Neural Inference at the Frontier of Energy, Space, and Time

Filipp Akopyan, Principal Research Scientist, Chip Design Team Lead

ACAT 2024

NorthPole Published in *Science*

October 20, 2023

Source: AAAS

AI computing reaches for the edge

A chip design integrates computation and memory to efficiently process data at low energy cost

By **Subramanian S. Iyer** and **Vwani Roychowdhury**

edge" that monitor and track in real time, for example, must communicate with powerful centralized AI models and then bring

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > news > article

NEWS | 19 October 2023

'Mind-blowing' IBM chip speeds up AI

IBM's NorthPole processor sidesteps need to access external memory, boosting computing power and saving energy.

Davide Castelvecchi

Source: Nature

SCIENCE & TECH

Microchip breakthrough may reshape the future of AI

IBM's new NorthPole may enable smarter, more efficient, network-independent devices that may even help the U.S. win the microchip war against China.

BY PATRICK TUCKER
SCIENCE & TECHNOLOGY EDITOR, DEFENSE ONE

OCTOBER 19, 2023

A prototype microchip design revealed today by IBM could pave the way for a world of much smarter devices that don't rely on the cloud or the internet for their intelligence. That could help soldiers who operate drones, ground robots, or augmented-reality gear against adversaries who can target electronic emissions. But the new chip—modeled loosely on the human brain—also paves the way for a different sort of AI, one that doesn't rely on big cloud and data companies like Amazon or Google.

Unlike traditional chips that separate memory from processing circuits, the NorthPole chip combines the two—like synapses in the brain that hold and process information based on their connection to other neurons. Writing in the journal *Science*, IBM researchers call it a "neural inference architecture that blurs this boundary by eliminating off-chip memory, intertwining compute with memory on-chip, and appearing externally as an active memory."

Source: Defense One

FORBES > INNOVATION > CLOUD

IBM Research Shows Off New NorthPole Neural Accelerator

Steve McDowell Contributor

Principal analyst and founding partner at NAND Research.

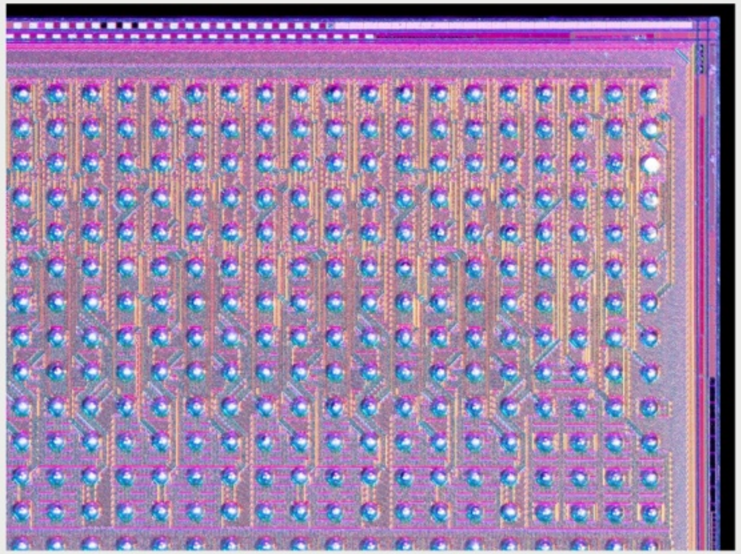
Follow

Source: Forbes

NEWS SEMICONDUCTORS

IBM Debuts Brain-Inspired Chip For Speedy, Efficient AI ›NorthPole is the top, says its maker

BY CHARLES Q. CHOI | 23 OCT 2023 | 3 MIN READ

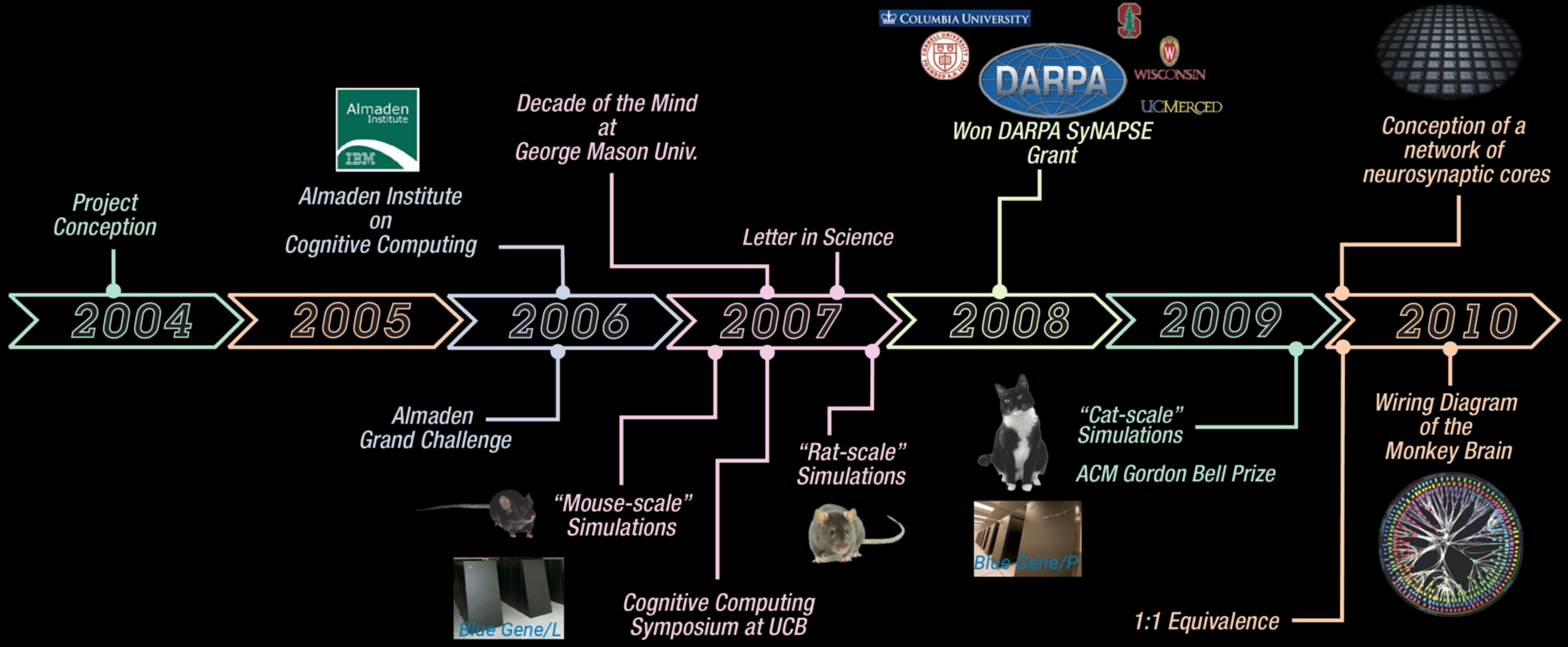


IBM's NorthPole AI chip packs 22 billion transistors into just 800 square millimeters. IBM

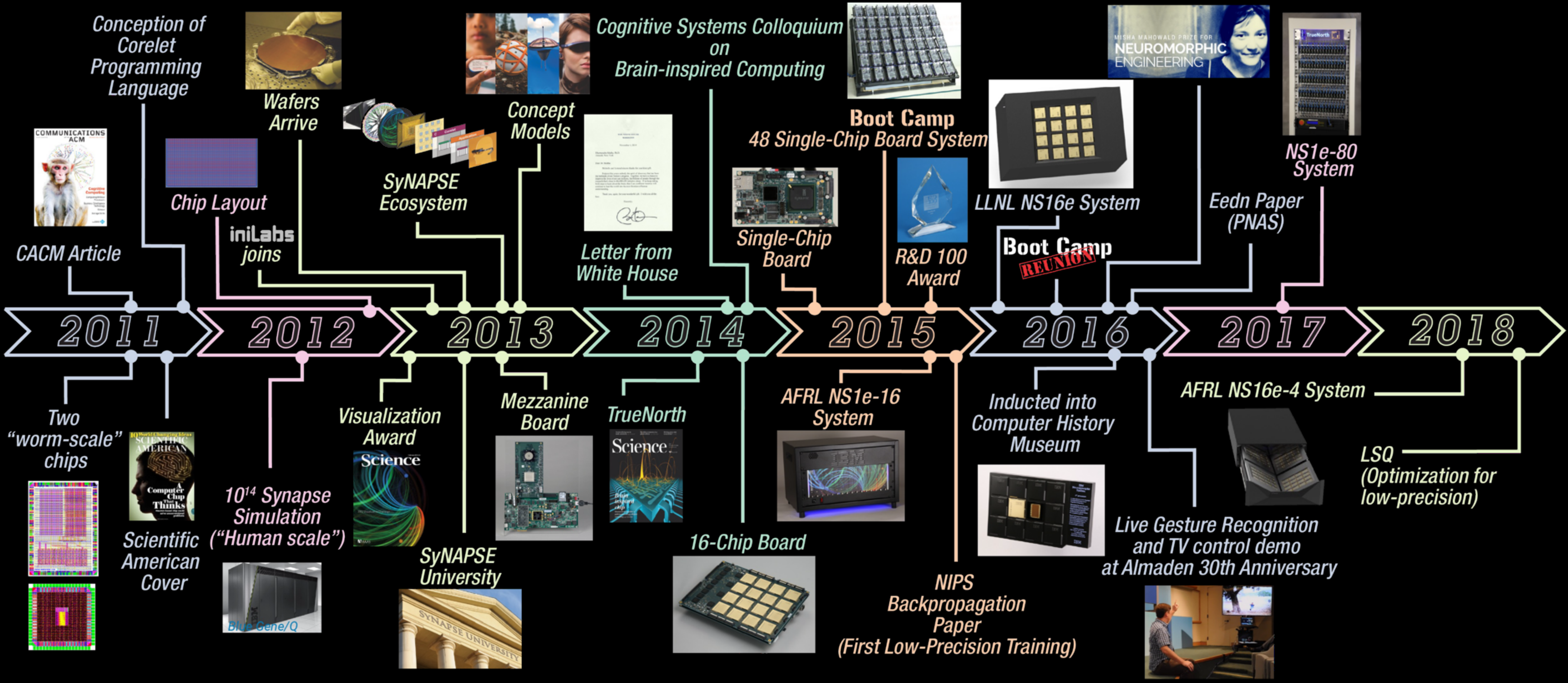
Source: IEEE Spectrum

Subu Iyer and Vwani Roychowdhury

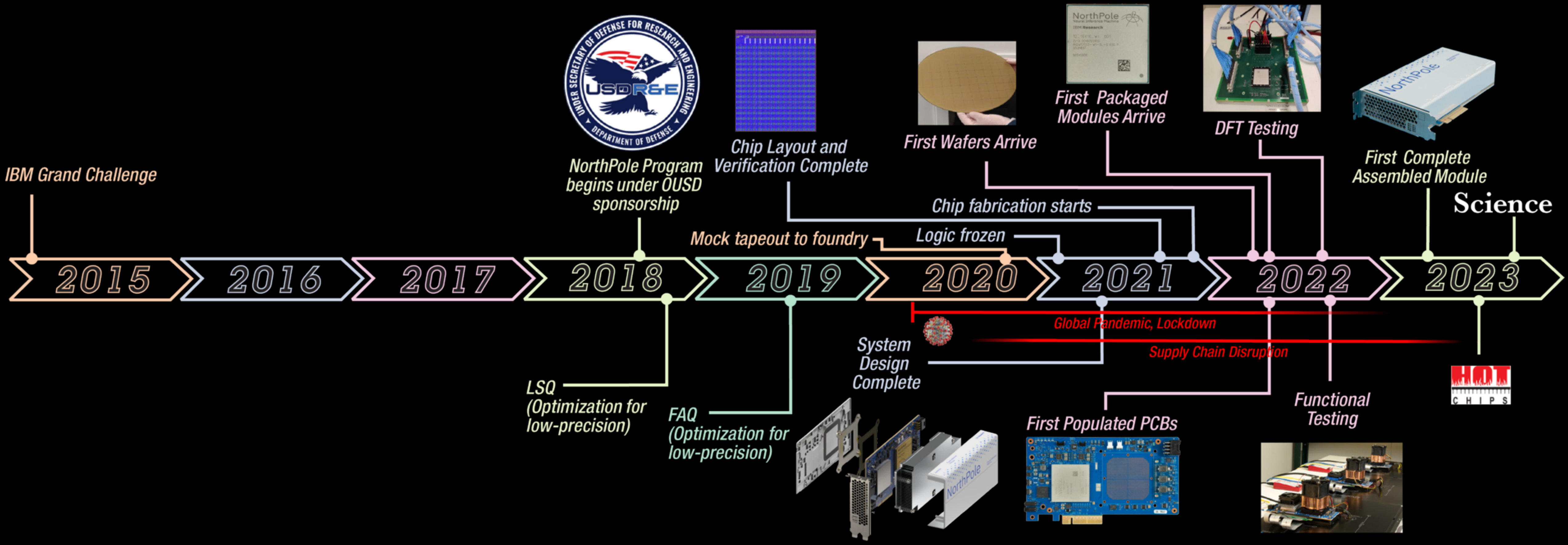
Supercomputing Era



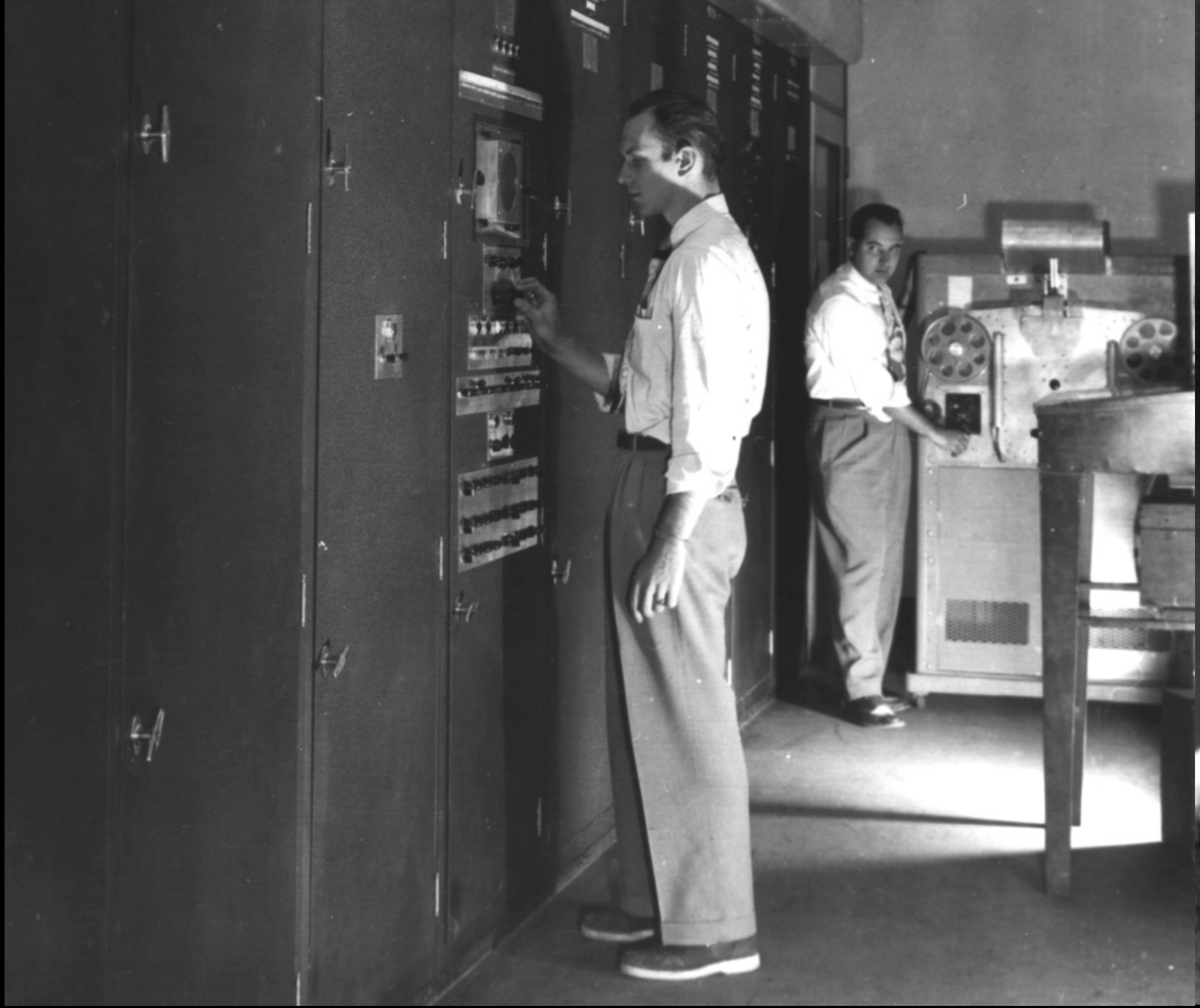
TrueNorth Era



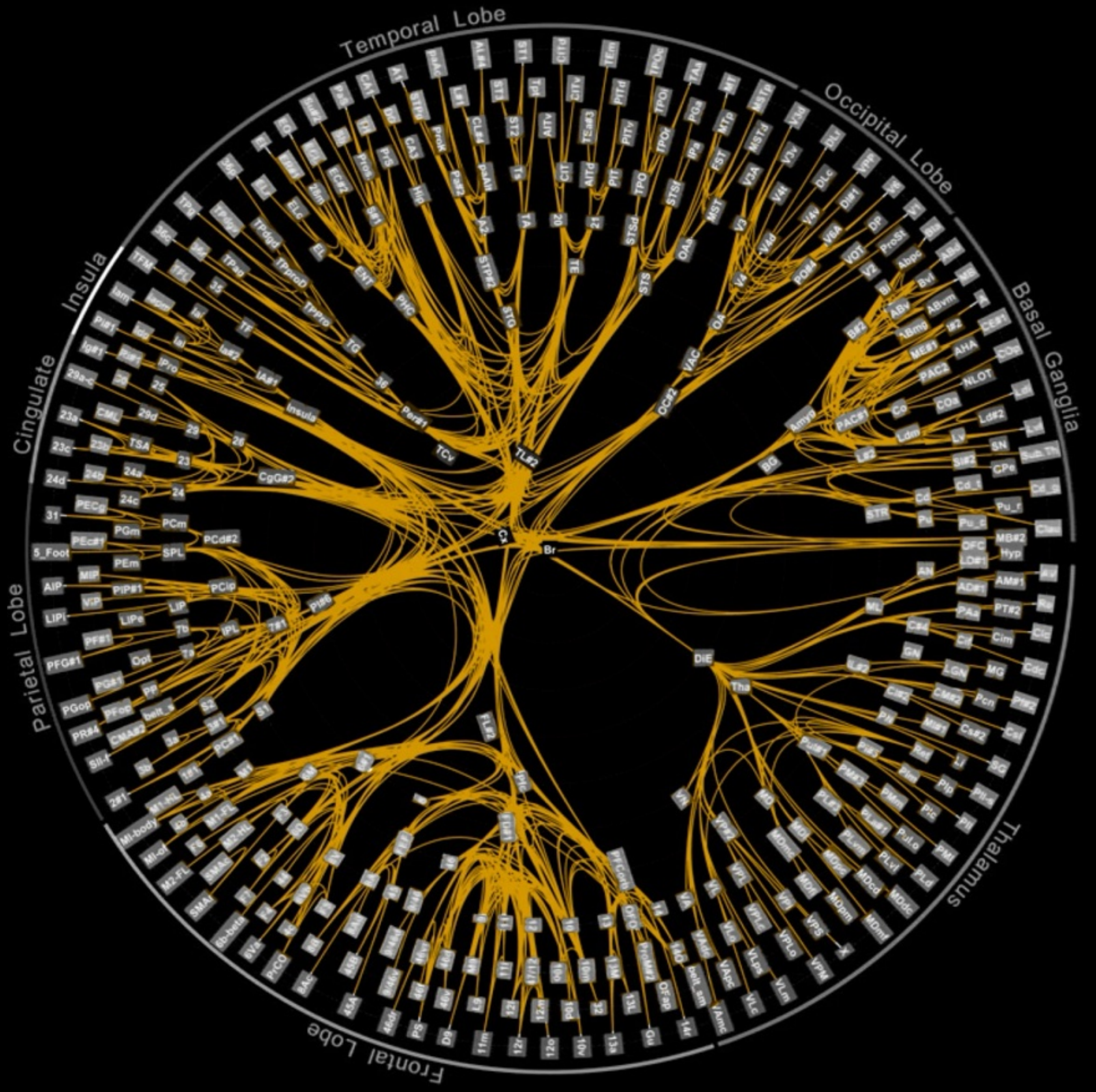
NorthPole Era



NorthPole is Inspired by the brain and optimized for silicon.



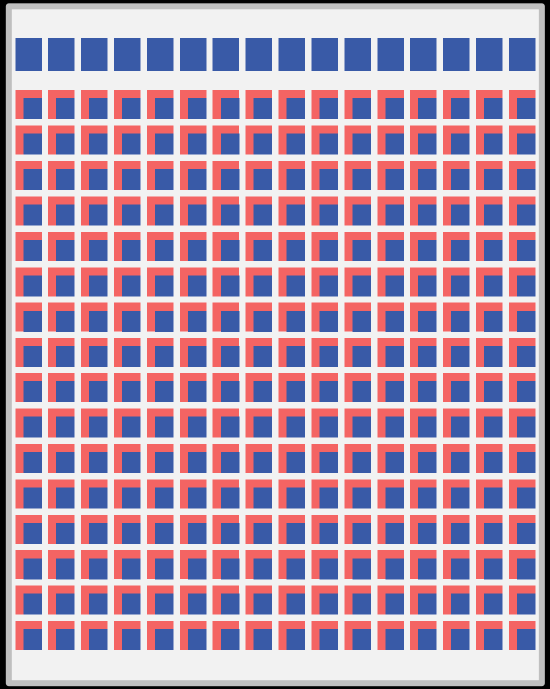
Computing is still dominated by EDVAC and its von Neumann architecture, which separates memory from compute.



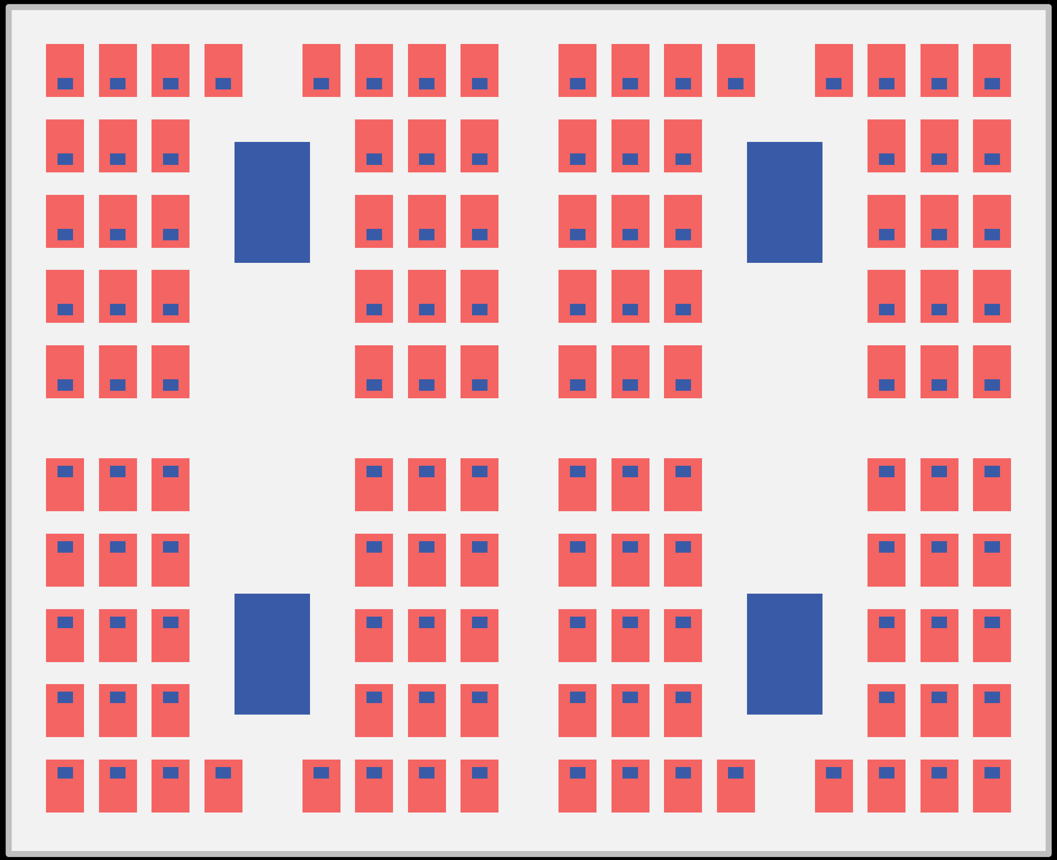
The brain is vastly more energy-efficient than modern computers, in part because it stores memory with compute in every neuron.

Compute | Memory

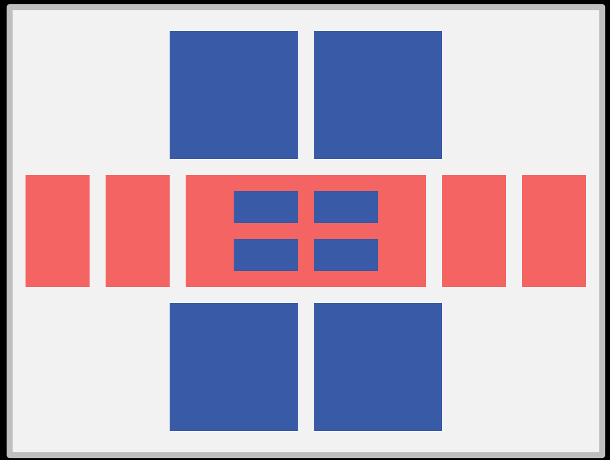
IBM NorthPole



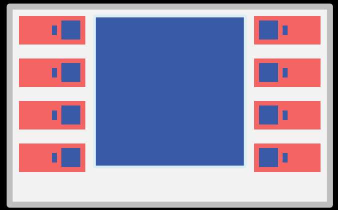
Other Contemporary AI Architectures



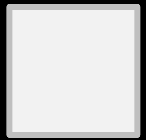
GPU (A100)



TPU



CPU (Zen 3)



1 billion transistors

Inspired by the brain, NorthPole stores memory near compute, with no centralized or off-chip memory, mitigating von Neumann bottleneck (unlike contemporary architectures).

NorthPole Axiomatic Architecture

Axiom 1: neural inference specialization

Axiom 2: brain-inspired low precision

Axiom 3: brain-inspired distributed, modular core array with massive compute parallelism within and among cores

Axiom 4: brain-inspired memory near compute

Axiom 5: brain-inspired networks-on-chip

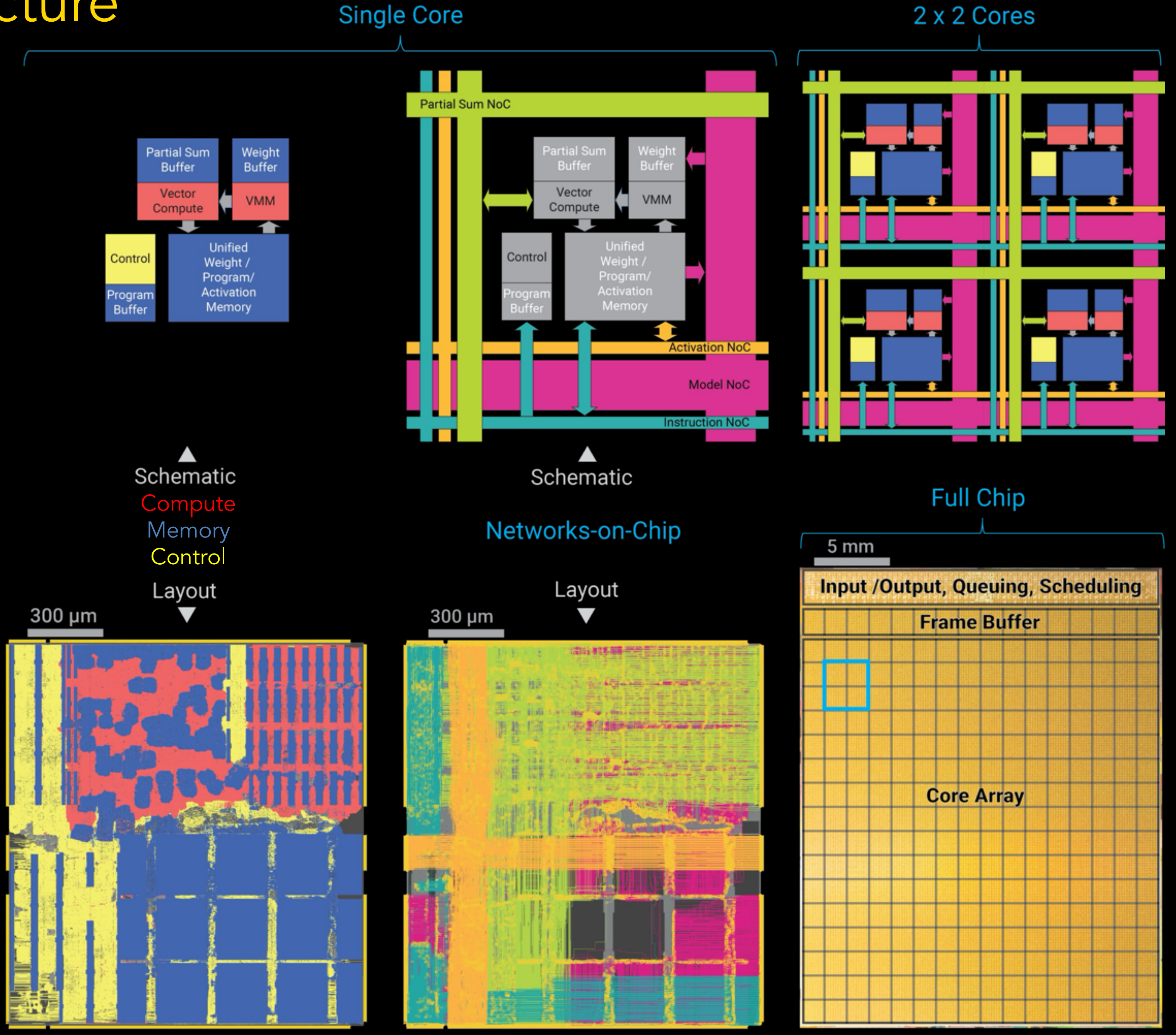
Axiom 6: silicon-optimized networks-on-chip

Axiom 7: stall-free, deterministic control

Axiom 8: co-optimized training algorithms

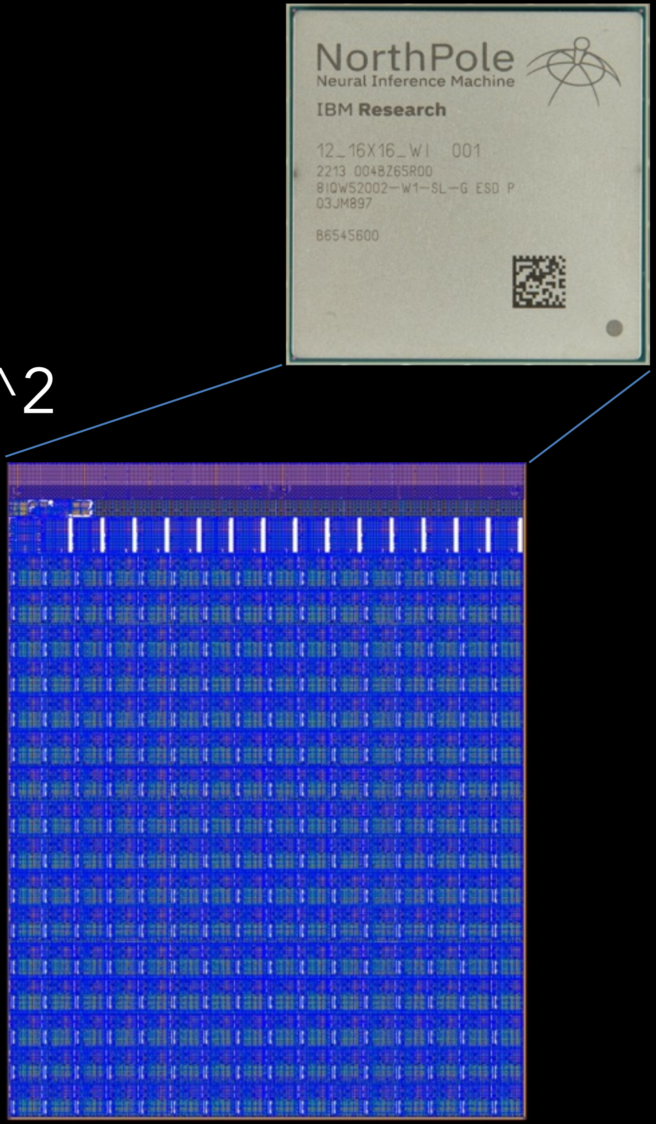
Axiom 9: co-designed software

Axiom 10: simplest usage model – write input, run network, read output



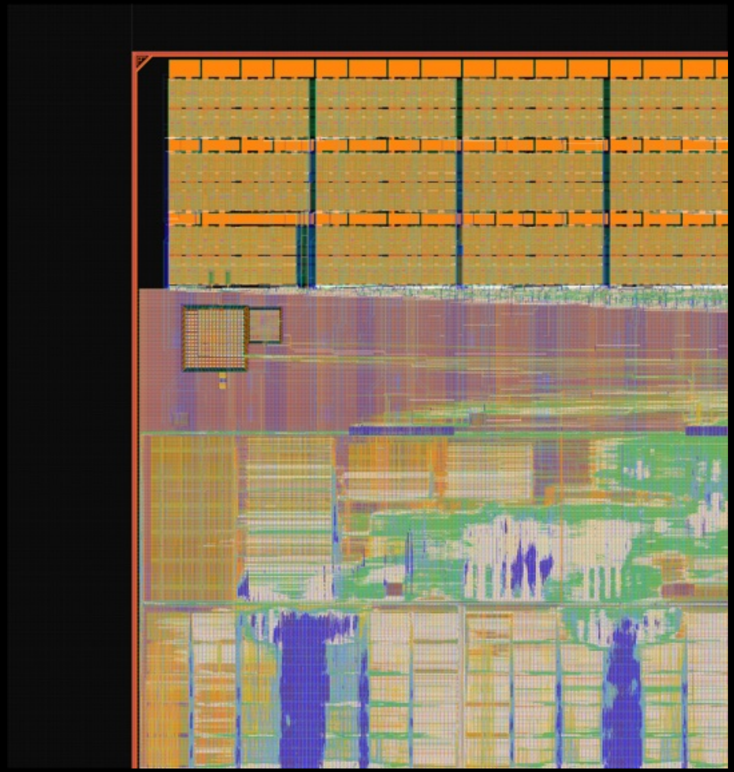
NorthPole Chip Implementation Data

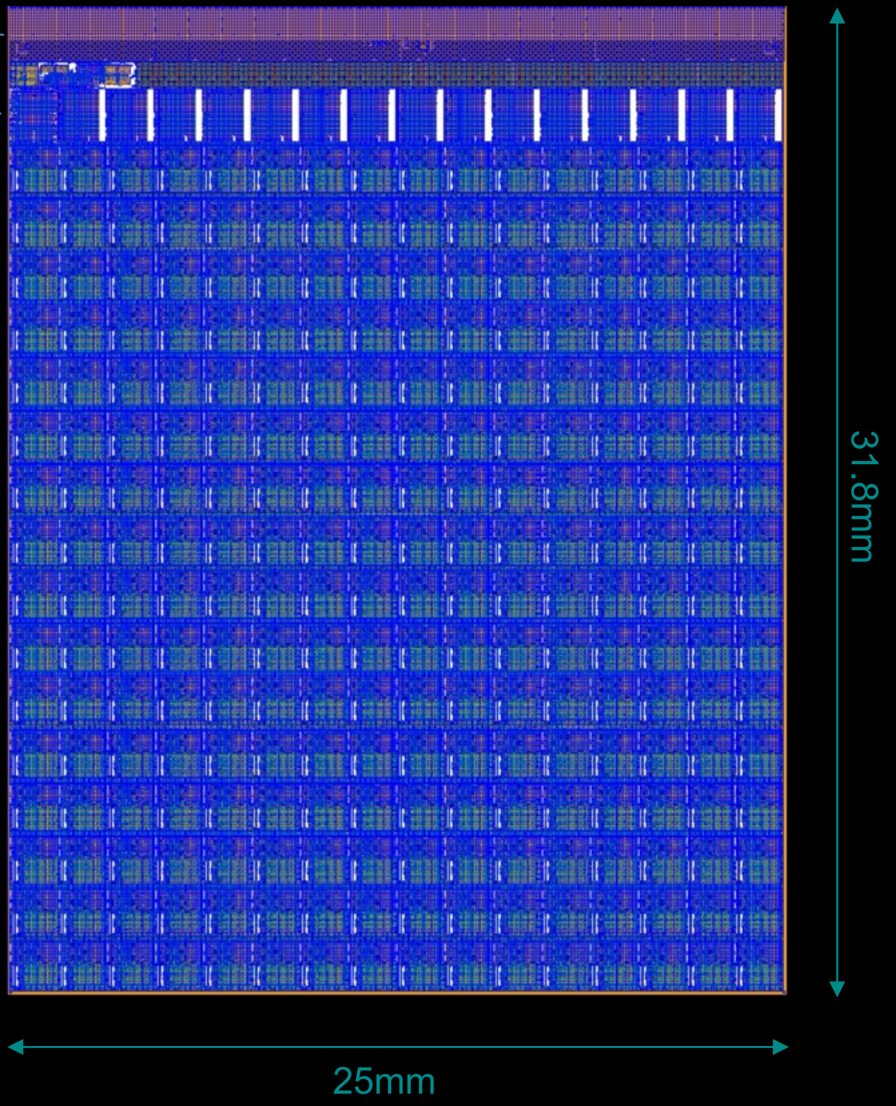
- Fabricated in a 12nm process
- Reticle Size: 25 mm x 31.8 mm
- Total Instances: ~565 million (plus physical cells)
- Transistor Count: ~21.8 billion (ignoring Fill) → 27.4 MT/mm^2
- Total nets: ~592 million
- Total Signal wire length: ~17.8 km
- Total Power/Ground wire length: 1.6 km
- Total number of vias: ~9 billion



Chip-level Manufactured GDS

Some Layers, Routing and Metal Fill are not shown for better visibility into the chip layout

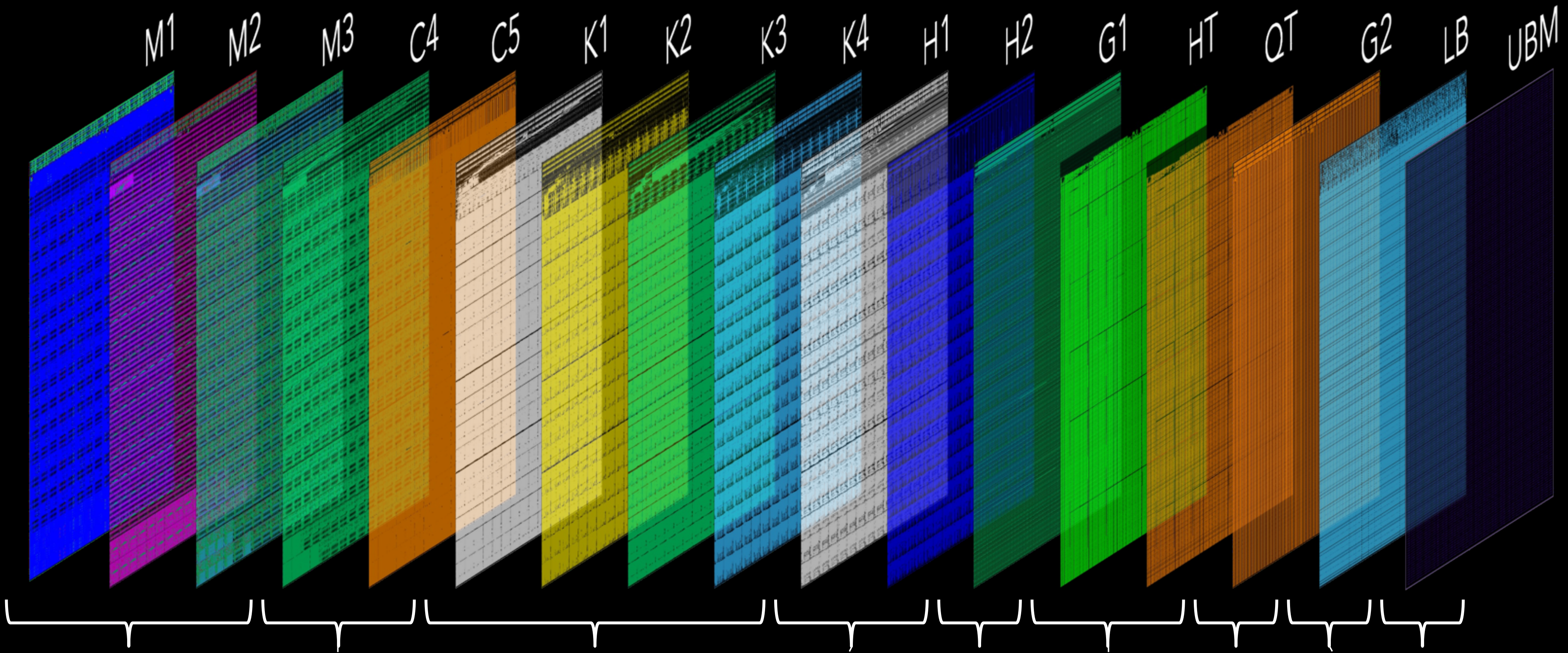




Selected Challenges:

- Largest-possible die size in this Process for maximum capacity
- Extreme high-density cell placement for compute parallelism
- Extreme high-density wiring for large bandwidth
- Hand-crafted clock trees for low skew and synchronization
- Hand-crafted pad placement and top routing for optimal I/O

NorthPole Signal Routing Layers – 13 Usable Metal Layers



M1, M2, M3: dense metals for local wiring

C4, C5: semi-dense metals for short-range wiring

K1, K2, K3, K4: medium-width metals for medium-range wiring

H1, H2: wide metals for long-range wiring

G1: wide metal for power and clock

HT, QT: Special plate-metal layers for metal-insulator-metal capacitors

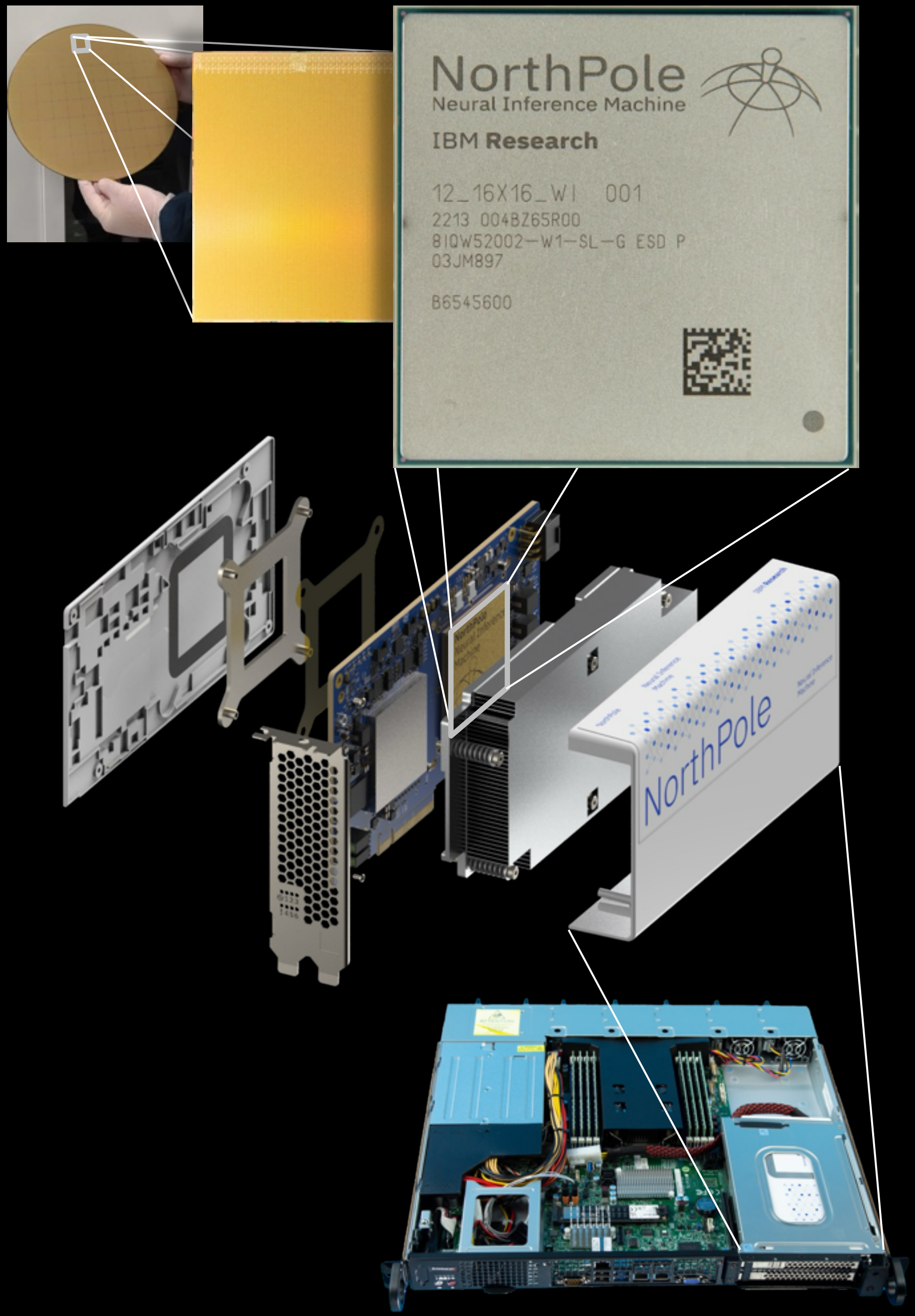
G2: wide metal for power and clock

LB: ultra-wide metal for wiring to bumps

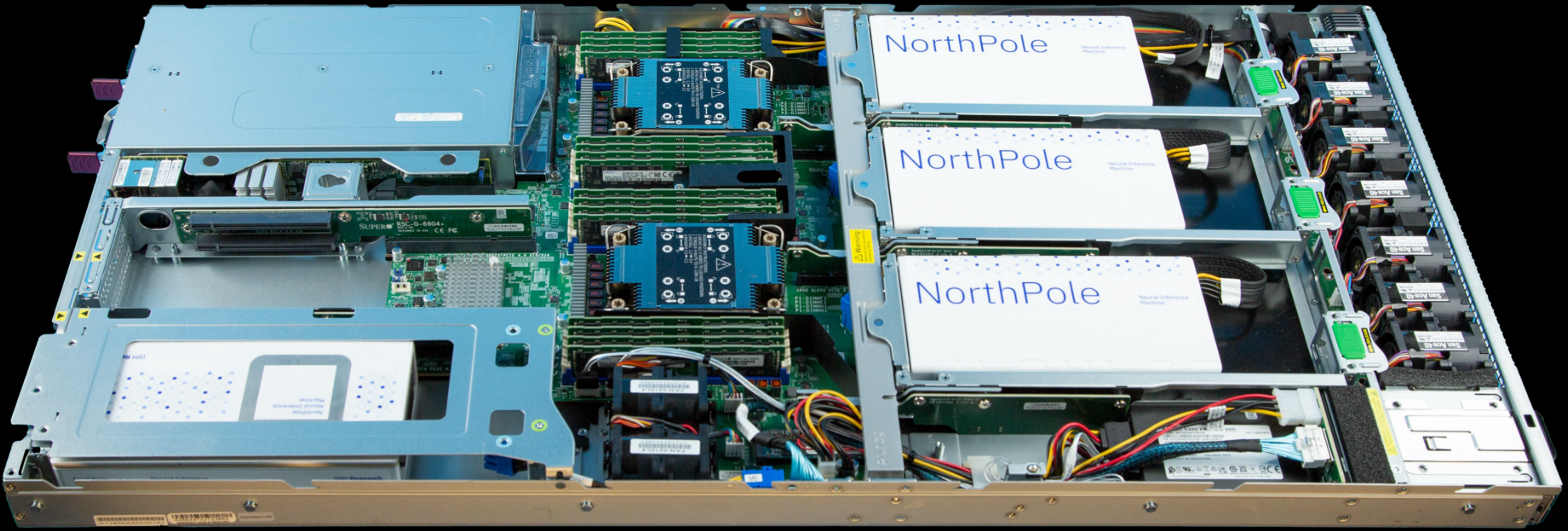
UBM: under-bump metall-ization

NorthPole Chip, Board, System Design

- Fully operational in first silicon implementation
- Compute: has 256 cores; has 2,048 (4,096 and 8,192) operations per core per cycle at 8-bit (at 4-bit and 2-bit, respectively) precision
- Memory: has 224MB of on-chip memory (192MB in core array, 32MB framebuffer for input-output)
- Communication: has over 4,096 wires crossing each core both horizontally and vertically
- Control: has 2,048 threads
- Deployed in a PCIe form factor research prototype printed circuit board
- Integrated into an end-to-end software toolchain



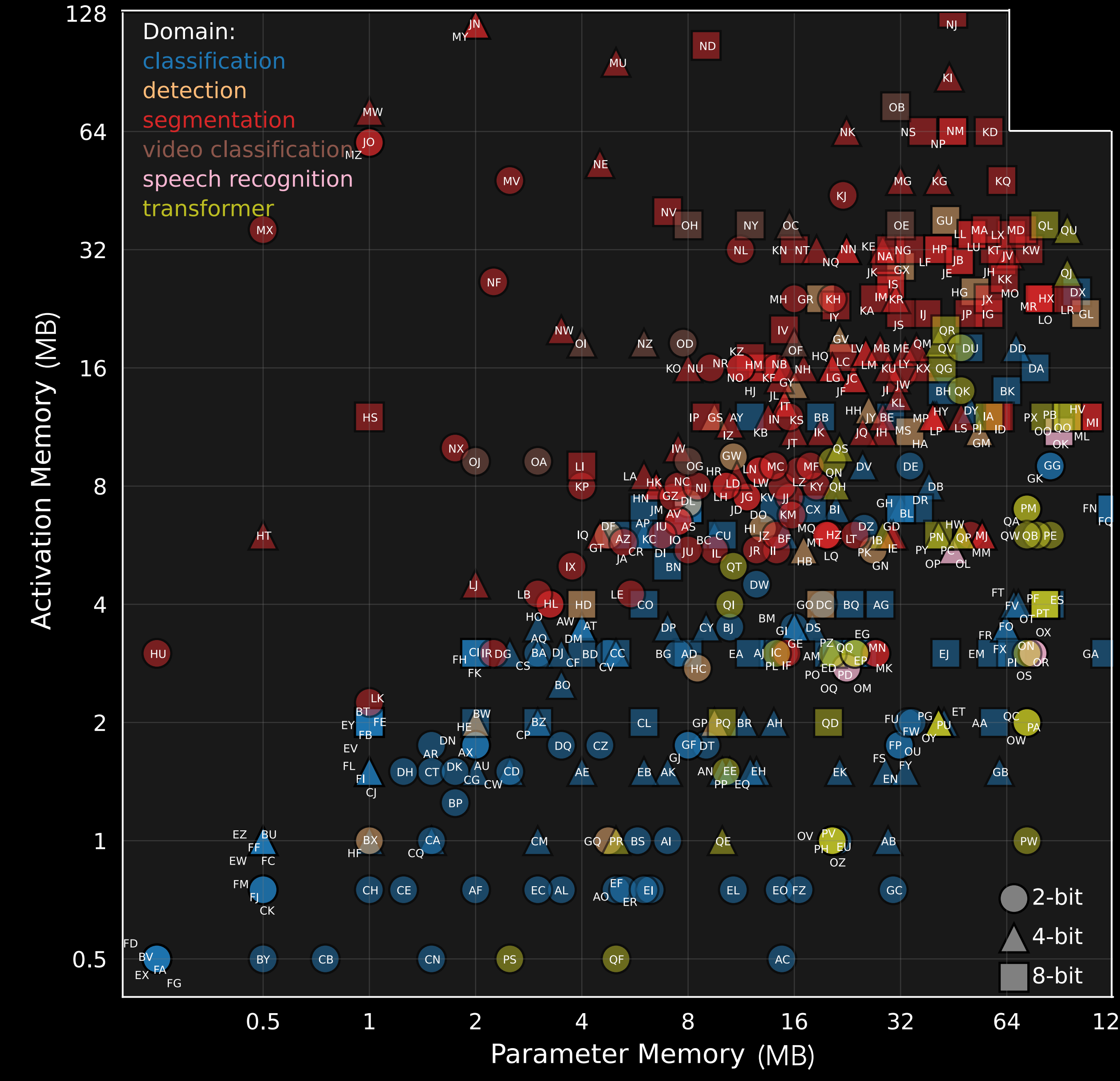
To scale-out, a model can be striped across chips, increasing FPS and parameter memory while keeping energy-, space-, and latency-efficiencies, with only low-bandwidth data tensors moving via PCIe



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

NorthPole Architecture trumps Moore’s Law.

NorthPole can support various AI applications



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

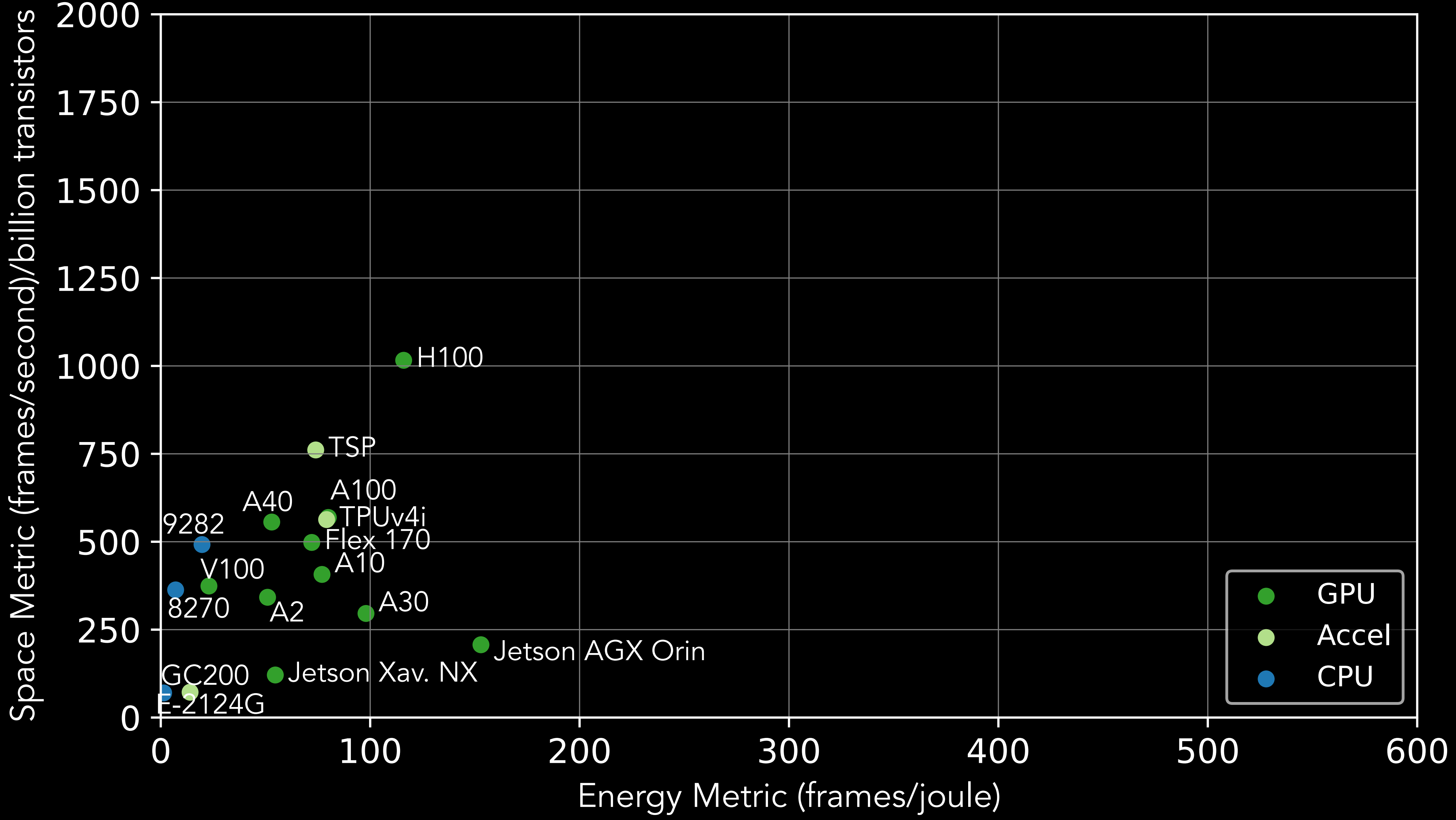
detection : GL. Faster-RCNN-8b GM. Faster-RCNN-4b GN. Faster-RCNN-2b GO. fasterrcnn_mobilenetv3_large_320_fpn-8b GP. fasterrcnn_mobilenetv3_large_320_fpn-4b GQ. fasterrcnn_mobilenetv3_large_320_fpn-2b GR. fasterrcnn_mobilenetv3_large_fpn-8b GS. fasterrcnn_mobilenetv3_large_fpn-4b GT. fasterrcnn_mobilenetv3_large_fpn-2b 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

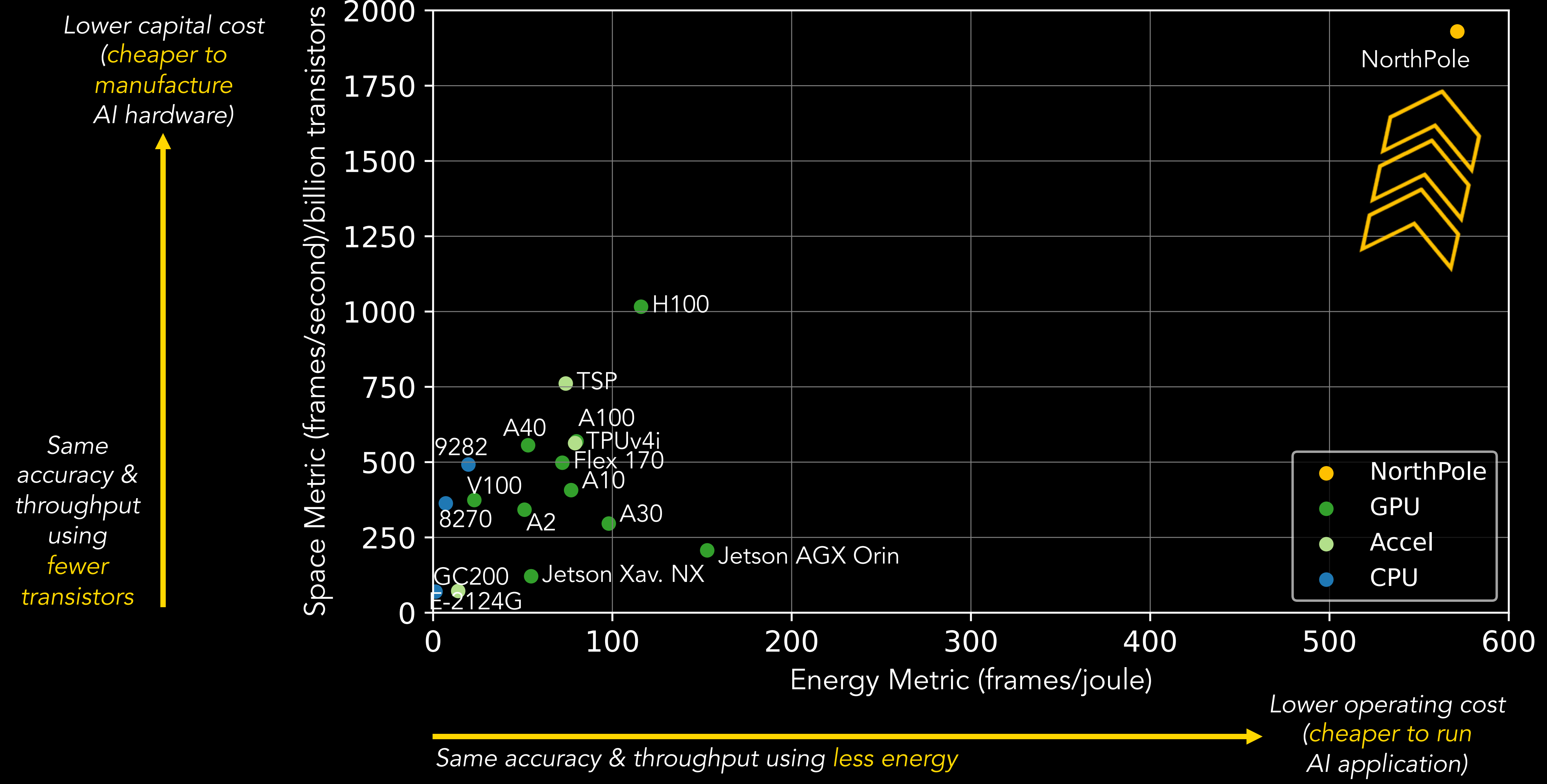
video classification: NY. mc3_18-8b NZ. mc3_18-4b OA. mc3_18-2b OB. r2plus1d_18-8b OC. r2plus1d_18-4b OD. r2plus1d_18-2b OE. r3d_18-8b OF. r3d_18-4b OG. r3d_18-2b OH. s3d-8b OI. s3d-4b OJ. s3d-2b

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

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



* Comparative numbers as published in papers, respective parties, 3rd parties (reference available for every datapoint).



- ResNet-50 AI benchmark for image classification, at state-of-art inference accuracy
- Similar Result on Yolo-v4 and BERT-Base AI benchmarks (see HotChips presentation)
- NorthPole outperforms ALL prevalent architectures, even those using advanced processes

NorthPole running Yolo-v4 on dashcam video.

NorthPole running PSPNet on dashcam video.

NorthPole processing a Yolo-v4 network with input from 2 cameras in real-time using under 5W of chip power.

Science paper: https://modha.org
HotChips: https://hotchips.org/

IBM