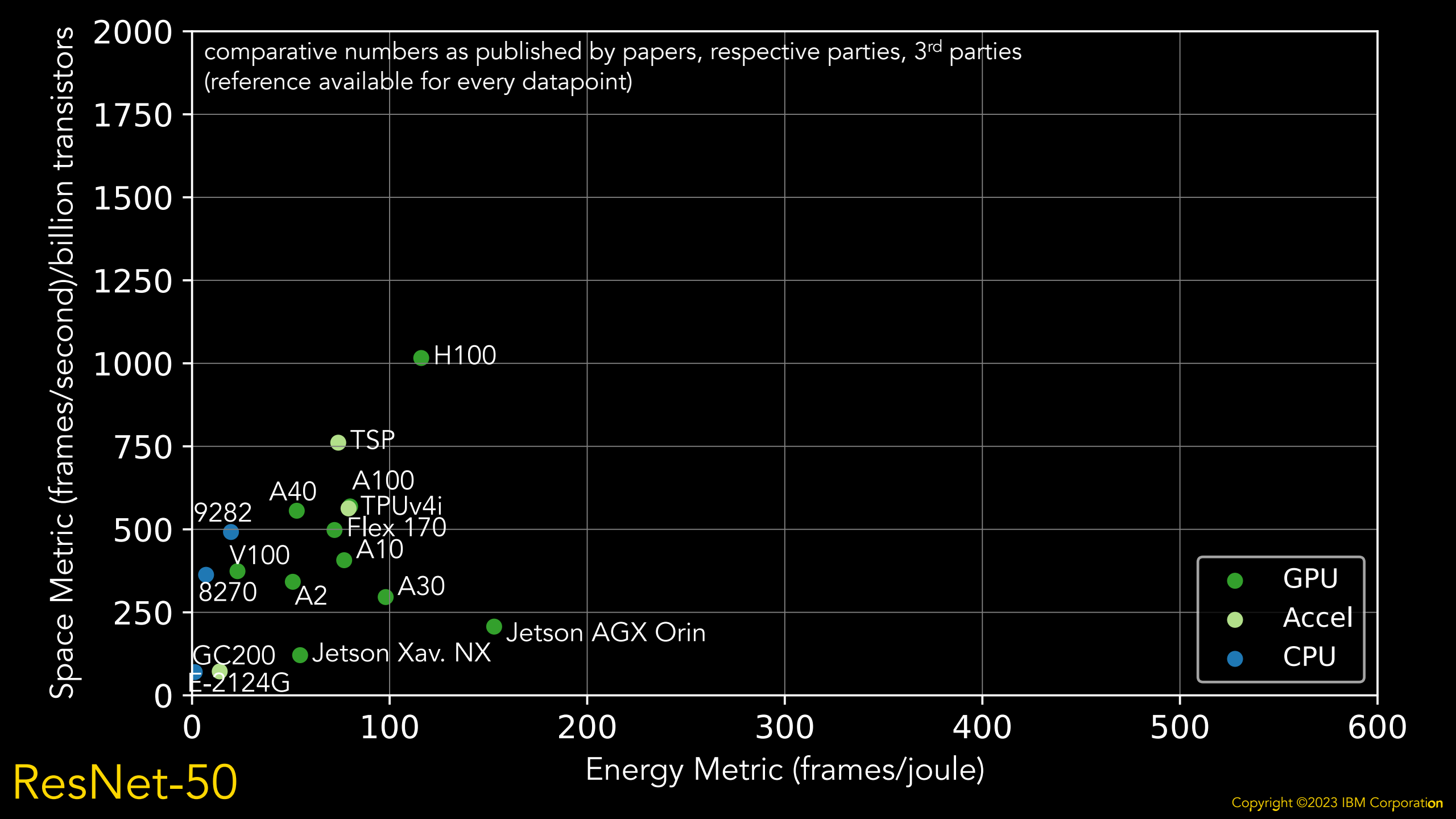


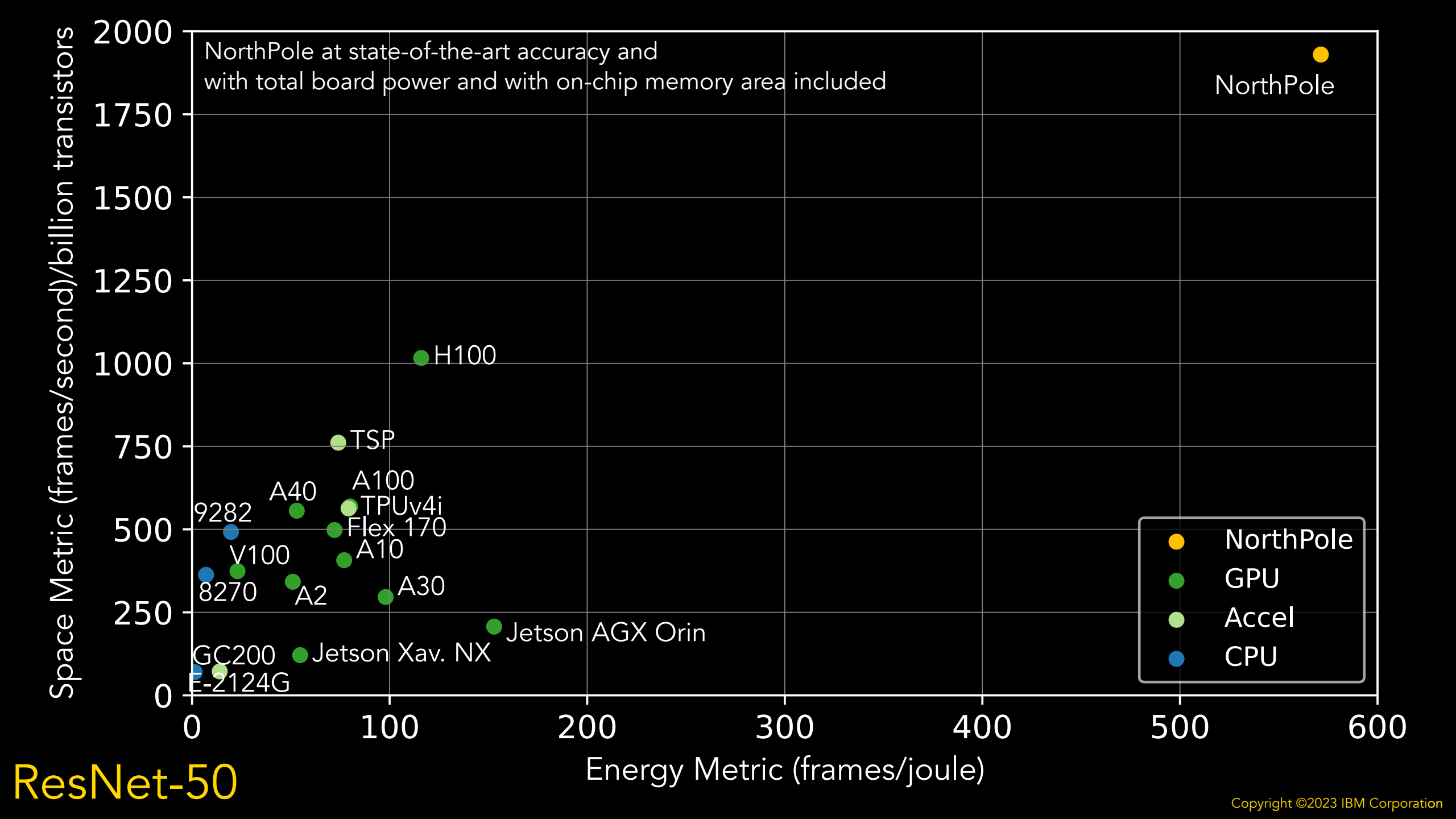




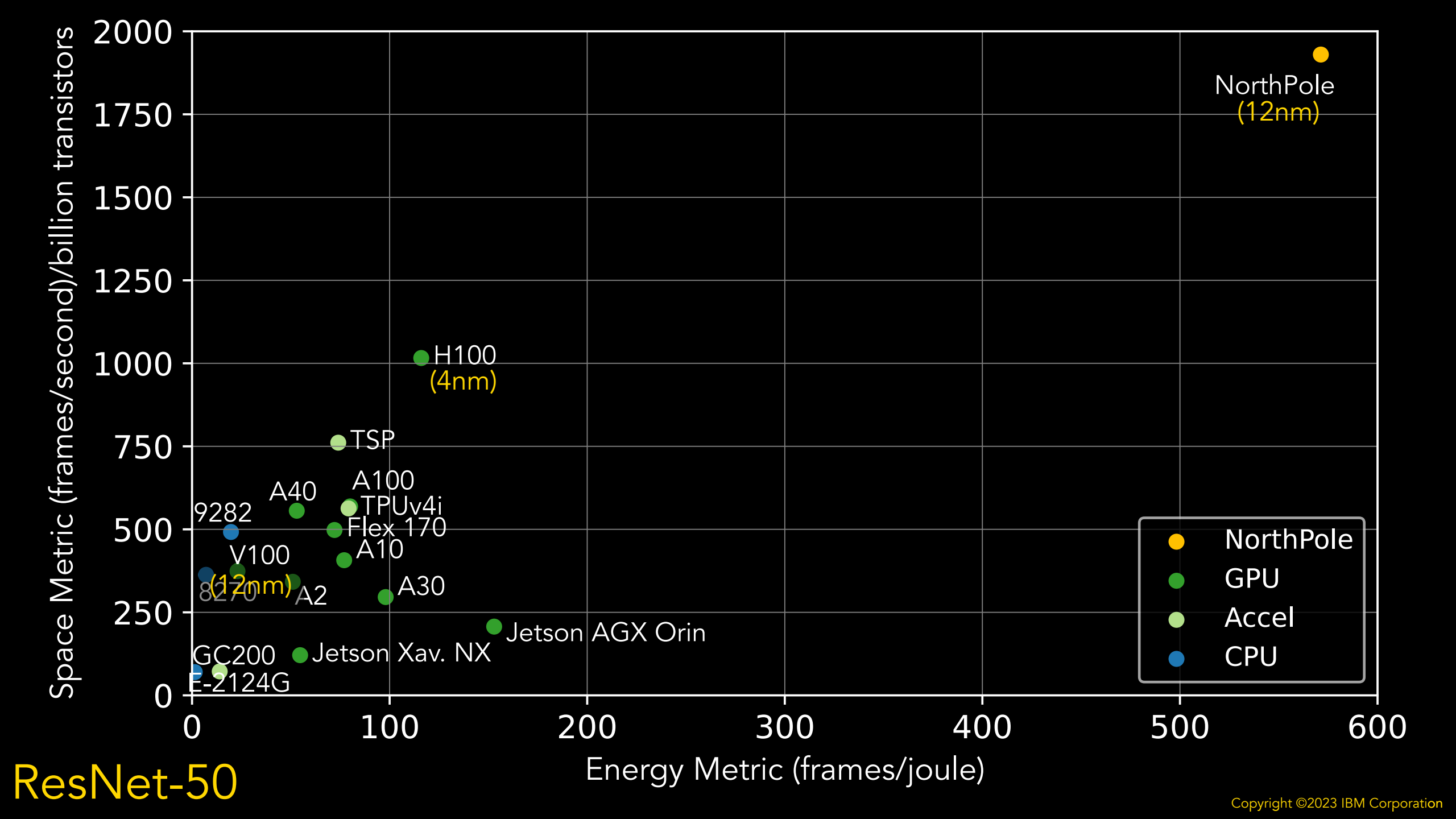


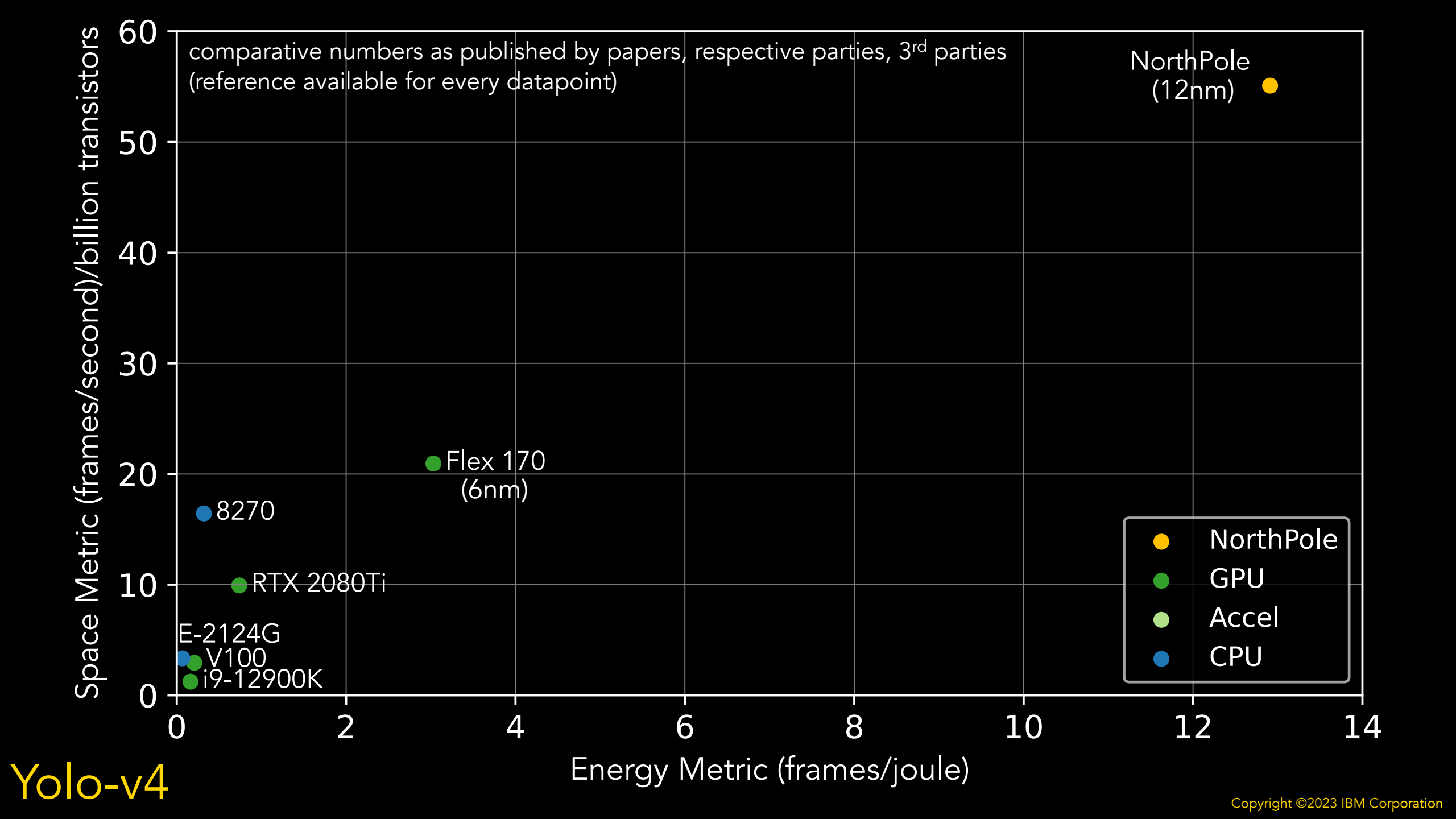




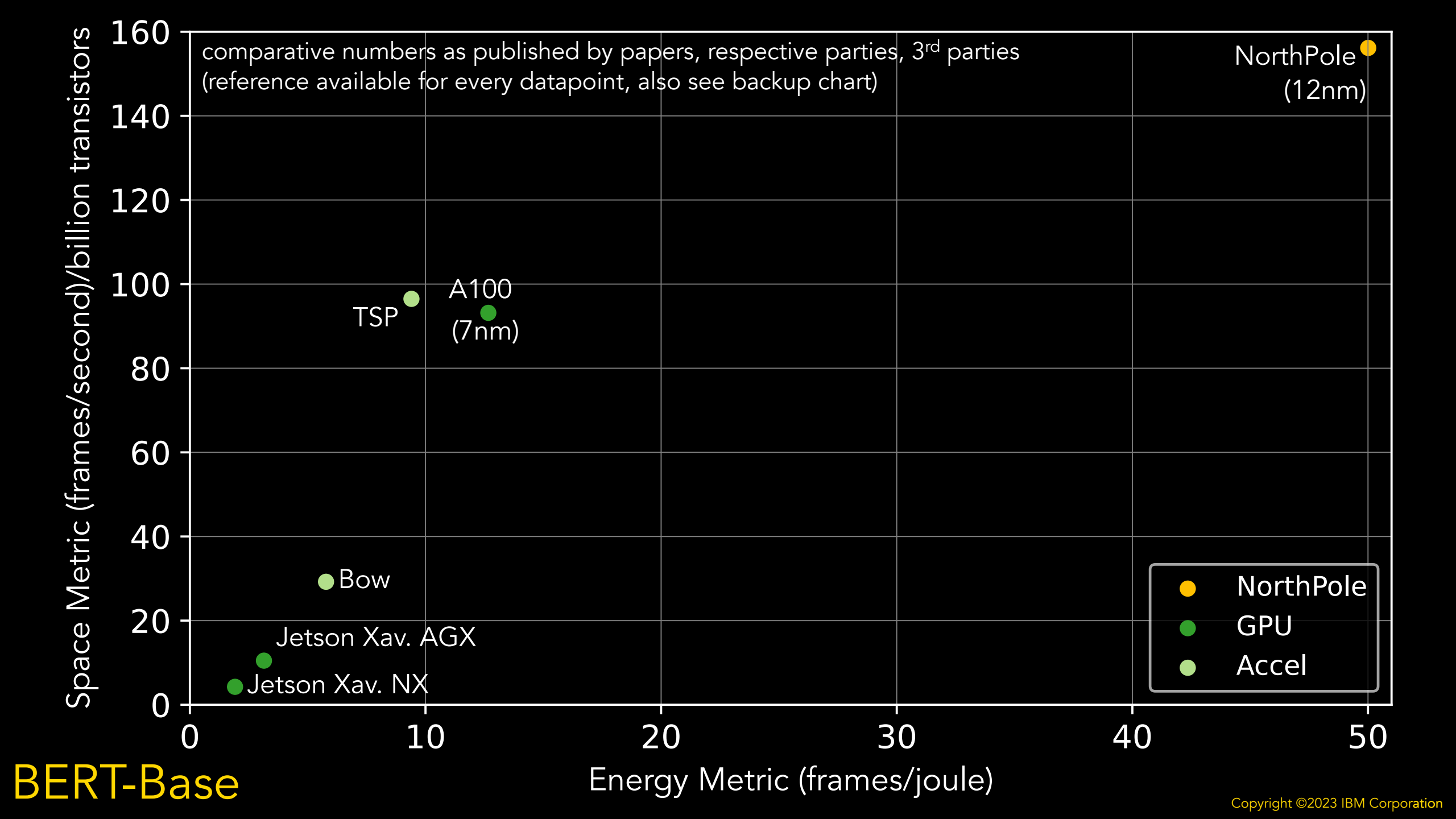


ResNet-50

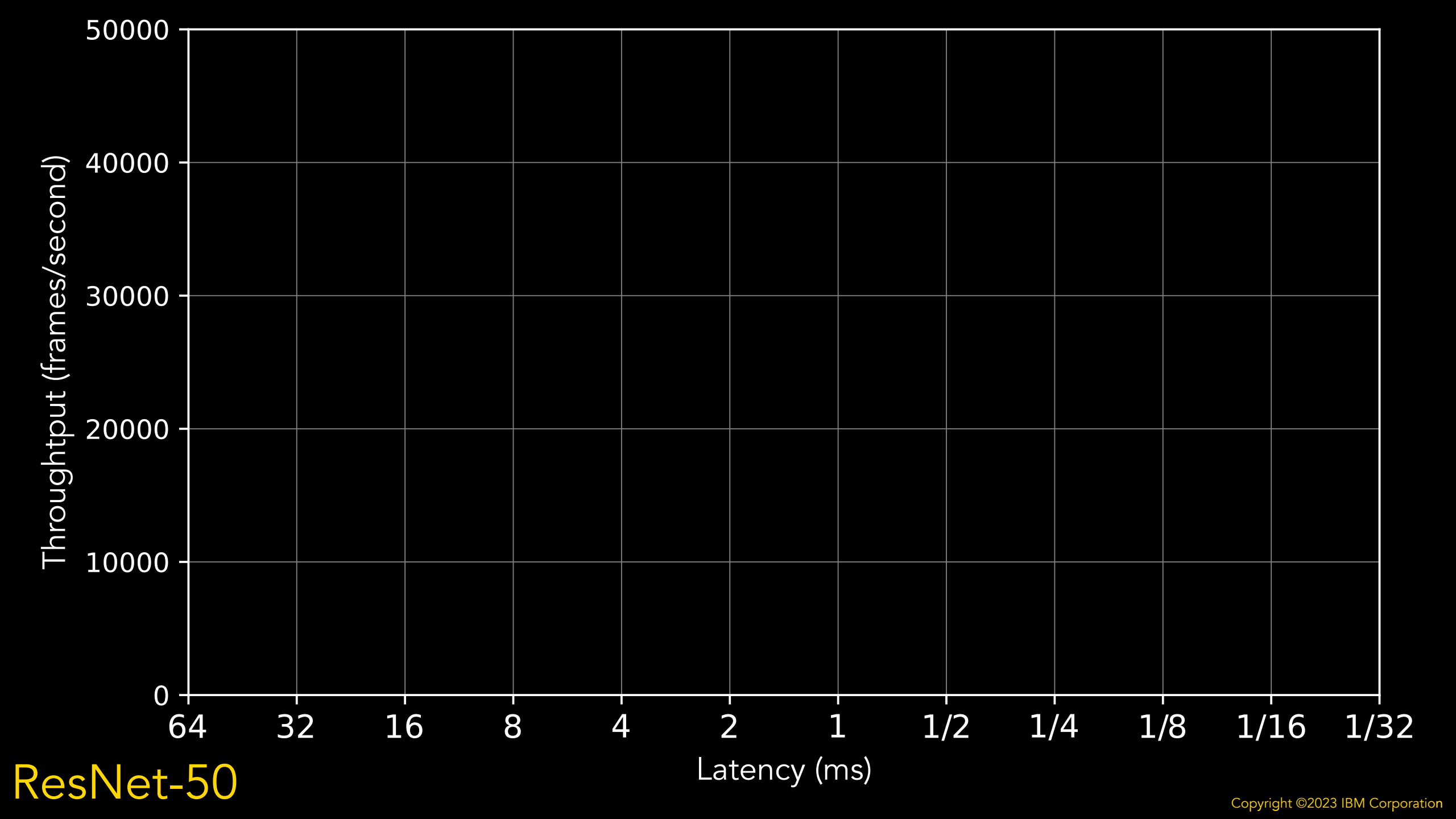


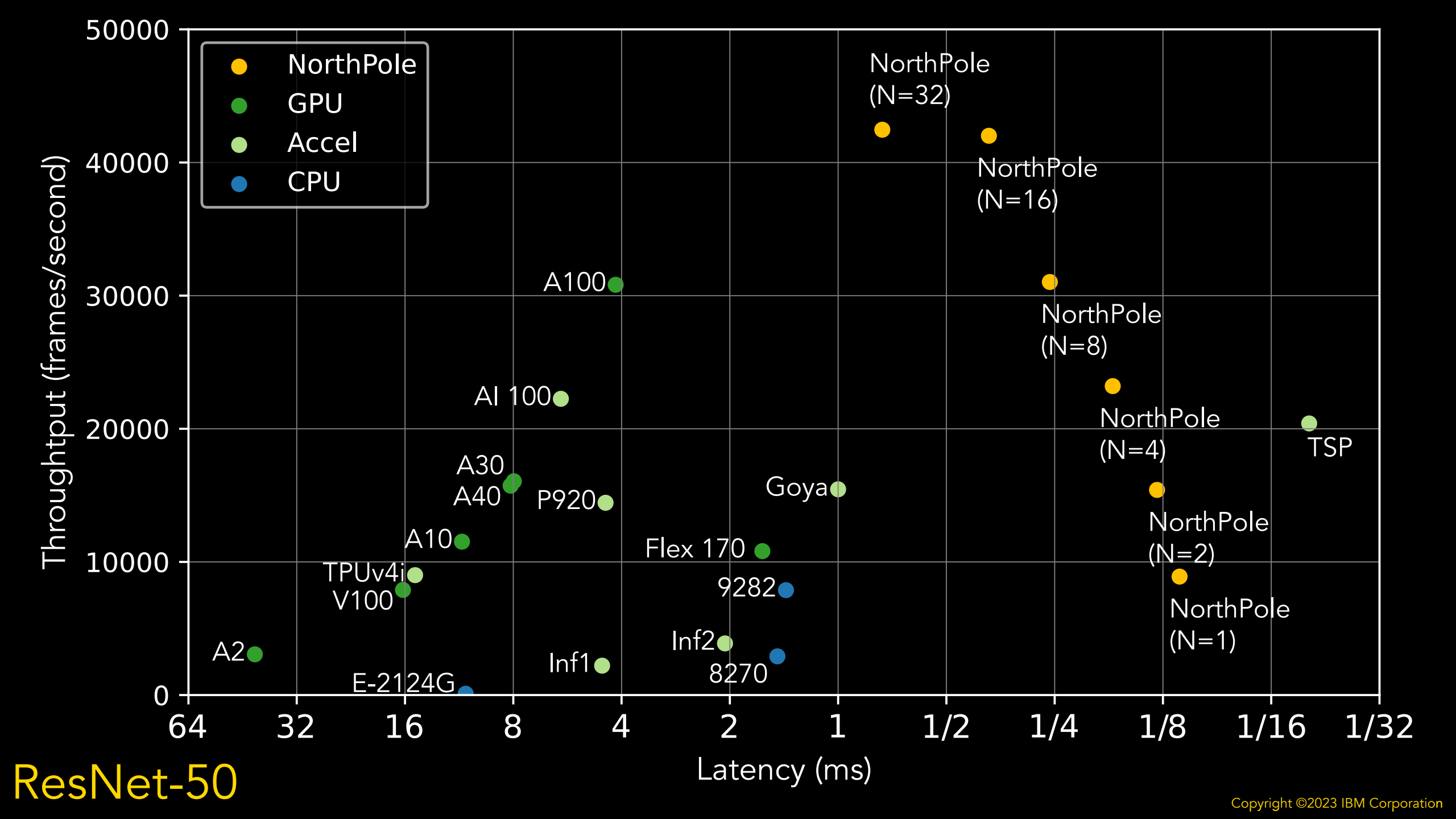


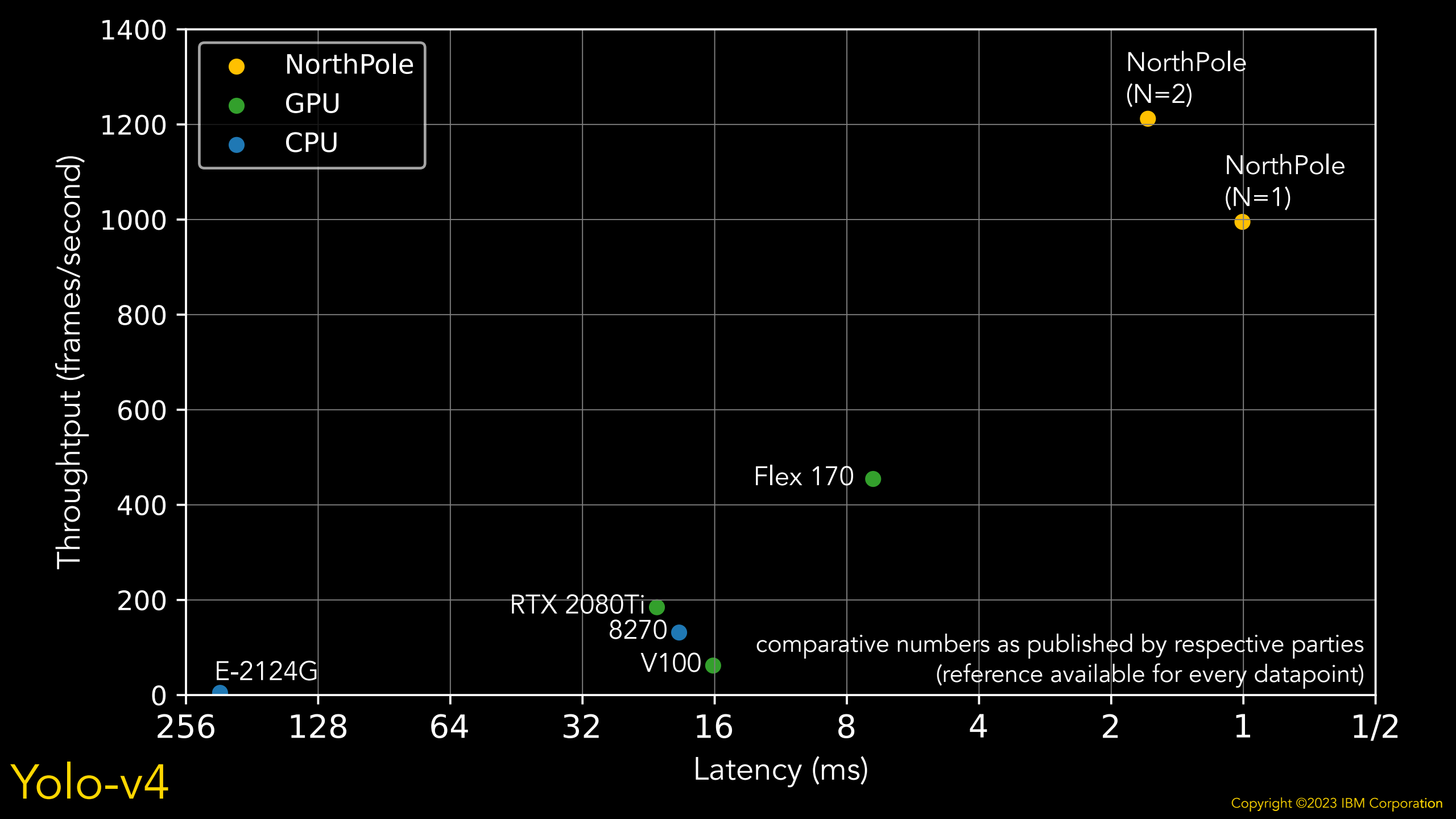
Yolo-v4



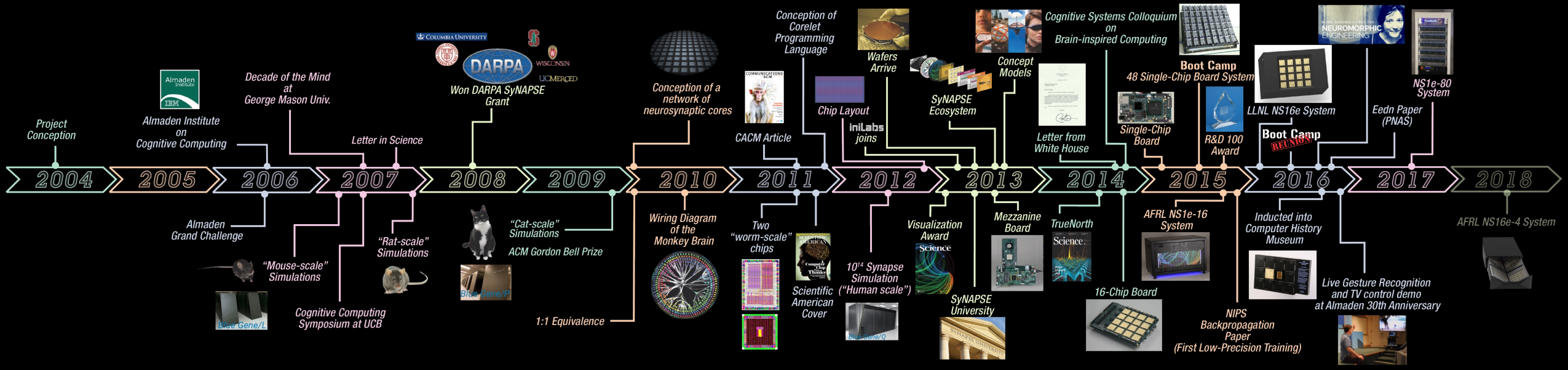
Why NorthPole?
Latency







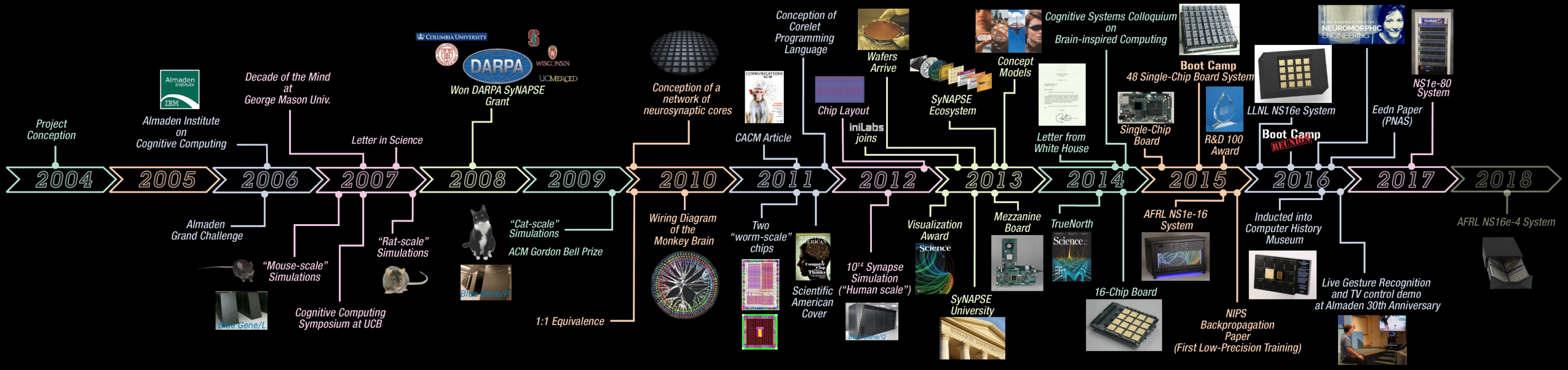
Context



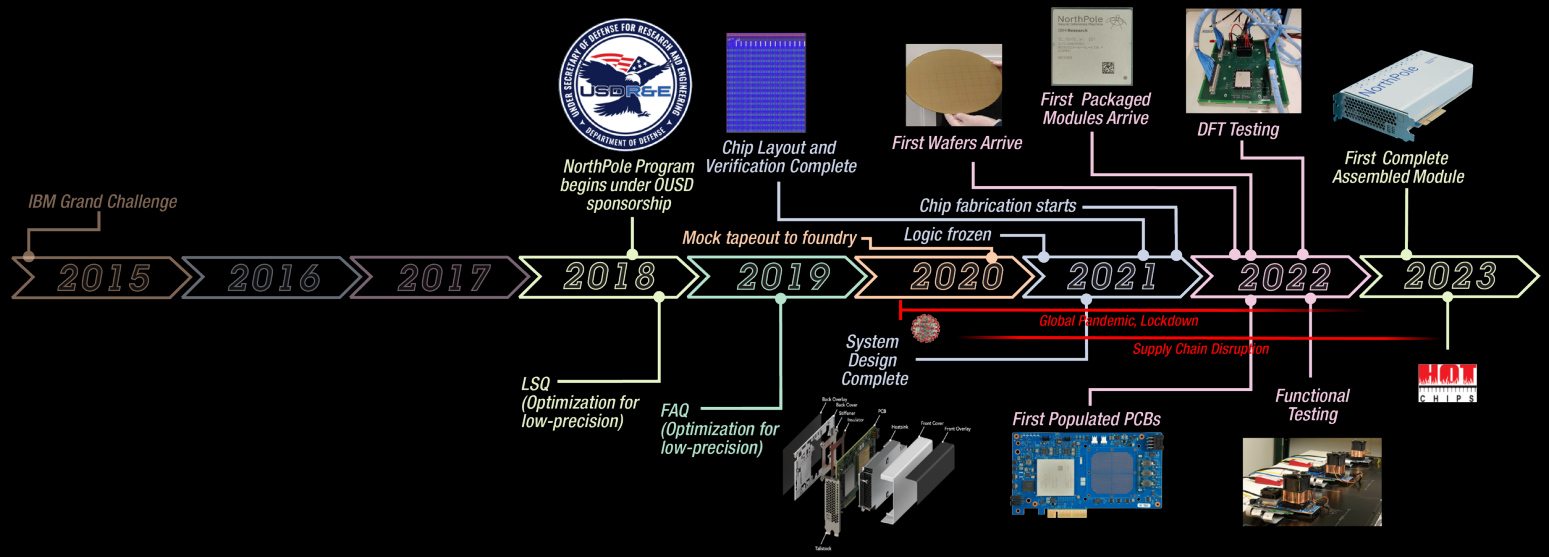
TrueNorth

2015 2016 2017 2018 2019 2020 2021 2022 2023

NorthPole in stealth mode since 2015



TrueNorth



NorthPole unveiled

NorthPole Architecture

Core-based design

Vector Matrix Multiplication (VMM)
- 8-, 4-, 2- bit precision
- 2048-4096-8192 Ops/cycle
- Mixed precision—right precision for each layer

Vector Compute Unit
- 256 Op/cycle
- FP16 precision

Activation Function Unit (not shown)
- 32 Op/cycle
- FP16 precision

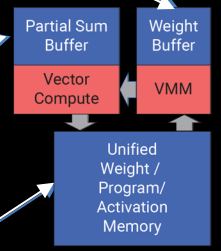
Fully pipelined operation

Memory-near-compute

Weight buffer
is near VMM

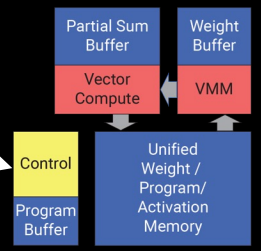
Partial Sum Buffer
is near Vector Compute Unit

768KB / core of unified memory
- weights (model)
- program
- neural activations



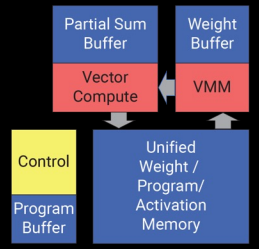
Concurrent, distributed control

Eight threads per core
- No VLIW



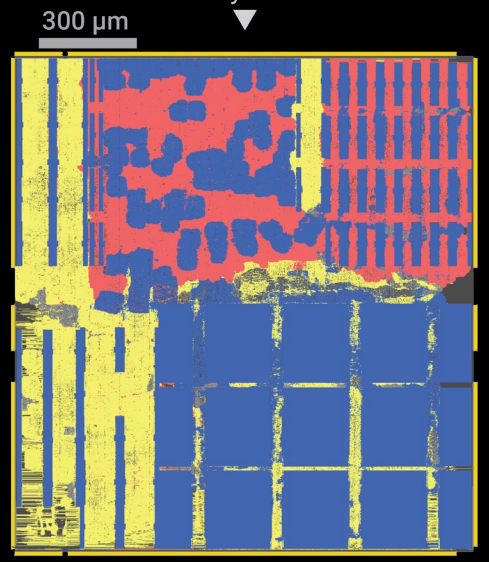
Fully prescheduled operation in the core array
- deterministic, predictable, verifiable
- no data-dependent conditional branching (breaking path with Turing's idea of conditional branching)
- no cache misses
- no stalls
- no speculative execution

Compute-intertwined-with-memory



Schematic

Layout

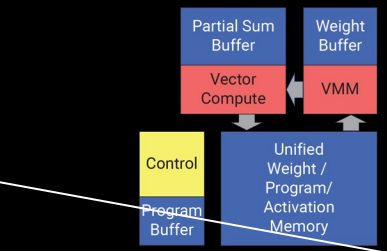


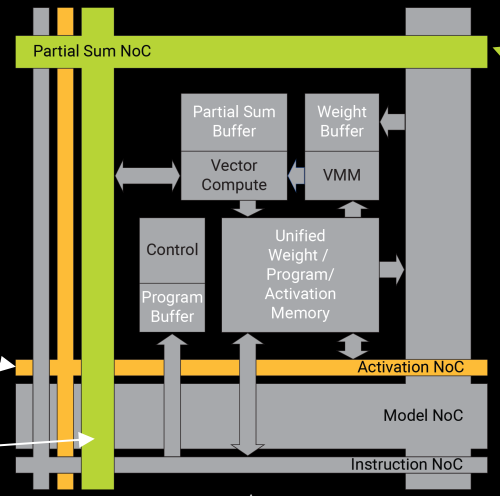
Four Networks-on-chip (NoC)

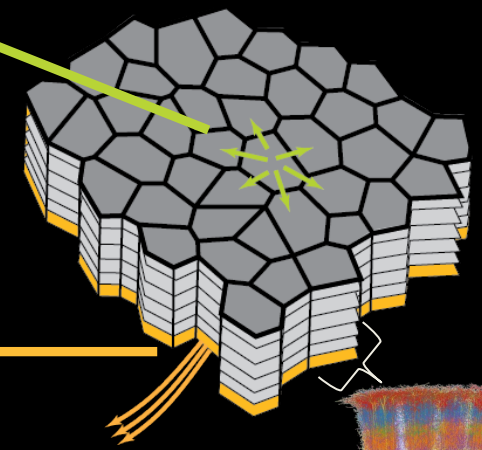
Unify distributed compute, memories

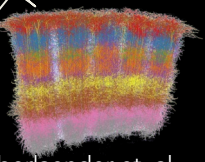
1. Activation NoC (ANoC) used between layers to reorganize neural activations

2. Partial Sum NoC (PNoC) used within a layer for neighboring cores to communicate – spatial computing









Oberlaender et. al.

Schematic

Compute

Memory

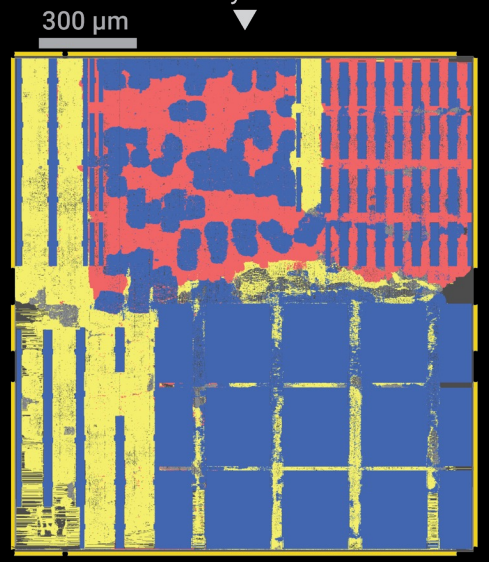
Control

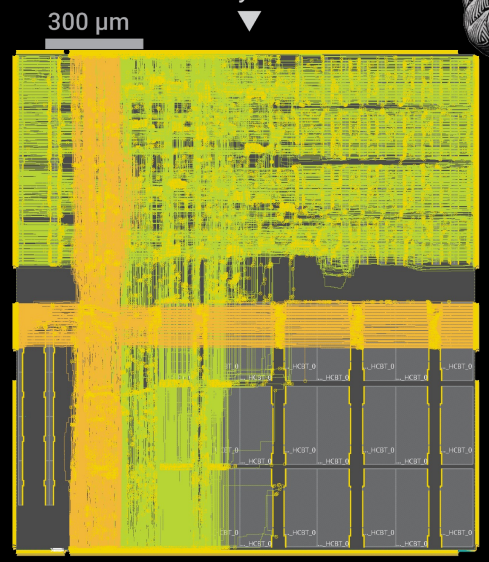
Layout

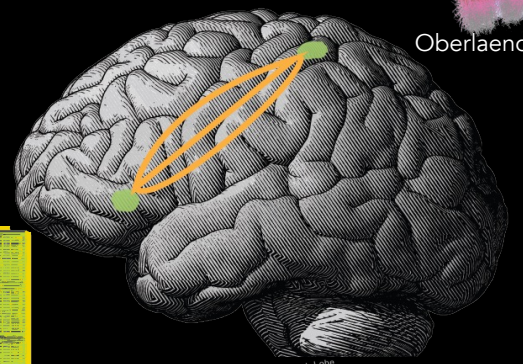
Schematic

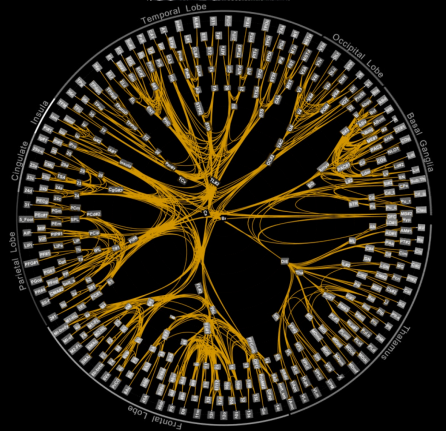
Networks-on-Chip

Layout









Four Networks-on-chip (NoC)

Unify distributed compute, memories

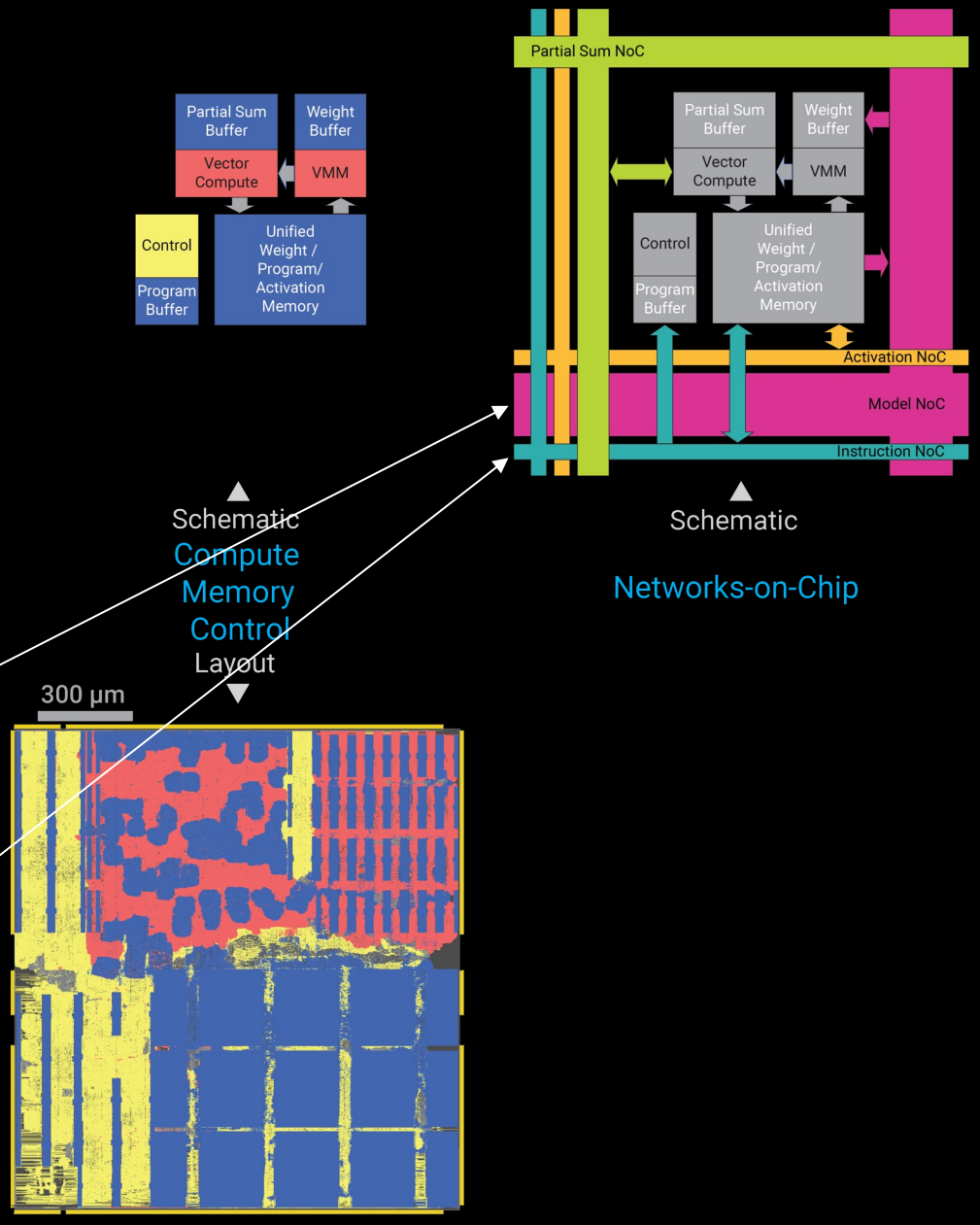
1. Activation NoC (ANoC) used between layers to reorganize neural activations

2. Partial Sum NoC (PNoC) used within a layer for neighboring cores to communicate – spatial computing

3. Model NoC (MNoC) delivers weights during layer execution

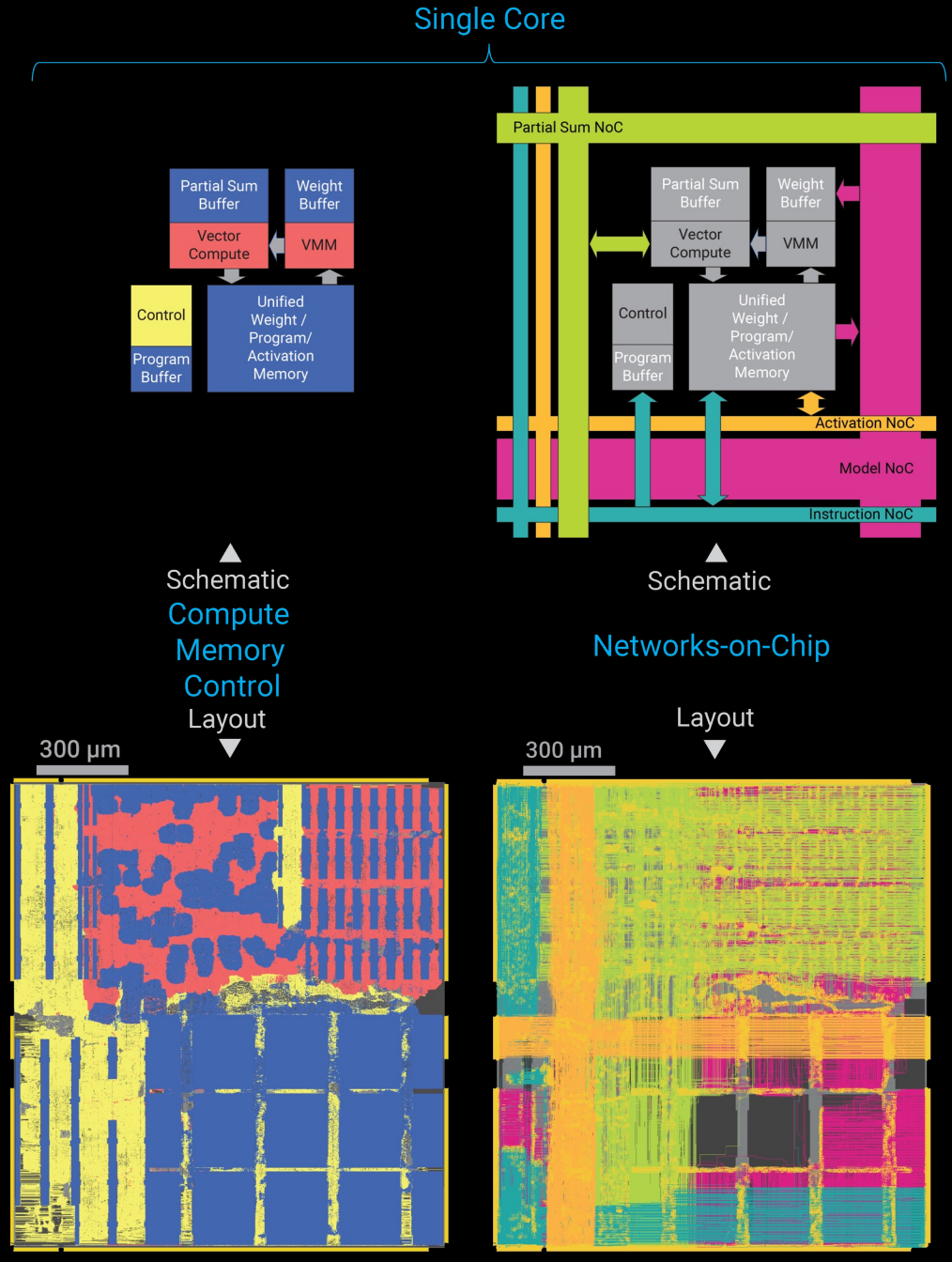
4. Instruction NoC (INoC) delivers program for each layer prior to layer start

MNoC/INoC enable reconfigurability – key to bridging brain-inspired computing with silicon



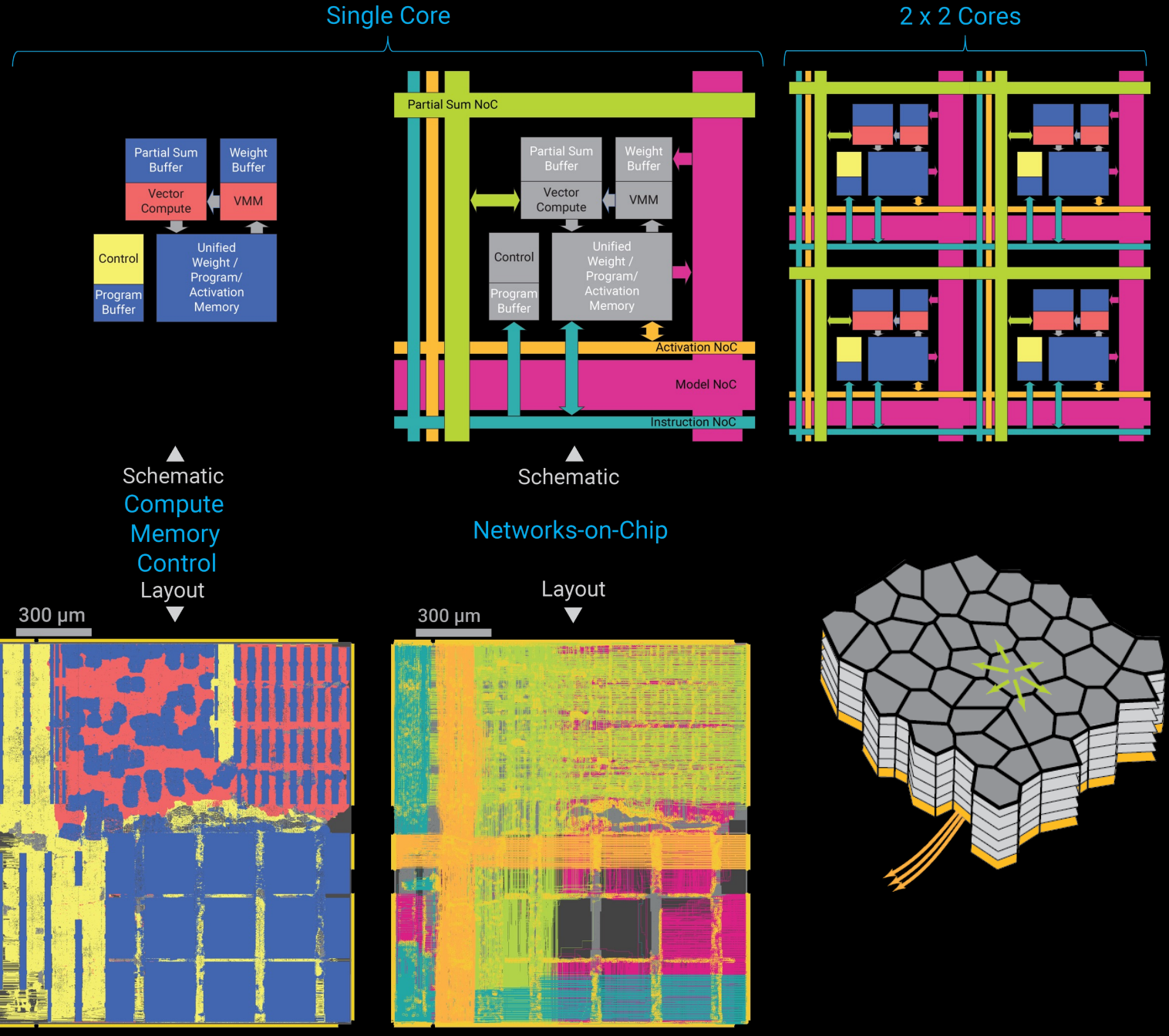
Dense interconnectivity

4,096 wires criss-cross each core in both dimensions

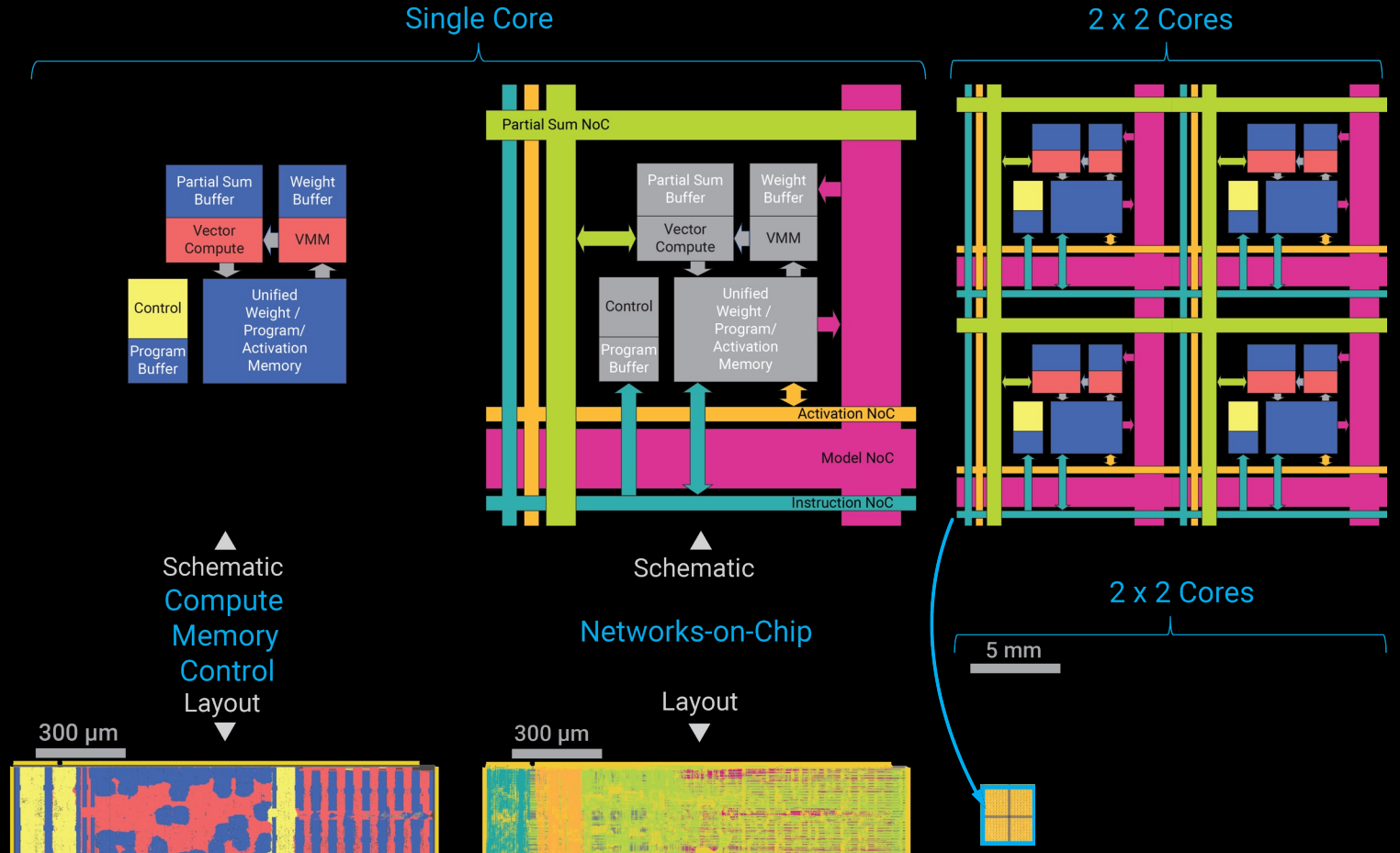


Distributed, modular core array

Cortex-like modularity enables homogeneous scalability in two-dimensions

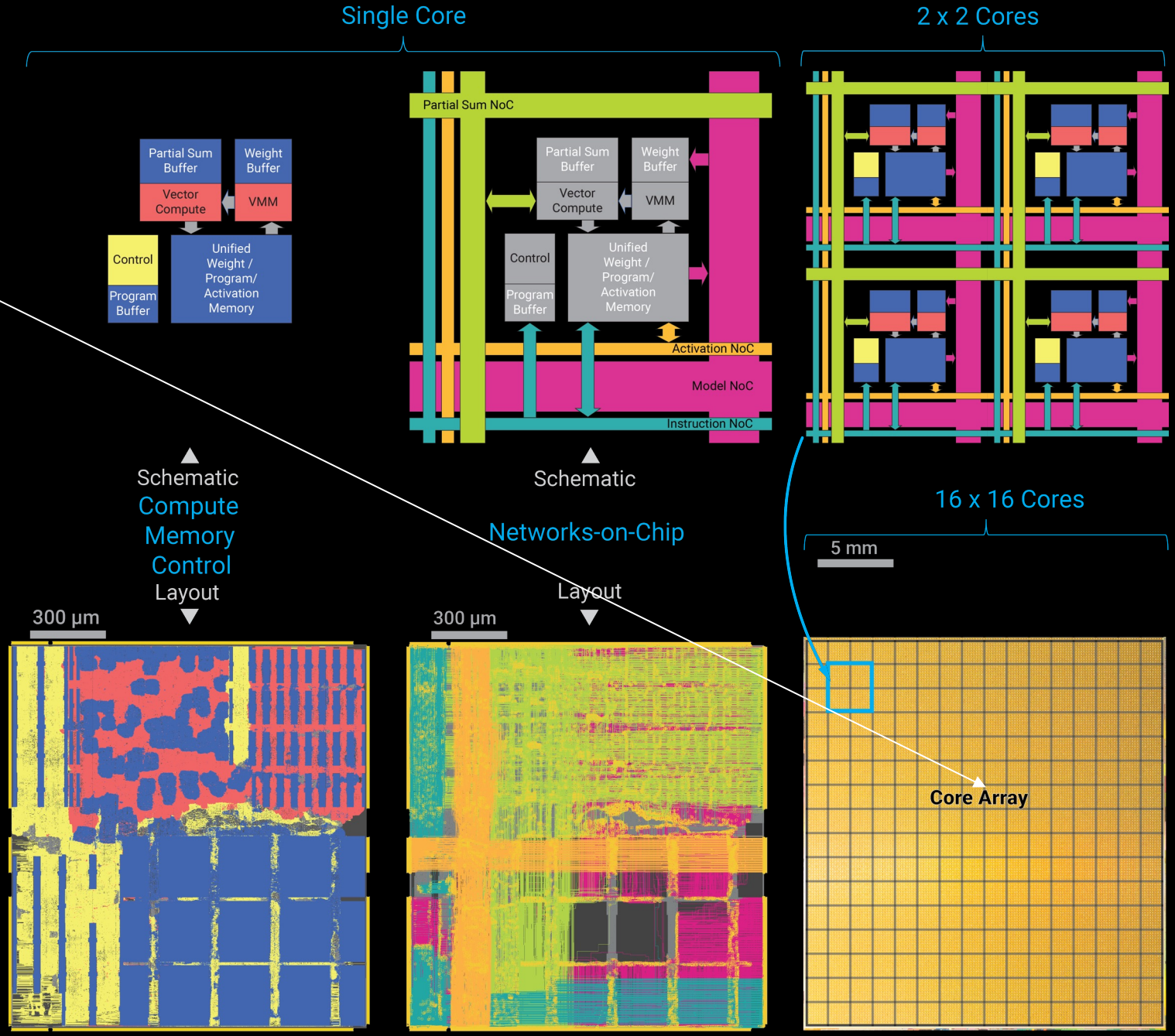


12nm Silicon Implementation



12nm Silicon Implementation

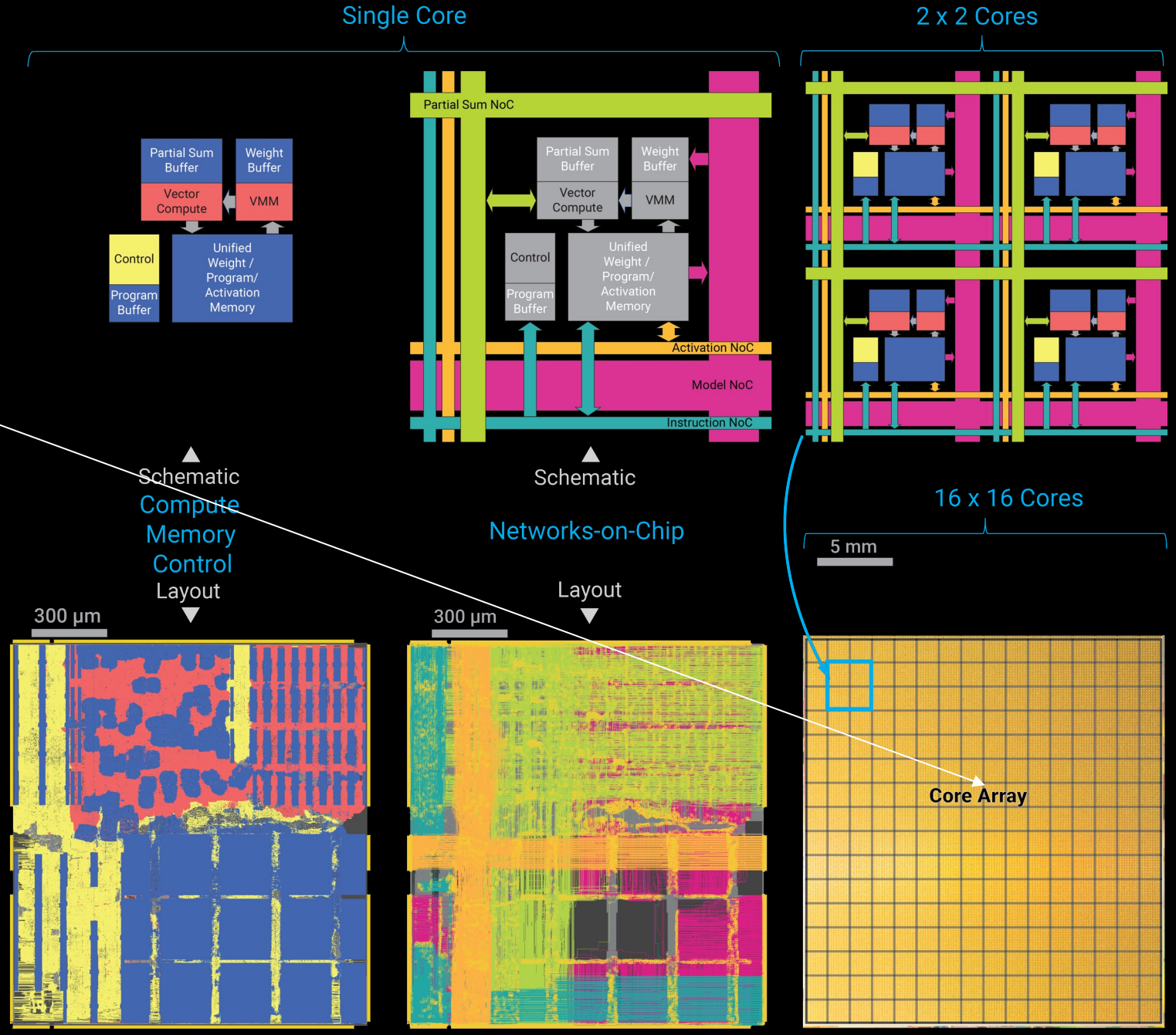
16×16 array of cores
Massive parallelism



12nm Silicon Implementation

16×16 array of cores
Massive parallelism

192MB of memory
for activations, model, program
distributed among cores

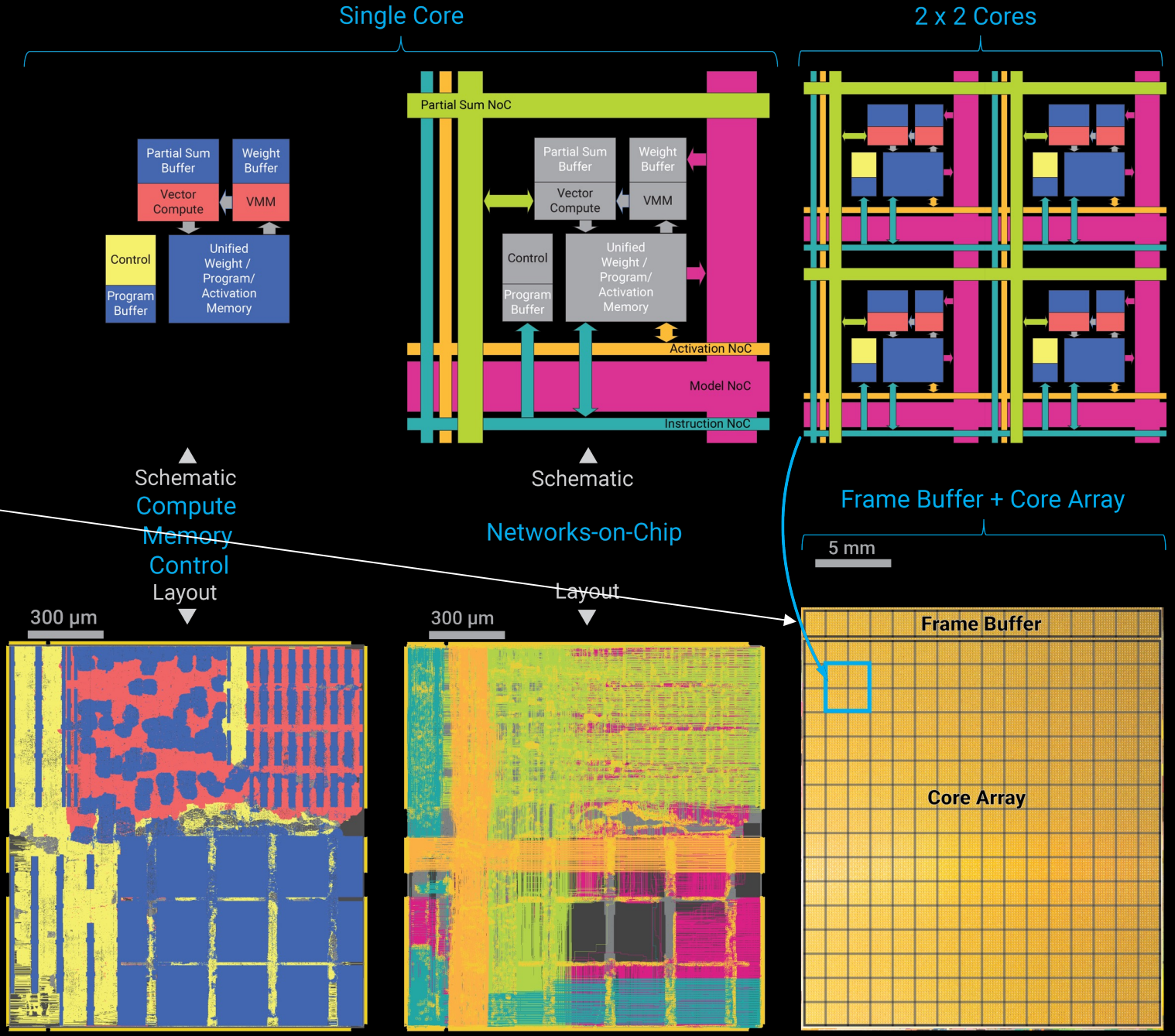


12nm Silicon Implementation

16×16 array of cores
Massive parallelism

192MB of memory
for activations, model, program
distributed among cores

32MB framebuffer for IO tensors



12nm Silicon Implementation

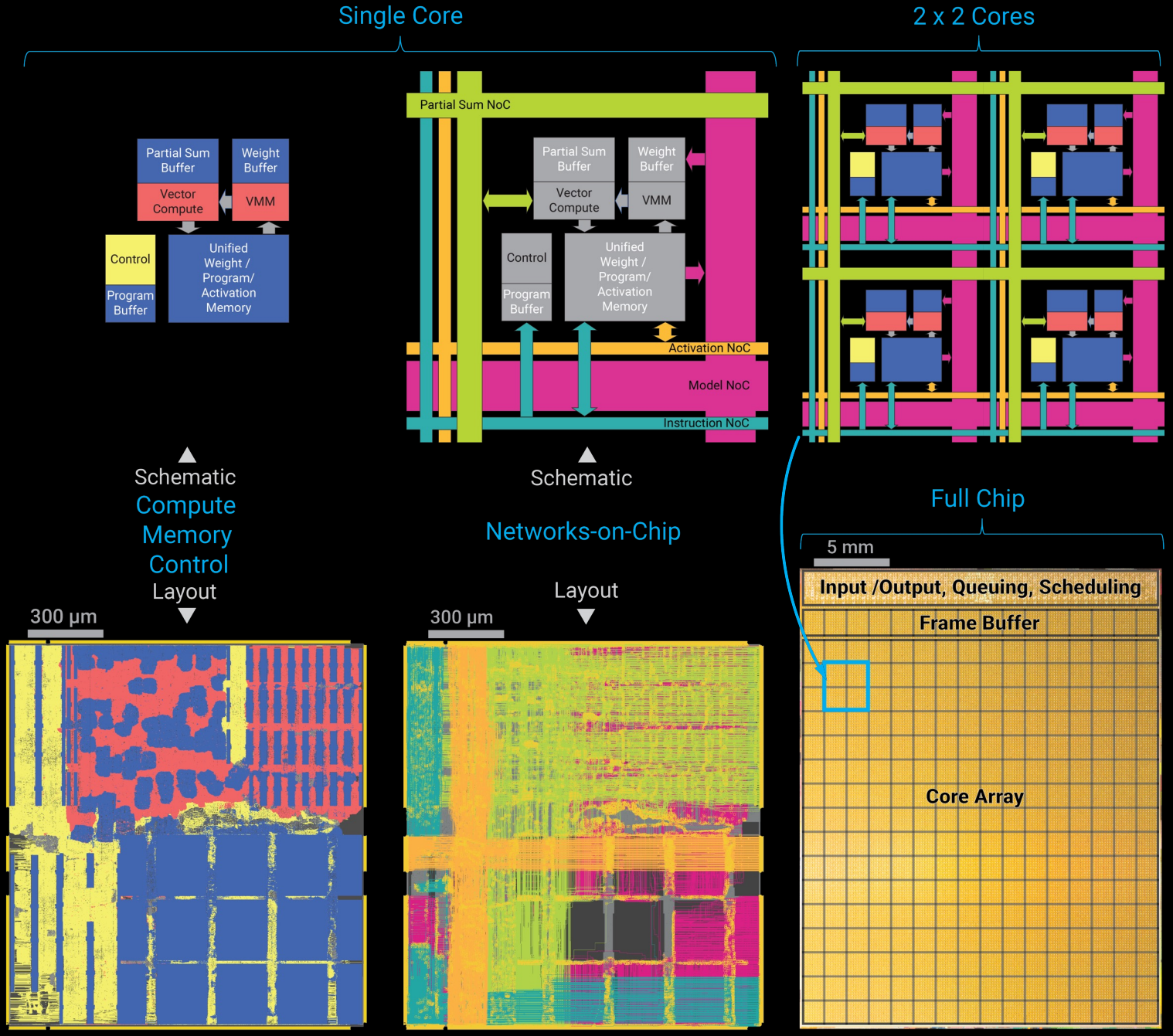
16X16 array of cores
Massive parallelism

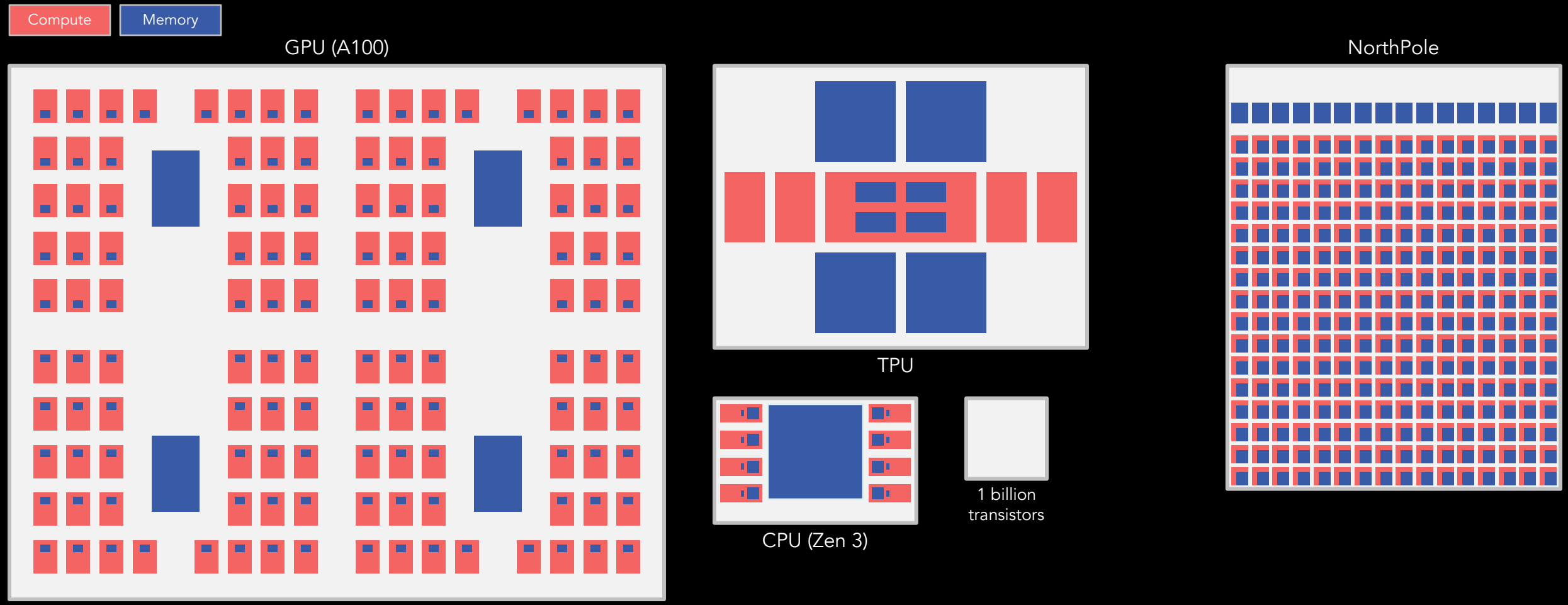
192MB of memory
for activations, model, program
distributed among cores

32MB framebuffer for IO tensors

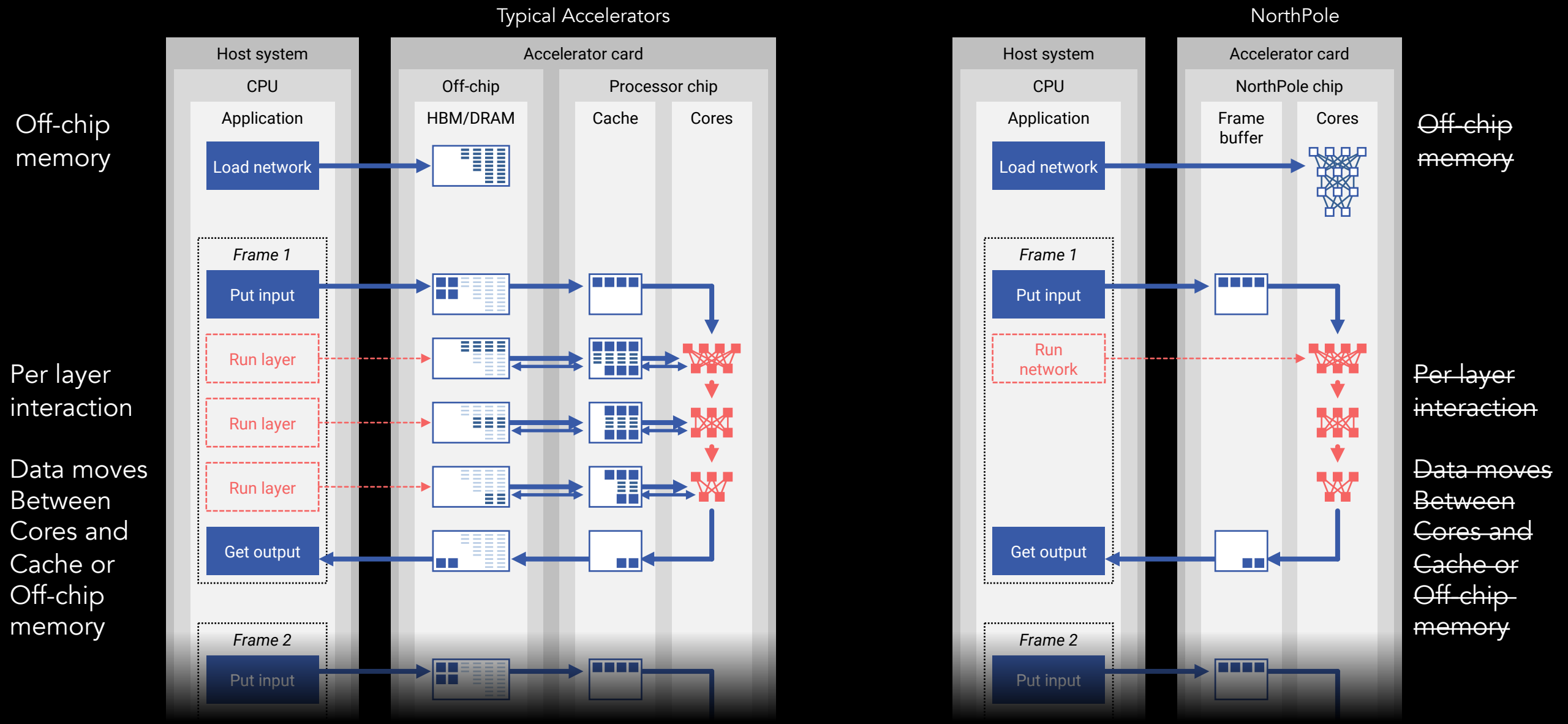
800mm^2 area, 22 Billion transistors

Functional in first-silicon





NorthPole has no centralized memory, no off-chip memory,
no von Neumann bottleneck



NorthPole has a simple usage/IO model: write tensor, run, read tensor – essentially an active memory
Entire network is on-chip, No layer-by-layer interaction, Minimum load on the host, Minimum IO bandwidth

Achieving State-of-the-art Accuracy with Mixed Precision

Quantization-aware-training to maximize accuracy via PyTorch extension

Two algorithms:

FAQ: Finetuning After Quantization

LSQ: Learned Step-size Quantization

Demonstrated on classification tasks

Network	Method	Precision (w,a)	Accuracy	Accuracy
ResNet-18	baseline	32,32	69.76	89.08
ResNet-18	**Apprentice**	**4,8**	**70.40**	-
ResNet-18	**FAQ (This paper)**	**8,8**	**70.02**	**89.32**
ResNet-18	**FAQ (This paper)**	**4,4**	**69.78±0.04**	**89.11±0.03**
ResNet-18	Joint Training	4,4	69.3	-
ResNet-18	UNIQ	4,8	67.02	-
ResNet-18	Distillation	4,32	64.20	-
ResNet-34	baseline	32,32	73.30	91.42
ResNet-34	**FAQ (This paper)**	**8,8**	**73.71**	**91.63**
ResNet-34	**FAQ (This paper)**	**4,4**	**73.31**	**91.32**
ResNet-34	UNIQ	4,32	73.1	-
ResNet-34	Apprentice	4,8	73.1	-
ResNet-34	UNIQ	4,8	71.09	-
ResNet-50	baseline	32,32	76.15	92.87
ResNet-50	**FAQ (This paper)**	**8,8**	**76.52**	**93.09**
ResNet-50	**FAQ (This paper)**	**4,4**	**76.27**	**92.89**
ResNet-50	**EL-Net**	**4,4**	**75.9**	**92.4**
ResNet-50	IOA	8,8	74.9	-
ResNet-50	Apprentice	4,8	74.7	-
ResNet-50	UNIQ	4,8	73.37	-
ResNet-152	baseline	32,32	78.31	94.06
ResNet-152	**FAQ (This paper)**	**4,4**	**78.64**	**94.12**
ResNet-152	**FAQ (This paper)**	**8,8**	**78.54**	**94.07**
Inception-v3	baseline	32,32	77.45	93.56
Inception-v3	**FAQ (This paper)**	**8,8**	**77.60**	**93.59**
Inception-v3	FAQ (This paper)	4,4	77.33	93.59
Inception-v3	IOA	8,8	74.2	92.2
Densenet-161	baseline	32,32	77.65	93.80
Densenet-161	**FAQ (This paper)**	**4,4**	**77.90**	**93.83**
Densenet-161	**FAQ (This paper)**	**8,8**	**77.84**	**93.91**
VGG-16bn	baseline	32,32	73.36	91.50
VGG-16bn	**FAQ (This paper)**	**4,4**	**73.87**	**91.67**
VGG-16bn	**FAQ (This paper)**	**8,8**	**73.66**	**91.56**

		Top-1 Accuracy @ Precision 3	4	8
ResNet-18	*Full precision: 70.5*			
	LSQ (Ours)	**67.6**	**70.2**	**71.1**	**71.1**
	QIL	65.7	69.2	70.1	
	FAQ			69.8	70.0
	LQ-Nets	64.9	68.2	69.3	
	PACT	64.4	68.1	69.2	
	NICE		67.7	69.8	
	Regularization	61.7		67.3	68.1
ResNet-34	*Full precision: 74.1*			
	LSQ (Ours)	**71.6**	**73.4**	**74.1**	**74.1**
	QIL	70.6	73.1	73.7	
	LQ-Nets	69.8	71.9		
	NICE		71.7	73.5	
	FAQ			73.3	73.7
ResNet-50	*Full precision: 76.9*			
	LSQ (Ours)	**73.7**	**75.8**	**76.7**	**76.8**
	PACT	72.2	75.3	76.5	
	NICE		75.1	76.5	
	FAQ			76.3	76.5
	LQ-Nets	71.5	74.2	75.1	
ResNet-101	*Full precision: 78.2*			
	LSQ (Ours)	**76.1**	**77.5**	**78.3**	**78.1**
ResNet-152	*Full precision: 78.9*			
	LSQ (Ours)	**76.9**	**78.2**	**78.5**	**78.5**
	FAQ			78.4	**78.5**
VGG-16bn	*Full precision: 73.4*			
	LSQ (Ours)	**71.4**	**73.4**	**74.0**	73.5
	FAQ			73.9	**73.7**
Squeeze Next-23-2x	*Full precision: 67.3*			
	LSQ (Ours)	**53.3**	**63.7**	**67.4**	**67.0**

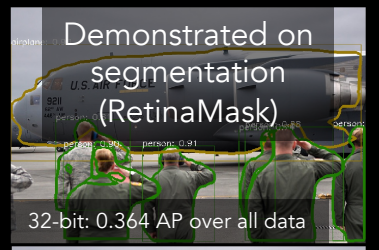
Demonstrated on speech recognition (Deep Speech 2)

Ground truth: *"the two doctors therefore entered the room alone"*

32-bit: "the two doctors therefore entered the room alone"
28.69 WER (low is better) over all data

8-bit: "the two doctors therefore entered the room alone"
28.44 WER (low is better) over all data

Demonstrated on segmentation (RetinaMask)



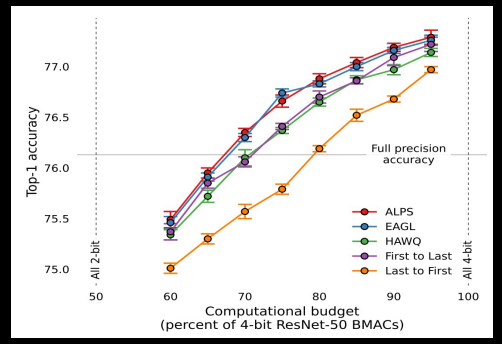
Selecting network layer precision to maximize accuracy / throughput

Two algorithms:

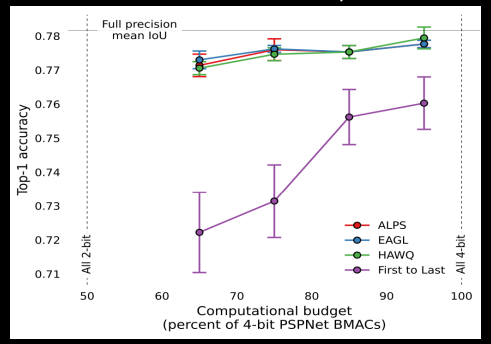
EAGL: Entropy Approximation Guided Layer selection

ALPS: Accuracy-aware Layer Precision Selection

ResNet-50 on ImageNet



PSPNet on Cityscapes



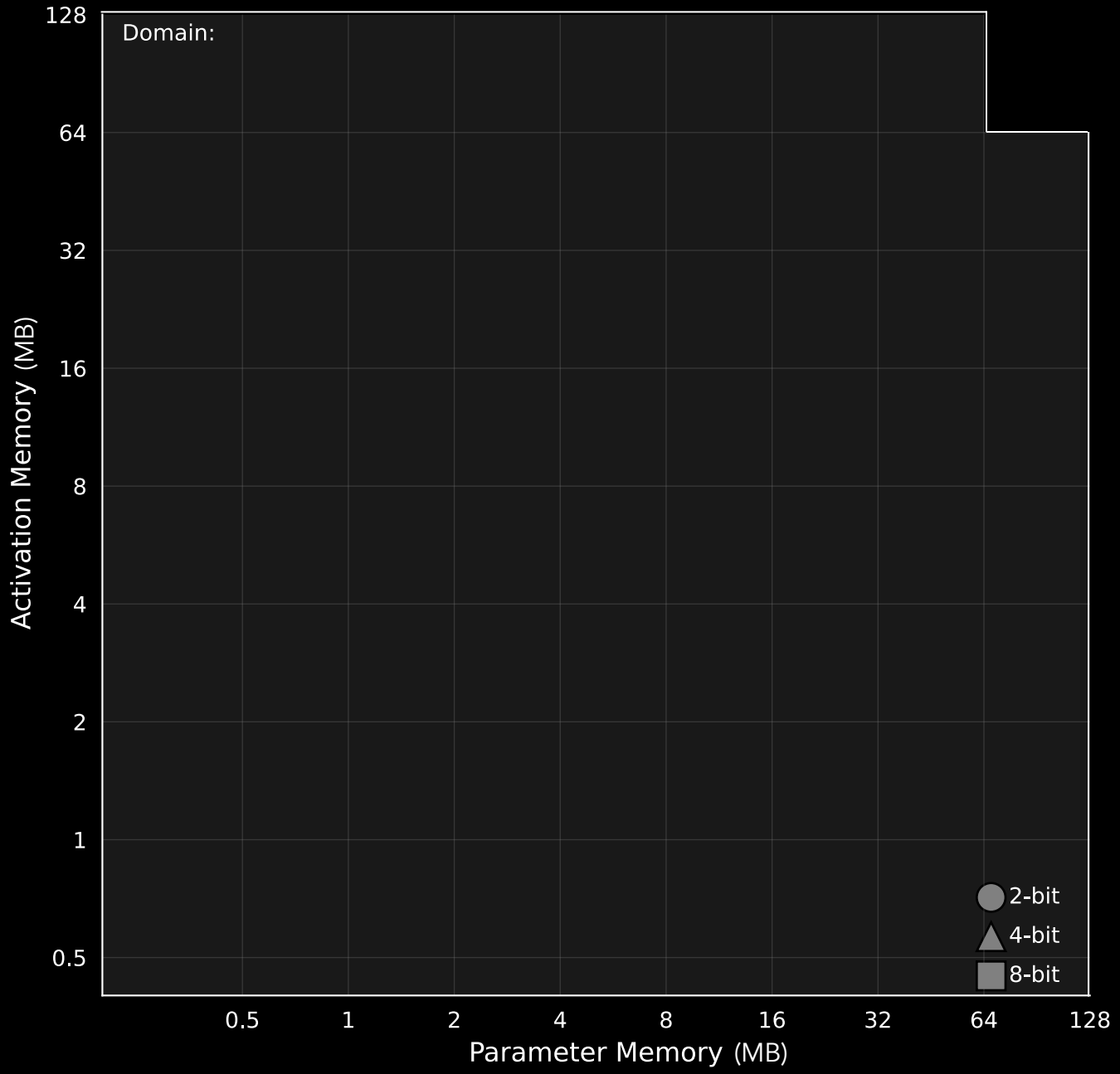
Metric Computation Cost

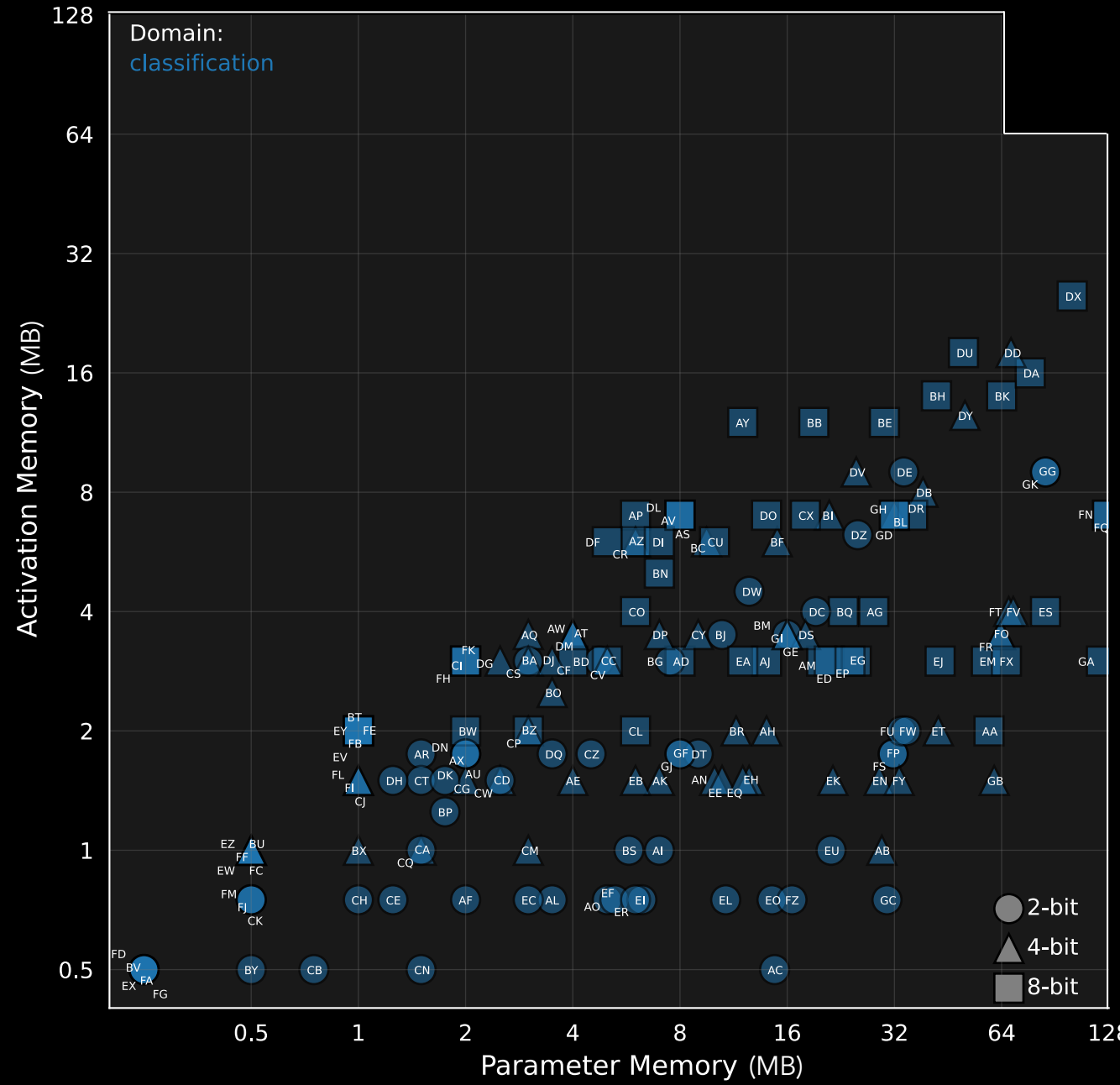
Method	ResNet-50	PSPNet
EAGL (Ours)	**3.15 CPU seconds**	**<1 CPU minute**
ALPS (Ours)	166 GPU hours	67 GPU hours
HAWQ-v3	2 GPU hours	1032 GPU hours

Low Precision ≠ Low fidelity

A Growing List of Implementable Networks

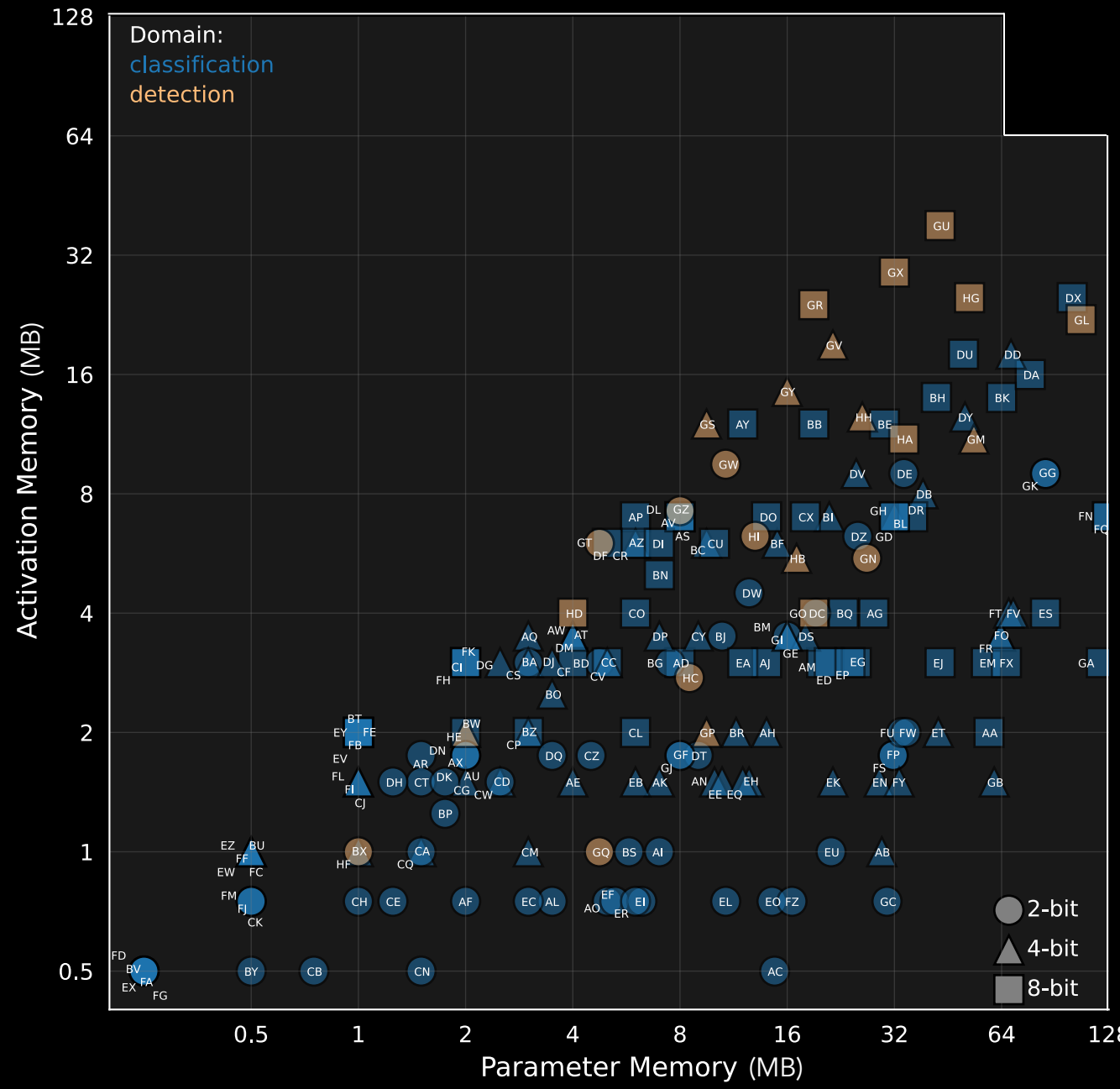
Networks:





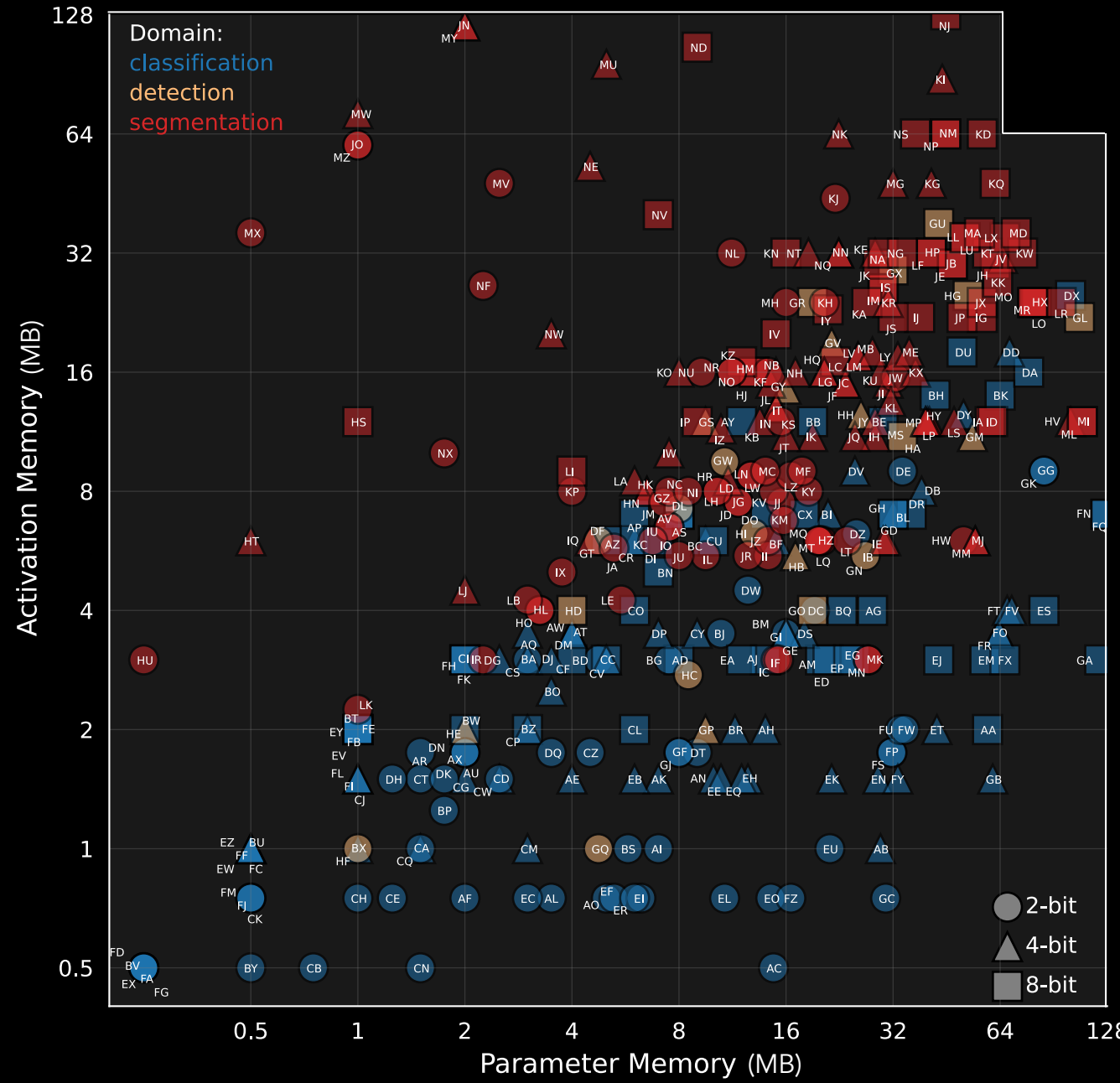
Networks:

classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-4b FU. vgg16_bn-2b FV. vgg19_bn-4b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b



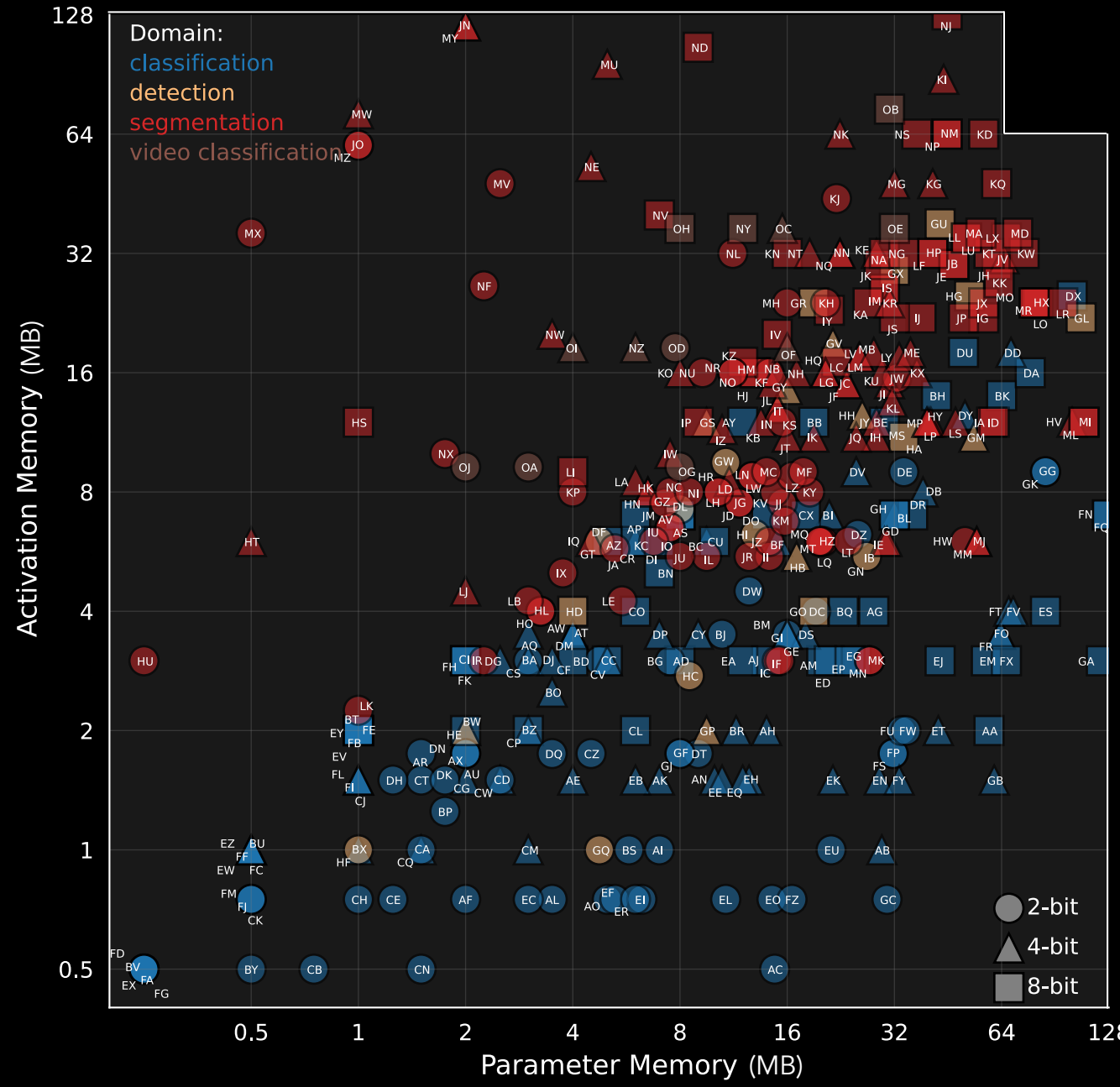
Networks:

classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-4b FU. vgg16_bn-2b FV. vgg19_bn-4b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b detection : GL. Faster-RCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b GU. maskrcnn_resnet50_fpn-8b GV. maskrcnn_resnet50_fpn-4b GW. maskrcnn_resnet50_fpn-2b GX. RetinaNet_r50_fpn-8b GY. RetinaNet_r50_fpn-4b GZ. RetinaNet_r50_fpn-2b HA. SSD-VGG-8b HB. SSD-VGG-4b HC. SSD-VGG-2b HD. ssdlite_mobilent_v3-8b HE. ssdlite_mobilent_v3-4b HF. ssdlite_mobilent_v3-2b HG. YOLOv4-8b HH. YOLOv4-4b HI. YOLOv4-2b



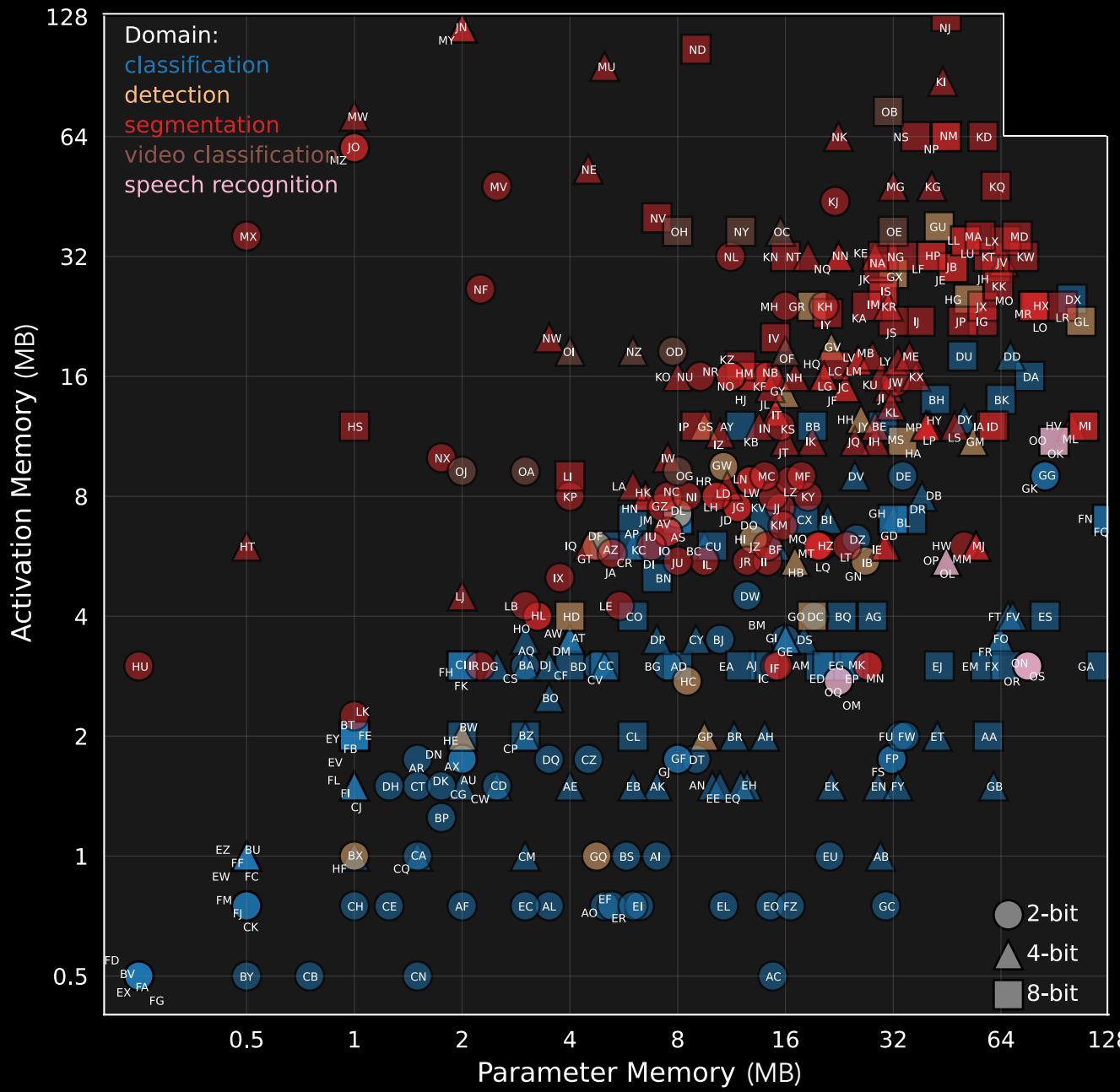
Networks:

classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-4b FU. vgg16_bn-2b FV. vgg19_bn-4b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b detection : GL. Faster-RCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b GU. maskrcnn_resnet50_fpn-8b GV. maskrcnn_resnet50_fpn-4b GW. maskrcnn_resnet50_fpn-2b GX. RetinaNet_r50_fpn-8b GY. RetinaNet_r50_fpn-4b GZ. RetinaNet_r50_fpn-2b HA. SSD-VGG-8b HB. SSD-VGG-4b HC. SSD-VGG-2b HD. ssdlite_mobilent_v3-8b HE. ssdlite_mobilent_v3-4b HF. ssdlite_mobilent_v3-2b HG. YOLOv4-8b HH. YOLOv4-4b HI. YOLOv4-2b segmentation: HJ. BiSeNet-8b HK. BiSeNet-4b HL. BiSeNet-2b HM. BiSeNet_Resnet18-8b HN. BiSeNet_Resnet18-4b HO. BiSeNet_Resnet18-2b HP. CoarseLinkNet50-8b HQ. CoarseLinkNet50-4b HR. CoarseLinkNet50-2b HS. DABNet-8b HT. DABNet-4b HU. DABNet-2b HV. DeepLabv2_ASPP-4b HW. DeepLabv2_ASPP-2b HX. DeepLabv2_FOV-8b HY. DeepLabv2_FOV-4b HZ. DeepLabv2_FOV-2b IA. DeepLabv3-8b IB. DeepLabv3-4b IC. DeepLabv3-2b ID. DeepLabv3_plus-8b IE. DeepLabv3_plus-4b IF. DeepLabv3_plus-2b IG. deeplabv3_resnt101-8b IH. deeplabv3_resnt101-4b II. deeplabv3_resnt101-2b IJ. deeplabv3_resnet50-8b IK. deeplabv3_resnet50-4b IL. deeplabv3_resnet50-2b IM. DenseASPP-8b IN. DenseASPP-4b IO. DenseASPP-2b IP. DenseASPP_121-8b IQ. DenseASPP_121-4b IR. DenseASPP_121-2b IS. DenseASPP_161-8b IT. DenseASPP_161-4b IU. DenseASPP_161-2b IV. DenseASPP_169-8b IW. DenseASPP_169-4b IX. DenseASPP_169-2b IY. DenseASPP_201-8b IZ. DenseASPP_201-4b JA. DenseASPP_201-2b JB. DUNet-8b JC. DUNet-4b JD. DUNet-2b JE. DUNet_Resnet101-8b JF. DUNet_Resnet101-4b JG. DUNet_Resnet101-2b JH. DUNet_Resnet152-8b JI. DUNet_Resnet152-4b JJ. DUNet_Resnet152-2b JK. DUNet_Resnet50-8b JL. DUNet_Resnet50-4b JM. DUNet_Resnet50-2b JN. FCDenseNet-4b JO. FCDenseNet-2b JP. fcn_resnet101-8b JQ. fcn_resnet101-4b JR. fcn_resnet101-2b JS. fcn_resnet50-8b JT. fcn_resnet50-4b JU. fcn_resnet50-2b JV. FCN32VGG-4b JW. FCN32VGG-2b JX. GCN-8b JY. GCN-4b JZ. GCN-2b KA. GCN_Densenet-8b KB. GCN_Densenet-4b KC. GCN_Densenet-2b KD. GCN_PSP-8b KE. GCN_PSP-4b KF. GCN_PSP-2b KG. GCN_Resnext-4b KH. GCN_Resnext-2b KI. GCNFuse-4b KJ. GCNFuse-2b KK. HRNetv2-8b KL. HRNetv2-4b KM. HRNetv2-2b KN. LinkDenseNet121-8b KO. LinkDenseNet121-4b KP. LinkDenseNet121-2b KQ. LinkDenseNet161-8b KR. LinkDenseNet161-4b KS. LinkDenseNet161-2b KT. LinkNet101-8b KU. LinkNet101-4b KV. LinkNet101-2b KW. LinkNet152-8b KX. LinkNet152-4b KY. LinkNet152-2b KZ. LinkNet18-8b LA. LinkNet18-4b LB. LinkNet18-2b LC. LinkNet34-8b LD. LinkNet34-4b LE. LinkNet34-2b LF. LinkNet50-8b LG. LinkNet50-4b LH. LinkNet50-2b LI. lraspp_mobilenet_v3_large-8b LJ. lraspp_mobilenet_v3_large-4b LK. lraspp_mobilenet_v3_large-2b LL. OCNet-8b LM. OCNet-4b LN. OCNet-2b LO. OCNet_ASP_Resnet101-8b LP. OCNet_ASP_Resnet101-4b LQ. OCNet_ASP_Resnet101-2b LR. OCNet_ASP_Resnet152-8b LS. OCNet_ASP_Resnet152-4b LT. OCNet_ASP_Resnet152-2b LU. OCNet_Base_Resnet101-8b LV. OCNet_Base_Resnet101-4b LW. OCNet_Base_Resnet101-2b LX. OCNet_Base_Resnet152-8b LY. OCNet_Base_Resnet152-4b LZ. OCNet_Base_Resnet152-2b MA. OCNet_Pyramid_Resnet101-8b MB. OCNet_Pyramid_Resnet101-4b MC. OCNet_Pyramid_Resnet101-2b MD. OCNet_Pyramid_Resnet152-8b ME. OCNet_Pyramid_Resnet152-4b MF. OCNet_Pyramid_Resnet152-2b MG. PSPNet-4b MH. PSPNet-2b MI. RefineNet4Cascade-8b MJ. RefineNet4Cascade-4b MK. RefineNet4Cascade-2b ML. RefineNet4CascadePoolingImproved-8b MM. RefineNet4CascadePoolingImproved-4b MN. RefineNet4CascadePoolingImproved-2b MO. ResNetDUC-8b MP. ResNetDUC-4b MQ. ResNetDUC-2b MR. ResNetDUCHDC-8b MS. ResNetDUCHDC-4b MT. ResNetDUCHDC-2b MU. Tiramisu103-4b MV. Tiramisu103-2b MW. Tiramisu57-4b MX. Tiramisu57-2b MY. Tiramisu67-4b MZ. Tiramisu67-2b NA. UNet-8b NB. UNet-4b NC. UNet-2b ND. Unet_Plus_Plus-8b NE. Unet_Plus_Plus-4b NF. Unet_Plus_Plus-2b NG. UNet1024-8b NH. UNet1024-4b NI. UNet1024-2b NJ. UNet128-8b NK. UNet128-4b NL. UNet128-2b NM. UNet256-8b NN. UNet256-4b NO. UNet256-2b NP. UNet512-8b NQ. UNet512-4b NR. UNet512-2b NS. UNet960-8b NT. UNet960-4b NU. UNet960-2b NV. UNetDilated-8b NW. UNetDilated-4b NX. UNetDilated-2b



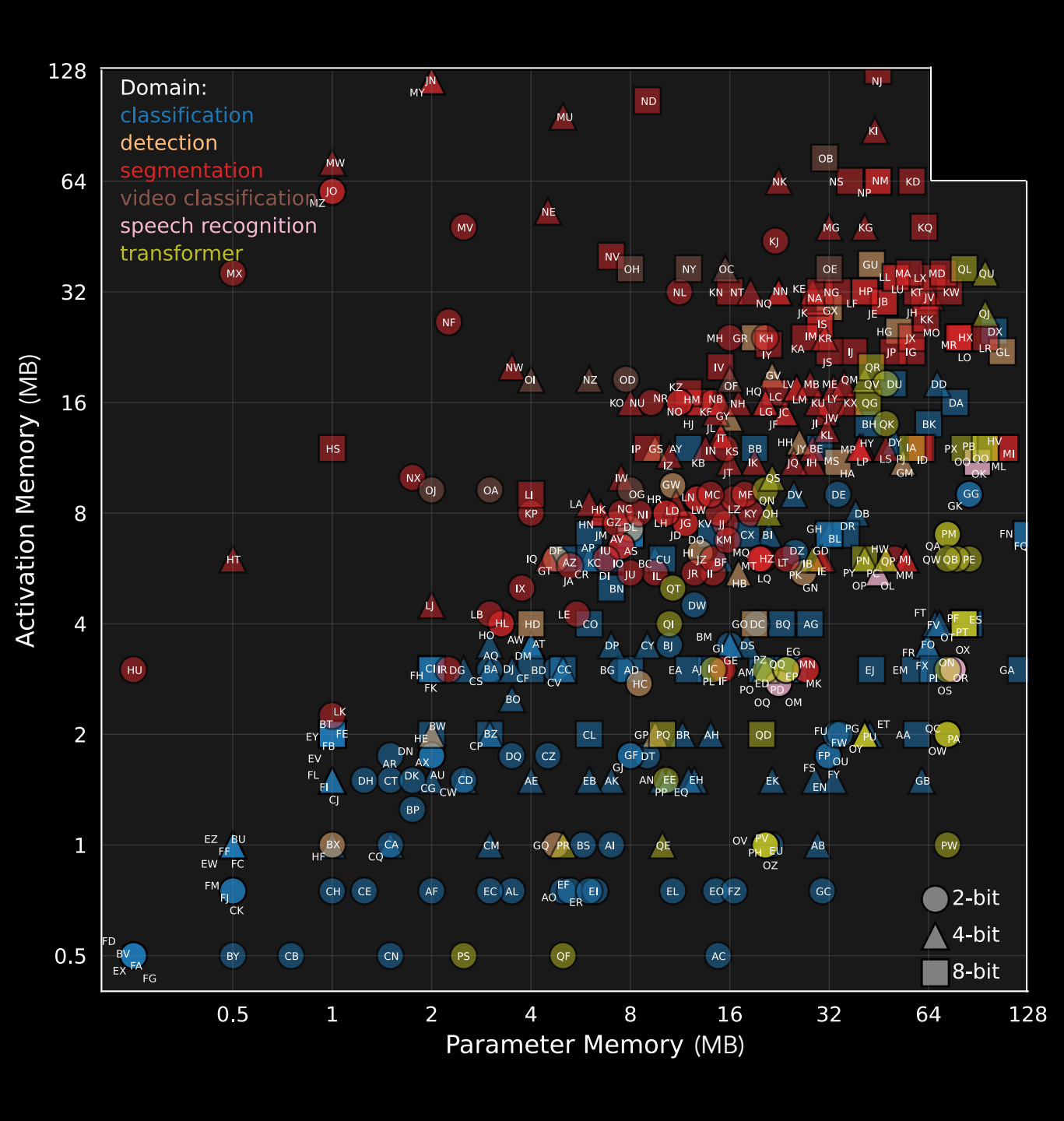
Networks:

classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-4b FU. vgg16_bn-2b FV. vgg19_bn-4b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b detection : GL. Faster-RCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b GU. maskrcnn_resnet50_fpn-8b GV. maskrcnn_resnet50_fpn-4b GW. maskrcnn_resnet50_fpn-2b GX. RetinaNet_r50_fpn-8b GY. RetinaNet_r50_fpn-4b GZ. RetinaNet_r50_fpn-2b HA. SSD-VGG-8b HB. SSD-VGG-4b HC. SSD-VGG-2b HD. ssdlite_mobilent_v3-8b HE. ssdlite_mobilent_v3-4b HF. ssdlite_mobilent_v3-2b HG. YOLOv4-8b HH. YOLOv4-4b HI. YOLOv4-2b segmentation: HJ. BiSeNet-8b HK. BiSeNet-4b HL. BiSeNet-2b HM. BiSeNet_Resnet18-8b HN. BiSeNet_Resnet18-4b HO. BiSeNet_Resnet18-2b HP. CoarseLinkNet50-8b HQ. CoarseLinkNet50-4b HR. CoarseLinkNet50-2b HS. DABNet-8b HT. DABNet-4b HU. DABNet-2b HV. DeepLabv2_ASPP-4b HW. DeepLabv2_ASPP-2b HX. DeepLabv2_FOV-8b HY. DeepLabv2_FOV-4b HZ. DeepLabv2_FOV-2b IA. DeepLabv3-8b IB. DeepLabv3-4b IC. DeepLabv3-2b ID. DeepLabv3_plus-8b IE. DeepLabv3_plus-4b IF. DeepLabv3_plus-2b IG. deeplabv3_resnt101-8b IH. deeplabv3_resnt101-4b II. deeplabv3_resnt101-2b IJ. deeplabv3_resnet50-8b IK. deeplabv3_resnet50-4b IL. deeplabv3_resnet50-2b IM. DenseASPP-8b IN. DenseASPP-4b IO. DenseASPP-2b IP. DenseASPP_121-8b IQ. DenseASPP_121-4b IR. DenseASPP_121-2b IS. DenseASPP_161-8b IT. DenseASPP_161-4b IU. DenseASPP_161-2b IV. DenseASPP_169-8b IW. DenseASPP_169-4b IX. DenseASPP_169-2b IY. DenseASPP_201-8b IZ. DenseASPP_201-4b JA. DenseASPP_201-2b JB. DUNet-8b JC. DUNet-4b JD. DUNet-2b JE. DUNet_Resnet101-8b JF. DUNet_Resnet101-4b JG. DUNet_Resnet101-2b JH. DUNet_Resnet152-8b JI. DUNet_Resnet152-4b JJ. DUNet_Resnet152-2b JK. DUNet_Resnet50-8b JL. DUNet_Resnet50-4b JM. DUNet_Resnet50-2b JN. FCDenseNet-4b JO. FCDenseNet-2b JP. fcn_resnet101-8b JQ. fcn_resnet101-4b JR. fcn_resnet101-2b JS. fcn_resnet50-8b JT. fcn_resnet50-4b JU. fcn_resnet50-2b JV. FCN32VGG-4b JW. FCN32VGG-2b JX. GCN-8b JY. GCN-4b JZ. GCN-2b KA. GCN_Densenet-8b KB. GCN_Densenet-4b KC. GCN_Densenet-2b KD. GCN_PSP-8b KE. GCN_PSP-4b KF. GCN_PSP-2b KG. GCN_Resnext-4b KH. GCN_Resnext-2b KI. GCNFuse-4b KJ. GCNFuse-2b KK. HRNetv2-8b KL. HRNetv2-4b KM. HRNetv2-2b KN. LinkDenseNet121-8b KO. LinkDenseNet121-4b KP. LinkDenseNet121-2b KQ. LinkDenseNet161-8b KR. LinkDenseNet161-4b KS. LinkDenseNet161-2b KT. LinkNet101-8b KU. LinkNet101-4b KV. LinkNet101-2b KW. LinkNet152-8b KX. LinkNet152-4b KY. LinkNet152-2b KZ. LinkNet18-8b LA. LinkNet18-4b LB. LinkNet18-2b LC. LinkNet34-8b LD. LinkNet34-4b LE. LinkNet34-2b LF. LinkNet50-8b LG. LinkNet50-4b LH. LinkNet50-2b LI. lraspp_mobilenet_v3_large-8b LJ. lraspp_mobilenet_v3_large-4b LK. lraspp_mobilenet_v3_large-2b LL. OCNet-8b LM. OCNet-4b LN. OCNet-2b LO. OCNet_ASP_Resnet101-8b LP. OCNet_ASP_Resnet101-4b LQ. OCNet_ASP_Resnet101-2b LR. OCNet_ASP_Resnet152-8b LS. OCNet_ASP_Resnet152-4b LT. OCNet_ASP_Resnet152-2b LU. OCNet_Base_Resnet101-8b LV. OCNet_Base_Resnet101-4b LW. OCNet_Base_Resnet101-2b LX. OCNet_Base_Resnet152-8b LY. OCNet_Base_Resnet152-4b LZ. OCNet_Base_Resnet152-2b MA. OCNet_Pyramid_Resnet101-8b MB. OCNet_Pyramid_Resnet101-4b MC. OCNet_Pyramid_Resnet101-2b MD. OCNet_Pyramid_Resnet152-8b ME. OCNet_Pyramid_Resnet152-4b MF. OCNet_Pyramid_Resnet152-2b MG. PSPNet-4b MH. PSPNet-2b MI. RefineNet4Cascade-8b MJ. RefineNet4Cascade-4b MK. RefineNet4Cascade-2b ML. RefineNet4CascadePoolingImproved-8b MM. RefineNet4CascadePoolingImproved-4b MN. RefineNet4CascadePoolingImproved-2b MO. ResNetDUC-8b MP. ResNetDUC-4b MQ. ResNetDUC-2b MR. ResNetDUCHDC-8b MS. ResNetDUCHDC-4b MT. ResNetDUCHDC-2b MU. Tiramisu103-4b MV. Tiramisu103-2b MW. Tiramisu57-4b MX. Tiramisu57-2b MY. Tiramisu67-4b MZ. Tiramisu67-2b NA. UNet-8b NB. UNet-4b NC. UNet-2b ND. Unet_Plus_Plus-8b NE. Unet_Plus_Plus-4b NF. Unet_Plus_Plus-2b NG. UNet1024-8b NH. UNet1024-4b NI. UNet1024-2b NJ. UNet128-8b NK. UNet128-4b NL. UNet128-2b NM. UNet256-8b NN. UNet256-4b NO. UNet256-2b NP. UNet512-8b NQ. UNet512-4b NR. UNet512-2b NS. UNet960-8b NT. UNet960-4b NU. UNet960-2b NV. UNetDilated-8b NW. UNetDilated-4b NX. UNetDilated-2b video classification: NY. mc3_18-8b NZ. mc3_18-4b OA. mc3_18-2b OB. r2plus1d_18-8b OC. r2plus1d_18-4b OD. r2plus1d_18-2b OE. r3d_18-8b OF. r3d_18-4b OG. r3d_18-2b OH. s3d-8b OI. s3d-4b OJ. s3d-2b



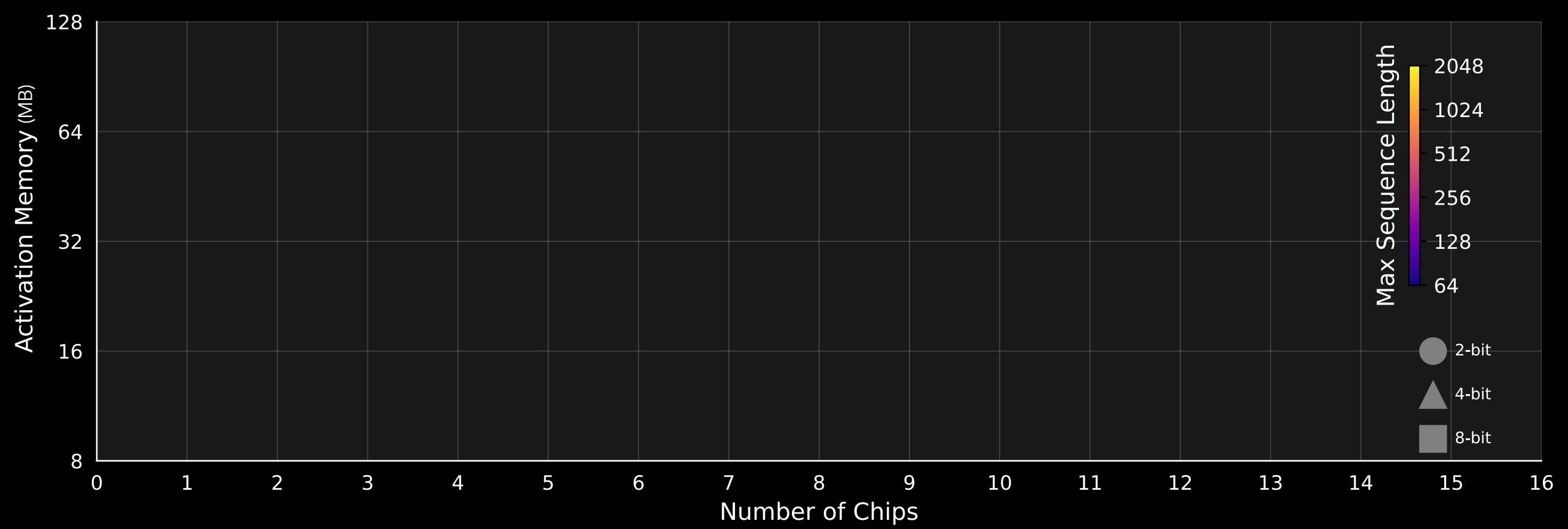
Networks:

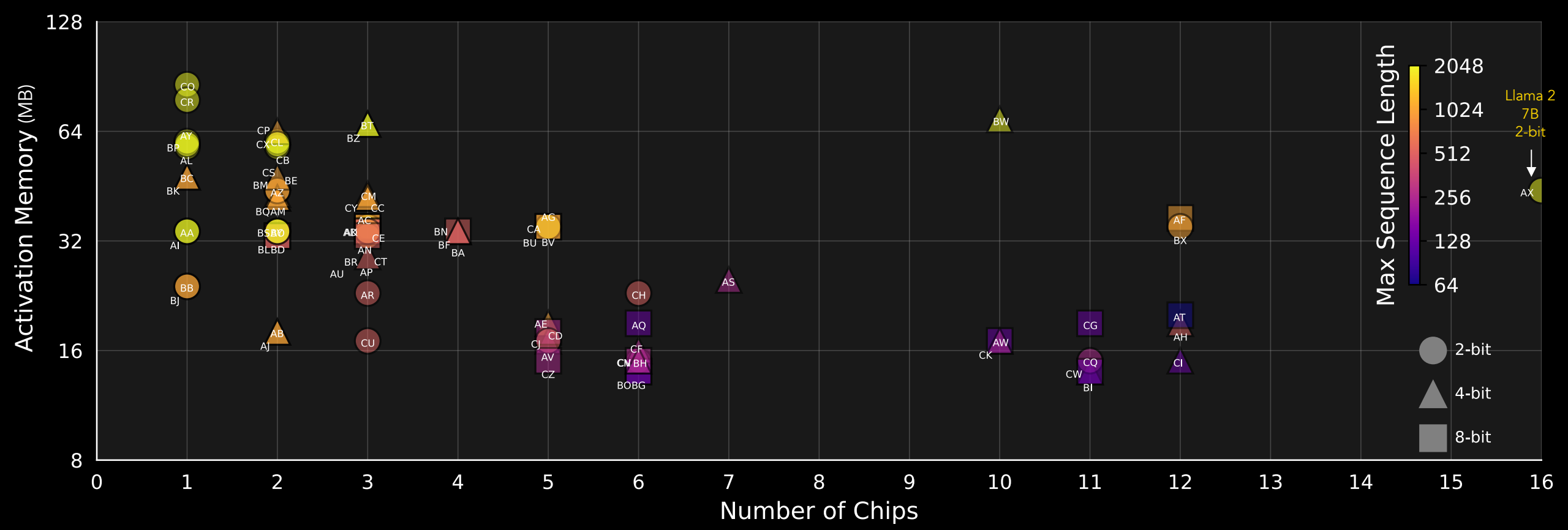
classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-4b FU. vgg16_bn-2b FV. vgg19_bn-4b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b detection: GL. Faster-RCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b GU. maskrcnn_resnet50_fpn-8b GV. maskrcnn_resnet50_fpn-4b GW. maskrcnn_resnet50_fpn-2b GX. RetinaNet_r50_fpn-8b GY. RetinaNet_r50_fpn-4b GZ. RetinaNet_r50_fpn-2b HA. SSD-VGG-8b HB. SSD-VGG-4b HC. SSD-VGG-2b HD. ssdlite_mobilent_v3-8b HE. ssdlite_mobilent_v3-4b HF. ssdlite_mobilent_v3-2b HG. YOLOv4-8b HH. YOLOv4-4b HI. YOLOv4-2b segmentation: HJ. BiSeNet-8b HK. BiSeNet-4b HL. BiSeNet-2b HM. BiSeNet_Resnet18-8b HN. BiSeNet_Resnet18-4b HO. BiSeNet_Resnet18-2b HP. CoarseLinkNet50-8b HQ. CoarseLinkNet50-4b HR. CoarseLinkNet50-2b HS. DABNet-8b HT. DABNet-4b HU. DABNet-2b HV. DeepLabv2_ASPP-4b HW. DeepLabv2_ASPP-2b HX. DeepLabv2_FOV-8b HY. DeepLabv2_FOV-4b HZ. DeepLabv2_FOV-2b IA. DeepLabv3-8b IB. DeepLabv3-4b IC. DeepLabv3-2b ID. DeepLabv3_plus-8b IE. DeepLabv3_plus-4b IF. DeepLabv3_plus-2b IG. deeplabv3_resnt101-8b IH. deeplabv3_resnt101-4b II. deeplabv3_resnt101-2b IJ. deeplabv3_resnet50-8b IK. deeplabv3_resnet50-4b IL. deeplabv3_resnet50-2b IM. DenseASPP-8b IN. DenseASPP-4b IO. DenseASPP-2b IP. DenseASPP_121-8b IQ. DenseASPP_121-4b IR. DenseASPP_121-2b IS. DenseASPP_161-8b IT. DenseASPP_161-4b IU. DenseASPP_161-2b IV. DenseASPP_169-8b IW. DenseASPP_169-4b IX. DenseASPP_169-2b IY. DenseASPP_201-8b IZ. DenseASPP_201-4b JA. DenseASPP_201-2b JB. DUNet-8b JC. DUNet-4b JD. DUNet-2b JE. DUNet_Resnet101-8b JF. DUNet_Resnet101-4b JG. DUNet_Resnet101-2b JH. DUNet_Resnet152-8b JI. DUNet_Resnet152-4b JJ. DUNet_Resnet152-2b JK. DUNet_Resnet50-8b JL. DUNet_Resnet50-4b JM. DUNet_Resnet50-2b JN. FCDenseNet-4b JO. FCDenseNet-2b JP. fcn_resnet101-8b JQ. fcn_resnet101-4b JR. fcn_resnet101-2b JS. fcn_resnet50-8b JT. fcn_resnet50-4b JU. fcn_resnet50-2b JV. FCN32VGG-4b JW. FCN32VGG-2b JX. GCN-8b JY. GCN-4b JZ. GCN-2b KA. GCN_Densenet-8b KB. GCN_Densenet-4b KC. GCN_Densenet-2b KD. GCN_PSP-8b KE. GCN_PSP-4b KF. GCN_PSP-2b KG. GCN_Resnext-4b KH. GCN_Resnext-2b KI. GCNFuse-4b KJ. GCNFuse-2b KK. HRNetv2-8b KL. HRNetv2-4b KM. HRNetv2-2b KN. LinkDenseNet121-8b KO. LinkDenseNet121-4b KP. LinkDenseNet121-2b KQ. LinkDenseNet161-8b KR. LinkDenseNet161-4b KS. LinkDenseNet161-2b KT. LinkNet101-8b KU. LinkNet101-4b KV. LinkNet101-2b KW. LinkNet152-8b KX. LinkNet152-4b KY. LinkNet152-2b KZ. LinkNet18-8b LA. LinkNet18-4b LB. LinkNet18-2b LC. LinkNet34-8b LD. LinkNet34-4b LE. LinkNet34-2b LF. LinkNet50-8b LG. LinkNet50-4b LH. LinkNet50-2b LI. lraspp_mobilenet_v3_large-8b LJ. lraspp_mobilenet_v3_large-4b LK. lraspp_mobilenet_v3_large-2b LL. OCNet-8b LM. OCNet-4b LN. OCNet-2b LO. OCNet_ASP_Resnet101-8b LP. OCNet_ASP_Resnet101-4b LQ. OCNet_ASP_Resnet101-2b LR. OCNet_ASP_Resnet152-8b LS. OCNet_ASP_Resnet152-4b LT. OCNet_ASP_Resnet152-2b LU. OCNet_Base_Resnet101-8b LV. OCNet_Base_Resnet101-4b LW. OCNet_Base_Resnet101-2b LX. OCNet_Base_Resnet152-8b LY. OCNet_Base_Resnet152-4b LZ. OCNet_Base_Resnet152-2b MA. OCNet_Pyramid_Resnet101-8b MB. OCNet_Pyramid_Resnet101-4b MC. OCNet_Pyramid_Resnet101-2b MD. OCNet_Pyramid_Resnet152-8b ME. OCNet_Pyramid_Resnet152-4b MF. OCNet_Pyramid_Resnet152-2b MG. PSPNet-4b MH. PSPNet-2b MI. RefineNet4Cascade-8b MJ. RefineNet4Cascade-4b MK. RefineNet4Cascade-2b ML. RefineNet4CascadePoolingImproved-8b MM. RefineNet4CascadePoolingImproved-4b MN. RefineNet4CascadePoolingImproved-2b MO. ResNetDUC-8b MP. ResNetDUC-4b MQ. ResNetDUC-2b MR. ResNetDUCHDC-8b MS. ResNetDUCHDC-4b MT. ResNetDUCHDC-2b MU. Tiramisu103-4b MV. Tiramisu103-2b MW. Tiramisu57-4b MX. Tiramisu57-2b MY. Tiramisu67-4b MZ. Tiramisu67-2b NA. UNet-8b NB. UNet-4b NC. UNet-2b ND. Unet_Plus_Plus-8b NE. Unet_Plus_Plus-4b NF. Unet_Plus_Plus-2b NG. UNet1024-8b NH. UNet1024-4b NI. UNet1024-2b NJ. UNet128-8b NK. UNet128-4b NL. UNet128-2b NM. UNet256-8b NN. UNet256-4b NO. UNet256-2b NP. UNet512-8b NQ. UNet512-4b NR. UNet512-2b NS. UNet960-8b NT. UNet960-4b NU. UNet960-2b NV. UNetDilated-8b NW. UNetDilated-4b NX. UNetDilated-2b video classification: NY. mc3_18-8b NZ. mc3_18-4b OA. mc3_18-2b OB. r2plus1d_18-8b OC. r2plus1d_18-4b OD. r2plus1d_18-2b OE. r3d_18-8b OF. r3d_18-4b OG. r3d_18-2b OH. s3d-8b OI. s3d-4b OJ. s3d-2b speech recognition: OK. hubert_base-8b OL. hubert_base-4b OM. hubert_base-2b ON. hubert_large-2b OO. wave2vec2_base-8b OP. wave2vec2_base-4b OQ. wave2vec2_base-2b OR. wav2vec2_large-2b OS. wav2vec2_large_lv60k-2b



Networks:

classification: AA. alexnet-8b AB. alexnet-4b AC. alexnet-2b AD. densenet121-8b AE. densenet121-4b AF. densenet121-2b AG. densenet161-8b AH. densenet161-4b AI. densenet161-2b AJ. densenet169-8b AK. densenet169-4b AL. densenet169-2b AM. densenet201-8b AN. densenet201-4b AO. densenet201-2b AP. efficientnet_b0-8b AQ. efficientnet_b0-4b AR. efficientnet_b0-2b AS. efficientnet_b1-8b AT. efficientnet_b1-4b AU. efficientnet_b1-2b AV. efficientnet_b1-8b AW. efficientnet_b1-4b AX. efficientnet_b1-2b AY. efficientnet_b3-8b AZ. efficientnet_b3-4b BA. efficientnet_b3-2b BB. efficientnet_b4-8b BC. efficientnet_b4-4b BD. efficientnet_b4-2b BE. efficientnet_b5-8b BF. efficientnet_b5-4b BG. efficientnet_b5-2b BH. efficientnet_b6-8b BI. efficientnet_b6-4b BJ. efficientnet_b6-2b BK. efficientnet_b7-8b BL. efficientnet_b7-4b BM. efficientnet_b7-2b BN. googlenet-8b BO. googlenet-4b BP. googlenet-2b BQ. inception_v3-8b BR. inception_v3-4b BS. inception_v3-2b BT. mnasnet0_5-8b BU. mnasnet0_5-4b BV. mnasnet0_5-2b BW. mnasnet0_75-8b BX. mnasnet0_75-4b BY. mnasnet0_75-2b BZ. mnasnet1_0-8b CA. mnasnet1_0-4b CB. mnasnet1_0-2b CC. mnasnet1_3-8b CD. mnasnet1_3-4b CE. mnasnet1_3-2b CF. mobilenet_v2-8b CG. mobilenet_v2-4b CH. mobilenet_v2-2b CI. mobilenet_v3_small-8b CJ. mobilenet_v3_small-4b CK. mobilenet_v3_small-2b CL. mobilenet_v3_large-8b CM. mobilenet_v3_large-4b CN. mobilenet_v3_large-2b CO. regnet_y_400mf-8b CP. regnet_y_400mf-4b CQ. regnet_y_400mf-2b CR. regnet_y_800mf-8b CS. regnet_y_800mf-4b CT. regnet_y_800mf-2b CU. regnet_y_1_6gf-8b CV. regnet_y_1_6gf-4b CW. regnet_y_1_6gf-2b CX. regnet_y_3_2gf-8b CY. regnet_y_3_2gf-4b CZ. regnet_y_3_2gf-2b DA. regnet_y_16gf-8b DB. regnet_y_16gf-4b DC. regnet_y_16gf-2b DD. regnet_y_32gf-4b DE. regnet_y_32gf-2b DF. regnet_x_400mf-8b DG. regnet_x_400mf-4b DH. regnet_x_400mf-2b DI. regnet_x_800mf-8b DJ. regnet_x_800mf-4b DK. regnet_x_800mf-2b DL. regnet_x_1_6gf-8b DM. regnet_x_1_6gf-4b DN. regnet_x_1_6gf-2b DO. regnet_x_3_2gf-8b DP. regnet_x_3_2gf-4b DQ. regnet_x_3_2gf-2b DR. regnet_x_8gf-8b DS. regnet_x_8gf-4b DT. regnet_x_8gf-2b DU. regnet_x_16gf-8b DV. regnet_x_16gf-4b DW. regnet_x_16gf-2b DX. regnet_x_32gf-8b DY. regnet_x_32gf-4b DZ. regnet_x_32gf-2b EA. resnet18-8b EB. resnet18-4b EC. resnet18-2b ED. resnet34-8b EE. resnet34-4b EF. resnet34-2b EG. resnet50-8b EH. resnet50-4b EI. resnet50-2b EJ. resnet101-8b EK. resnet101-4b EL. resnet101-2b EM. resnet152-8b EN. resnet152-4b EO. resnet152-2b EP. resnext50_32x4d-8b EQ. resnext50_32x4d-4b ER. resnext50_32x4d-2b ES. resnext101_32x8d-8b ET. resnext101_32x8d-4b EU. resnext101_32x8d-2b EV. shufflenet_v2_x0_5-8b EW. shufflenet_v2_x0_5-4b EX. shufflenet_v2_x0_5-2b EY. shufflenet_v2_x1_0-8b EZ. shufflenet_v2_x1_0-4b FA. shufflenet_v2_x1_0-2b FB. shufflenet_v2_x1_5-8b FC. shufflenet_v2_x1_5-4b FD. shufflenet_v2_x1_5-2b FE. shufflenet_v2_x2_0-8b FF. shufflenet_v2_x2_0-4b FG. shufflenet_v2_x2_0-2b FH. squeezenet1_0-8b FI. squeezenet1_0-4b FJ. squeezenet1_0-2b FK. squeezenet1_1-8b FL. squeezenet1_1-4b FM. squeezenet1_1-2b FN. vgg11_bn-8b FO. vgg11_bn-4b FP. vgg11_bn-2b FQ. vgg13_bn-8b FR. vgg13_bn-4b FS. vgg13_bn-2b FT. vgg16_bn-8b FU. vgg16_bn-4b FV. vgg19_bn-8b FW. vgg19_bn-2b FX. wide_resnet50_2-8b FY. wide_resnet50_2-4b FZ. wide_resnet50_2-2b GA. wide_resnet101_2-8b GB. wide_resnet101_2-4b GC. wide_resnet101_2-2b GD. vit_l_16-8b GE. vit_l_16-4b GF. vit_l_16-2b GG. vit_l_32-2b GH. vit_b_16-8b GI. vit_b_16-4b GJ. vit_b_16-2b GK. vit_b_32-2b detection: GL. FasterRCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b GU. maskrcnn_resnet50_fpn-8b GV. maskrcnn_resnet50_fpn-4b GW. maskrcnn_resnet50_fpn-2b GX. RetinaNet-r50_fpn-8b GY. RetinaNet-r50_fpn-4b GZ. RetinaNet-r50_fpn-2b HA. SSD-VGG-8b HB. SSD-VGG-4b HC. SSD-VGG-2b HD. ssdlite_mobilent_v3-8b HE. ssdlite_mobilent_v3-4b HF. ssdlite_mobilent_v3-2b HG. YOLOv4-8b HH. YOLOv4-4b HI. YOLOv4-2b segmentation: HJ. BiSeNet-8b HK. BiSeNet-4b HL. BiSeNet-2b HM. BiSeNet_Resnet18-8b HN. BiSeNet_Resnet18-4b HO. BiSeNet_Resnet18-2b HP. CoarseLinkNet50-8b HQ. CoarseLinkNet50-4b HR. CoarseLinkNet50-2b HS. DABNet-8b HT. DABNet-4b HU. DABNet-2b HV. DeepLabv2_ASPP-4b HW. DeepLabv2_ASPP-2b HX. DeepLabv2_FOV-8b HY. DeepLabv2_FOV-4b HZ. DeepLabv2_FOV-2b IA. DeepLabv3-8b IB. DeepLabv3-4b IC. DeepLabv3-2b ID. DeepLabv3_plus-8b IE. DeepLabv3_plus-4b IF. DeepLabv3_plus-2b IG. deeplabv3_resnt101-8b IH. deeplabv3_resnt101-4b II. deeplabv3_resnt101-2b IJ. deeplabv3_resnet50-8b IK. deeplabv3_resnet50-4b IL. deeplabv3_resnet50-2b IM. DenseASPP-8b IN. DenseASPP-4b IO. DenseASPP-2b IP. DenseASPP_121-8b IQ. DenseASPP_121-4b IR. DenseASPP_121-2b IS. DenseASPP_161-8b IT. DenseASPP_161-4b IU. DenseASPP_161-2b IV. DenseASPP_169-8b IW. DenseASPP_169-4b IX. DenseASPP_169-2b IY. DenseASPP_201-8b IZ. DenseASPP_201-4b JA. DenseASPP_201-2b JB. DUNet-8b JC. DUNet-4b JD. DUNet-2b JE. DUNet_Resnet101-8b JF. DUNet_Resnet101-4b JG. DUNet_Resnet101-2b JH. DUNet_Resnet152-8b JI. DUNet_Resnet152-4b JJ. DUNet_Resnet152-2b JK. DUNet_Resnet50-8b JL. DUNet_Resnet50-4b JM. DUNet_Resnet50-2b JN. FCDenseNet-4b JO. FCDenseNet-2b JP. fcn_resnet101-8b JQ. fcn_resnet101-4b JR. fcn_resnet101-2b JS. fcn_resnet50-8b JT. fcn_resnet50-4b JU. fcn_resnet50-2b JV. FCN32VGG-4b JW. FCN32VGG-2b JX. GCN-8b JY. GCN-4b JZ. GCN-2b KA. GCN_Densenet-8b KB. GCN_Densenet-4b KC. GCN_Densenet-2b KD. GCN_PSP-8b KE. GCN_PSP-4b KF. GCN_PSP-2b KG. GCN_Resnext-4b KH. GCN_Resnext-2b KI. GCNFuse-4b KJ. GCNFuse-2b KK. HRNetv2-8b KL. HRNetv2-4b KM. HRNetv2-2b KN. LinkDenseNet121-8b KO. LinkDenseNet121-4b KP. LinkDenseNet121-2b KQ. LinkDenseNet161-8b KR. LinkDenseNet161-4b KS. LinkDenseNet161-2b KT. LinkNet101-8b KU. LinkNet101-4b KV. LinkNet101-2b KW. LinkNet152-8b KX. LinkNet152-4b KY. LinkNet152-2b KZ. LinkNet18-8b LA. LinkNet18-4b LB. LinkNet18-2b LC. LinkNet34-8b LD. LinkNet34-4b LE. LinkNet34-2b LF. LinkNet50-8b LG. LinkNet50-4b LH. LinkNet50-2b LI. lraspp_mobilenet_v3_large-8b LJ. lraspp_mobilenet_v3_large-4b LK. lraspp_mobilenet_v3_large-2b LL. OCNet-8b LM. OCNet-4b LN. OCNet-2b LO. OCNet_ASP_Resnet101-8b LP. OCNet_ASP_Resnet101-4b LQ. OCNet_ASP_Resnet101-2b LR. OCNet_ASP_Resnet152-8b LS. OCNet_ASP_Resnet152-4b LT. OCNet_ASP_Resnet152-2b LU. OCNet_Base_Resnet101-8b LV. OCNet_Base_Resnet101-4b LW. OCNet_Base_Resnet101-2b LX. OCNet_Base_Resnet152-8b LY. OCNet_Base_Resnet152-4b LZ. OCNet_Base_Resnet152-2b MA. OCNet_Pyramid_Resnet101-8b MB. OCNet_Pyramid_Resnet101-4b MC. OCNet_Pyramid_Resnet101-2b MD. OCNet_Pyramid_Resnet152-8b ME. OCNet_Pyramid_Resnet152-4b MF. OCNet_Pyramid_Resnet152-2b MG. PSPNet-4b MH. PSPNet-2b MI. RefineNet4Cascade-8b MJ. RefineNet4Cascade-4b MK. RefineNet4Cascade-2b ML. RefineNet4CascadePoolingImproved-8b MM. RefineNet4CascadePoolingImproved-4b MN. RefineNet4CascadePoolingImproved-2b MO. ResNetDUC-8b MP. ResNetDUC-4b MQ. ResNetDUC-2b MR. ResNetDUCHDC-8b MS. ResNetDUCHDC-4b MT. ResNetDUCHDC-2b MU. Tiramisu103-4b MV. Tiramisu103-2b MW. Tiramisu57-4b MX. Tiramisu57-2b MY. Tiramisu67-4b MZ. Tiramisu67-2b NA. UNet-8b NB. UNet-4b NC. UNet-2b ND. Unet_Plus_Plus-8b NE. Unet_Plus_Plus-4b NF. Unet_Plus_Plus-2b NG. UNet1024-8b NH. UNet1024-4b NI. UNet1024-2b NJ. UNet128-8b NK. UNet128-4b NL. UNet128-2b NM. UNet256-8b NN. UNet256-4b NO. UNet256-2b NP. UNet512-8b NQ. UNet512-4b NR. UNet512-2b NS. UNet960-8b NT. UNet960-4b NU. UNet960-2b NV. UNetDilated-8b NW. UNetDilated-4b NX. UNetDilated-2b video classification: NY. mc3_18-8b NZ. mc3_18-4b OA. mc3_18-2b OB. r2plus1d_18-8b OC. r2plus1d_18-4b OD. r2plus1d_18-2b OE. r3d_18-8b OF. r3d_18-4b OG. r3d_18-2b OH. s3d-8b OI. s3d-4b OJ. s3d-2b speech recognition: OK. hubert_base-8b OL. hubert_base-4b OM. hubert_base-2b ON. hubert_large-2b OO. wave2vec2_base-8b OP. wave2vec2_base-4b OQ. wave2vec2_base-2b OR. wav2vec2_large-2b OS. wav2vec2_large_lv60k-2b transformer: OT. Albert-base-v1-8b OU. Albert-base-v1-4b OV. Albert-base-v1-2b OW. Albert-large-v1-2b OX. Albert-base-v2-8b OY. Albert-base-v2-4b OZ. Albert-base-v2-2b PA. Albert-large-v2-2b PB. BART-base-8b PC. BART-base-4b PD. BART-base-2b PE. BART-large-2b PF. BERT-base-8b PG. BERT-base-4b PH. BERT-base-2b PI. BERT-large-2b PJ. Blenderbot-small-8b PK. Blenderbot-small-4b PL. Blenderbot-small-2b PM. Bloom-2b PN. DistilBERT-base-8b PO. DistilBERT-base-4b PP. DistilBERT-base-2b PQ. Electra-small(discriminator)-8b PR. Electra-small(discriminator)-4b PS. Electra-small(discriminator)-2b PT. Electra-base(discriminator)-8b PU. Electra-base(discriminator)-4b PV. Electra-base(discriminator)-2b PW. Electra-large(discriminator)-2b PX. GPT2-small-8b PY. GPT2-small-4b PZ. GPT2-small-2b QA. GPT2-medium-2b QB. M2M100-2b QC. MegatronBERT-2b QD. MobileBERT-8b QE. MobileBERT-4b QF. MobileBERT-2b QG. MT5-small-8b QH. MT5-small-4b QI. MT5-small-2b QJ. MT5-base-4b QK. MT5-base-2b QL. Nezha-8b QM. Nezha-4b QN. Nezha-2b QO. PLBart-base-8b QP. PLBart-base-4b QQ. PLBart-base-2b QR. T5-small-8b QS. T5-small-4b QT. T5-small-2b QU. T5-base-4b QV. T5-base-2b QW. XGLM-2b



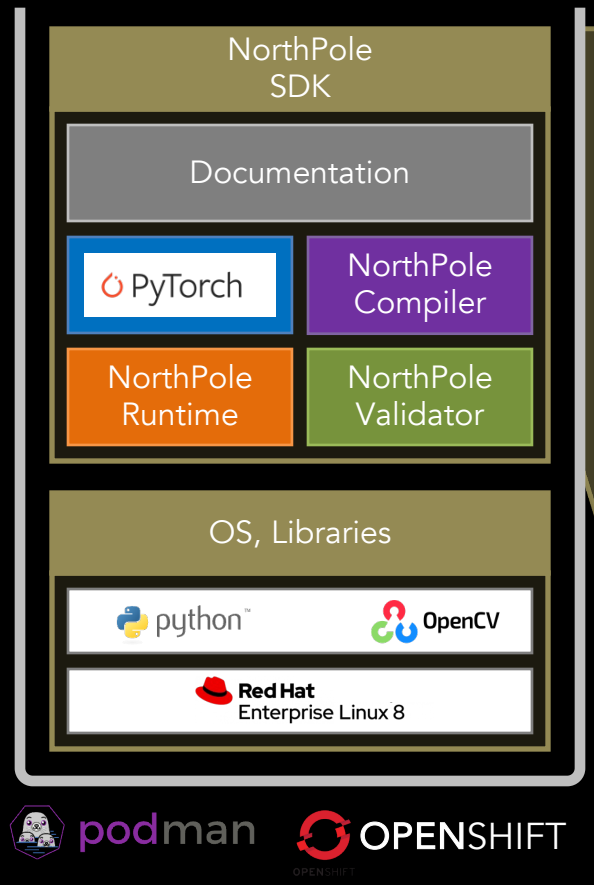


Transformer models supportable by potential multi-chip NorthPole. A64 (A128) indicates model as run on NorthPole in 64 (128) MB activation memory / 128 (64) MB weight memory configuration.

AA. albert-large-v1-2b(A64)
AB. albert-large-v1-4b(A64)
AC. albert-large-v1-8b(A64)
AD. albert-xlarge-v1-2b(A64)
AE. albert-xlarge-v1-4b(A64)
AF. albert-xlarge-v1-8b(A64)
AG. albert-xxlarge-v1-2b(A64)
AH. albert-xxlarge-v1-4b(A64)
AI. BertLarge-2b(A64)
AJ. BertLarge-4b(A64)
AK. BertLarge-8b(A64)
AL. gpt2-Medium-2b(A64)
AM. gpt2-Medium-4b(A64)
AN. gpt2-Medium-8b(A64)
AO. gpt2-Large-2b(A64)
AP. gpt2-Large-4b(A64)
AQ. gpt2-Large-8b(A64)
AR. gpt2-XL-2b(A64)
AS. gpt2-XL-4b(A64)
AT. gpt2-XL-8b(A64)
AU. gpt-neo-2b(A64)
AV. gpt-neo-4b(A64)
AW. gpt-neo-8b(A64)
AX. llama-2b(A64)
AY. m2m100-2b(A64)
AZ. m2m100-4b(A64)
BA. m2m100-8b(A64)
BB. mt5-base-2b(A64)
BC. mt5-base-4b(A64)
BD. mt5-base-8b(A64)
BE. mt5-large-2b(A64)
BF. mt5-large-4b(A64)
BG. mt5-large-8b(A64)
BH. mt5-xl-2b(A64)
BI. mt5-xl-4b(A64)
BJ. t5-base-2b(A64)
BK. t5-base-4b(A64)
BL. t5-base-8b(A64)
BM. t5-large-2b(A64)
BN. t5-large-4b(A64)
BO. t5-large-8b(A64)
BP. xglm-2b(A64)
BQ. xglm-4b(A64)
BR. xglm-8b(A64)
BS. albert-large-v1-2b(A128)
BT. albert-large-v1-4b(A128)
BU. albert-large-v1-8b(A128)
BV. albert-xlarge-v1-2b(A128)
BW. albert-xlarge-v1-4b(A128)
BX. albert-xxlarge-v1-2b(A128)
BY. BertLarge-2b(A128)
BZ. BertLarge-4b(A128)
CA. BertLarge-8b(A128)
CB. gpt2-medium-2b(A128)
CC. gpt2-medium-4b(A128)
CD. gpt2-medium-8b(A128)
CE. gpt2-large-2b(A128)
CF. gpt2-large-4b(A128)
CG. gpt2-large-8b(A128)
CH. gpt2-xl-2b(A128)
CI. gpt2-xl-4b(A128)
CJ. gpt-neo-2b(A128)
CK. gpt-neo-4b(A128)
CL. m2m100-2b(A128)
CM. m2m100-4b(A128)
CN. m2m100-8b(A128)
CO. mt5-base-2b(A128)
CP. mt5-base-4b(A128)
CQ. mt5-xl-2b(A128)
CR. t5-base-2b(A128)
CS. t5-base-4b(A128)
CT. t5-base-8b(A128)
CU. t5-large-2b(A128)
CV. t5-large-4b(A128)
CW. t5-large-8b(A128)
CX. xglm-564M-2b(A128)
CY. xglm-564M-4b(A128)
CZ. xglm-564M-8b(A128)

Networks can be implemented from a large set of possibilities
as more matrix multiplication primitives are added to the software toolchain

NorthPole End-to-end Toolchain



Container preinstalled with SDK

Train Flow (Offline)

PyTorch — PyTorch API to adapt network for NorthPole and train on GPU

NorthPole Compiler — Compiler to export network to hardware-ready model

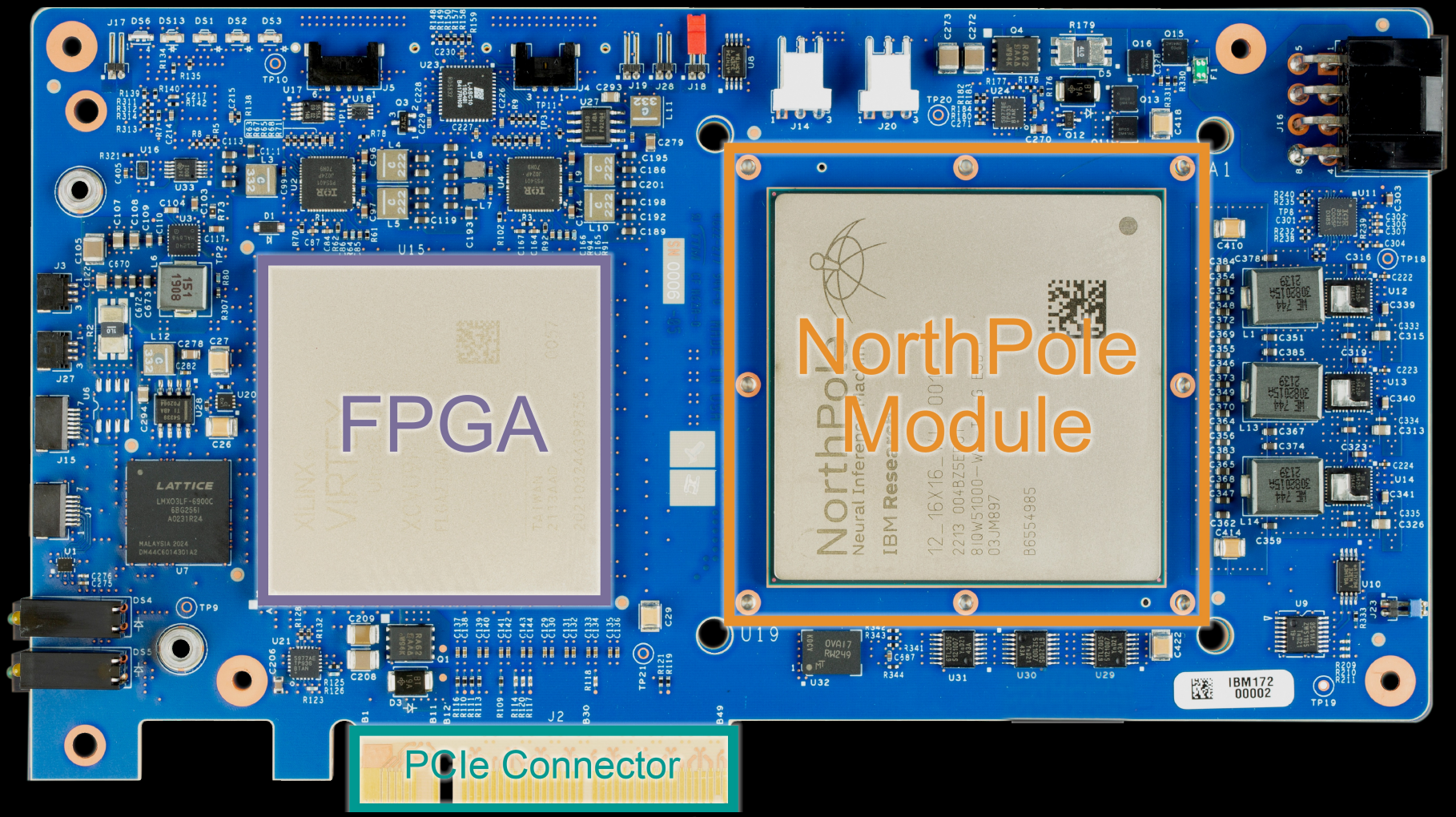
Run Flow

NorthPole Runtime — Runtime API to deploy model on NorthPole

NorthPole Validator — Validator to emulate NorthPole in software

Example applications with full source code, pretrained networks

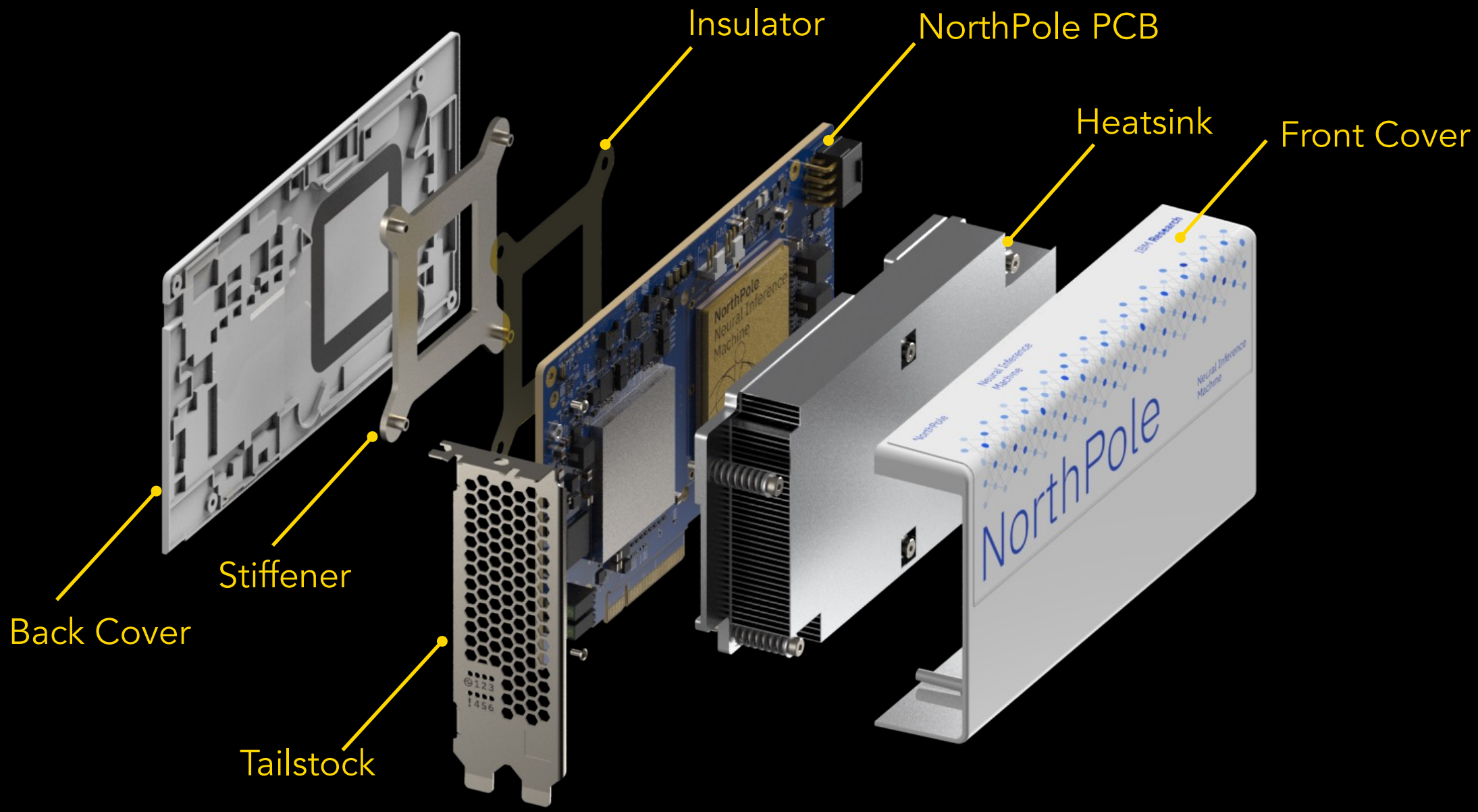
NorthPole Systems



(Research Prototype)

FPGA is only used for PCIe bridge

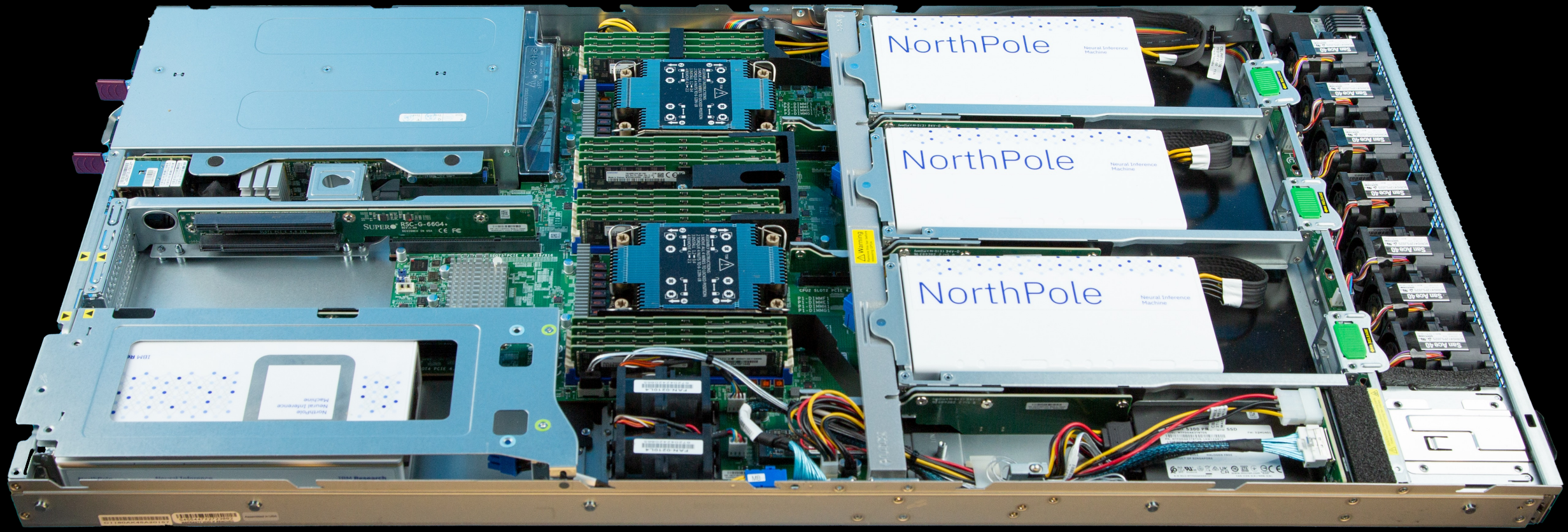
No HBM/external memory, No CPU Cores



NorthPole PCIe assembly (Research Prototype)

Single NorthPole assembly in a 1U server (Research Prototype)

To scale-out, a model can be striped across chips,
increasing FPS and parameter memory
while keeping energy-, space-, and latency-efficiencies,
with only low-bandwidth data tensors moving via PCIe



Four NorthPole assemblies in a server (Research Prototype)
… 8, 10, 12, 16 assemblies in a server are possible

NorthPole … is specialized to inference
… performs at the frontier of energy, space, and time
… can support many deep networks in vision, speech, and natural language
… has brain-inspired and silicon-optimized architecture
… has modular, tileable architecture – like the cortex
… has massive parallelism – like the cortex
… has mixed-precision – like the cortex
… has memory-near-compute – like the cortex
… has no off-chip / centralized memory and no von Neumann bottleneck – like the cortex
… has only three commands: write tensor(s), run network, read tensor(s) – is an active memory
… has minimum IO bandwidth requirement
… has minimum load on the host
… has two dense brain-inspired networks-on-chip
… has two dense silicon-optimized networks-on-chip
… has no VLIW
… has pre-scheduled, deterministic operation in the core array free from cache-misses
… has unscheduled, input-driven operation in the framebuffer for queuing and isolation
… has co-designed mixed-precision training algorithms
… has an end-to-end software toolchain
… has a current PCIe implementation with many possible custom boards
… has an easy scale-out implementation
… has significant headroom in terms of system scaling, silicon scaling, architecture innovations

Thank you!

dmodha@us.ibm.com

Notes on BERT-base Performance Comparison

1. Comparative approaches use a sequence length of 128. Their performance metrics are scaled by a factor of 3x, scaled to the compute required for a sequence length of 384.

2. The 3x scaling factor from sequence length of 128 to 384 was validated based on A100 GPU performance numbers on BERT-large, which are reported for both sequence lengths.

3. This is a reasonable and a conservative upper bound, as the compute and communication required by the network scale by a factor of 3x.

4. It does not account for the fact that the longer sequence network may not fit in chip memories sized for the shorter sequence length, or similar caching effects. This would lead to scaling worse than a factor of 3x.