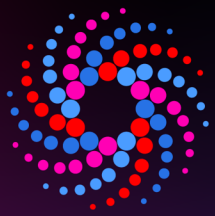


A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads

Murali Emani
ALCF, Argonne National Laboratory
memani@anl.gov

13th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS) 2022

Work of many

Collaboration between Argonne, Cerebras, SambaNova, Graphcore, and Groq

Murali Emani* memani@anl.gov
Zhen Xie* zhen.xie@anl.gov
Siddhisanket Raskar* sraskar@anl.gov
Varuni Sastry* vsastry@anl.gov
William Arnold* arnoldw@anl.gov
Bruce Wilson* wilsonb@anl.gov

Rajeev Thakur* thakur@anl.gov
Venkatram Vishwanath* venkat@anl.gov
Zhengchun Liu* zhengchun.liu@anl.gov
Michael E. Papka*‖ papka@anl.gov
Cindy Orozco Bohorquez† cindy@cerebras.net

Rick Weisner‡ rick.weisner@sambanova.ai
Karen Li‡ xiaoyan.li@sambanova.ai
Yongning Sheng‡ yongning.sheng@sambanova.ai
Yun Du‡ yun.du@sambanova.ai

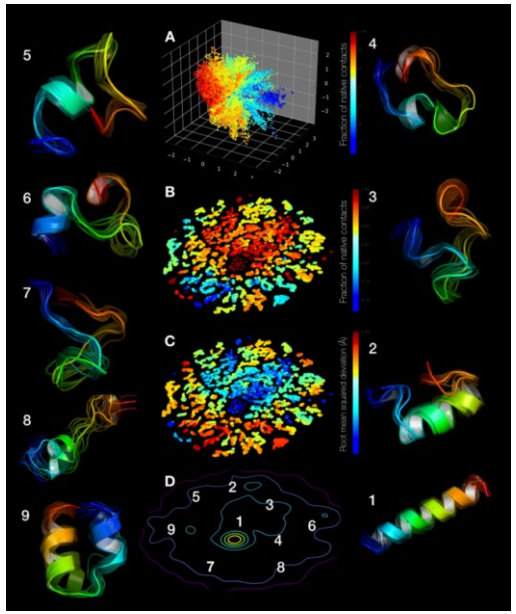
Jian Zhang‡ jian.zhang@sambanova.ai
Alexander Tsyplikhin§ alext@graphcore.ai
Gurdaman Khaira§ damank@graphcore.ai
Jeremy Fowers¶ jfowers@groq.com
Ramakrishnan Sivakumar¶ rsivakumar@groq.com

Victoria Godsoe¶ vgodsoe@groq.com
Adrian Macias¶ am@groq.com
Chetan Tekur¶ ctekur@groq.com
Matthew Boyd¶ matt@groq.com

*Argonne National Laboratory, Lemont, IL 60439, USA, †Cerebras Systems, Sunnyvale, CA 95085, USA,
‡SambaNova Systems Inc., Palo Alto, CA 94303, USA, §Graphcore Inc., Palo Alto, CA 94301, USA,
¶Groq Inc., Mountain View, CA 94041, USA, ‖University of Illinois, Chicago, IL 60637, USA

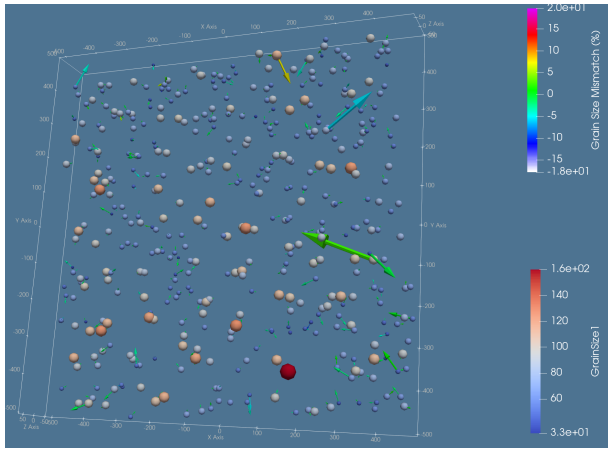
Surge of Scientific Machine Learning

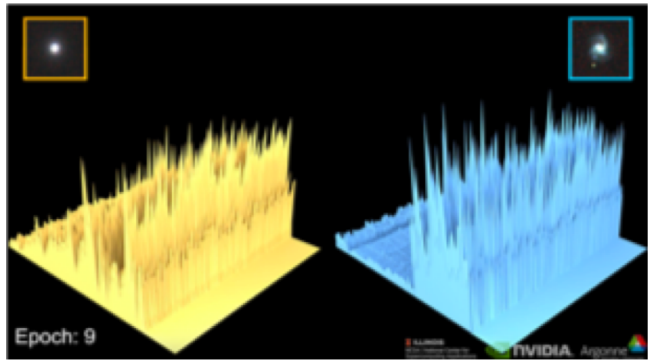
- Simulations/ surrogate models
  - Replace, in part, or guide simulations with AI-driven surrogate models
- Data-driven models
  - Use data to build models without simulations
- Co-design of experiments
  - AI-driven experiments



Protein-folding

Braggs Peak





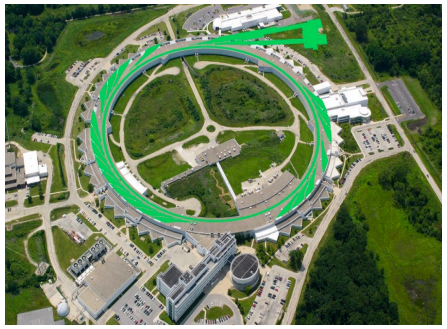
Galaxy Classification

Design infrastructure to facilitate and accelerate AI for Science (AI4S) applications

Integrating AI Systems in Facilities

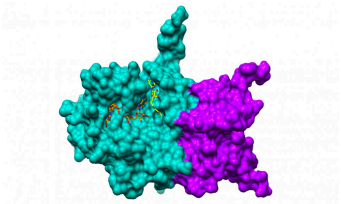
Computing Facility

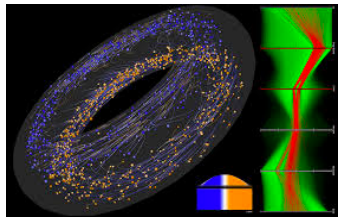
Experimental Facility



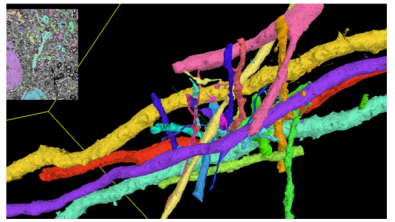
AI-Accelerators

AI-Edge accelerator





Simulations



Data-driven Models

ALCF AI Testbed

https://www.alcf.anl.gov/alcf-ai-testbed

Cerebras (CS-2)

SambaNova

Graphcore

Habana

Groq

- Infrastructure of next-generation machines with AI hardware accelerators
- Provide a platform to evaluate usability and performance of AI4S applications
- Understand how to integrate AI systems with supercomputers to accelerate science

	Cerebras CS-2	SambaNova Cardinal SN10	Groq GroqCard	GraphCore GC200 IPU	Habana Gaudi1	NVIDIA A100
Compute Units	850,000 Cores	640 PCUs	5120 vector ALUs	1472 IPUs	8 TPC + GEMM engine	6912 Cuda Cores
On-Chip Memory	40 GB	>300MB	230MB	900MB	24 MB	192KB L1 40MB L2
Process	7nm	7nm	14nm	7nm	7nm	7nm
System Size	2 Nodes	2 nodes (8 cards per node)	4 nodes (8 cards per node)	1 node (8 cards per node)	2 nodes (8 cards per node)	Several systems
Software Stack Support	Tensorflow, Pytorch	SambaFlow, Pytorch	GroqAPI, ONNX	Tensorflow, Pytorch, PopArt	Synapse AI, TensorFlow and PyTorch	Tensorflow, Pytorch, etc
Interconnect	Ethernet-based	Infiniband	RealScale ™	IPU Link	Ethernet-based	NVLink

Challenges

- Understand how these systems perform for different workloads given diverse hardware and software characteristics
- What are the unique capabilities of each evaluated system
- Opportunities and potential for integrating AI accelerators with HPC computing facilities

Approach

- Perform a comprehensive evaluation with a diverse set of Deep Learning (DL) models:
  - *DL primitives*: GEMM, Conv2D, ReLU, and RNN
  - *Benchmarks*: U-Net, BERT-Large, ResNet-50
  - *AI4S applications*: BraggNN and Uno
  - Scalability and Collective communications

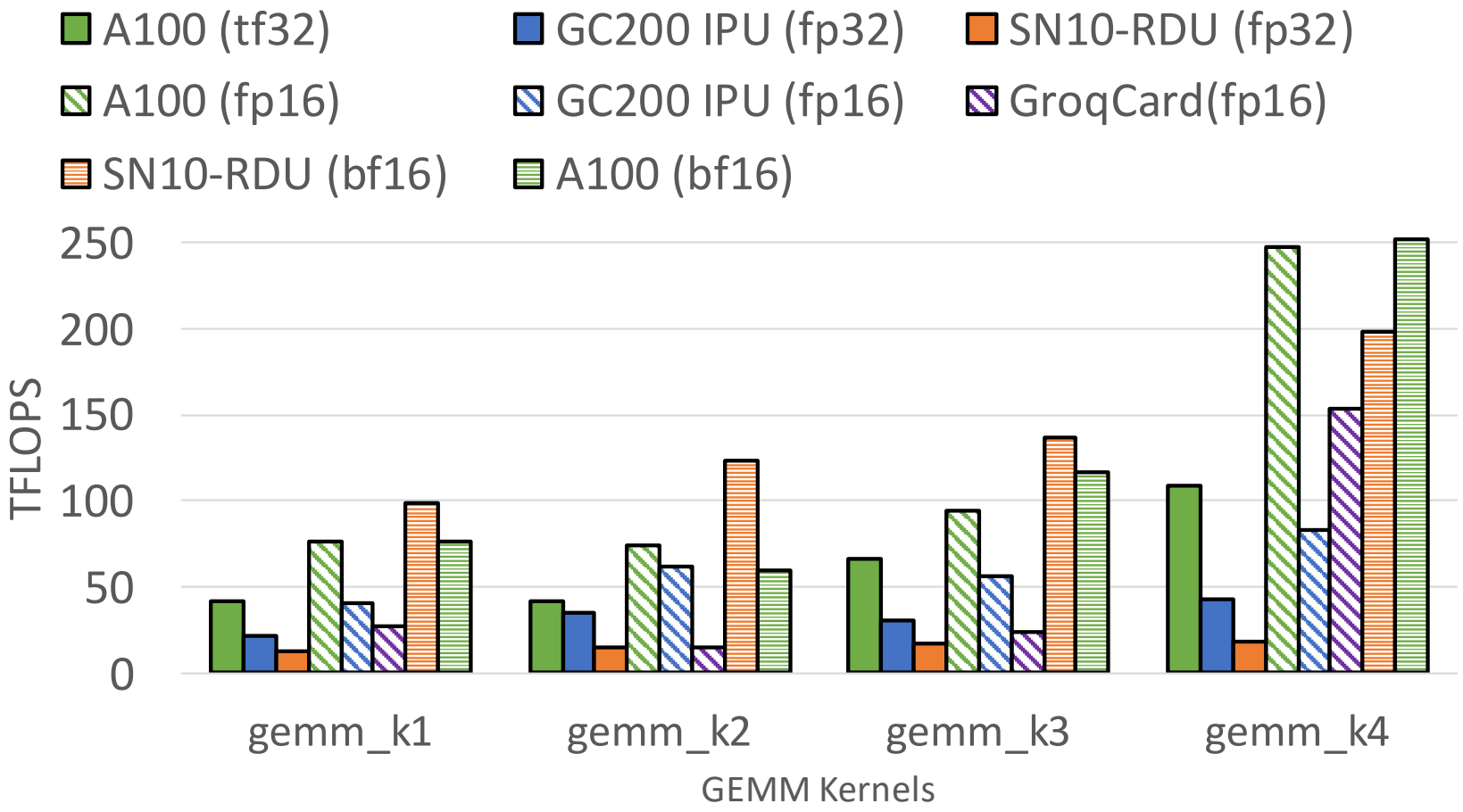
Approach

- Perform a comprehensive evaluation with a diverse set of Deep Learning (DL) models:
  - *DL primitives*: GEMM, Conv2D, ReLU, and RNN
  - *Benchmarks*: U-Net, BERT-Large, ResNet-50
  - *AI4S applications*: BraggNN and Uno
  - Scalability and Collective communications

- Evaluated SambaNova, Cerebras, Graphcore, Groq systems and Nvidia A100 as a baseline*

* run out-of-box.

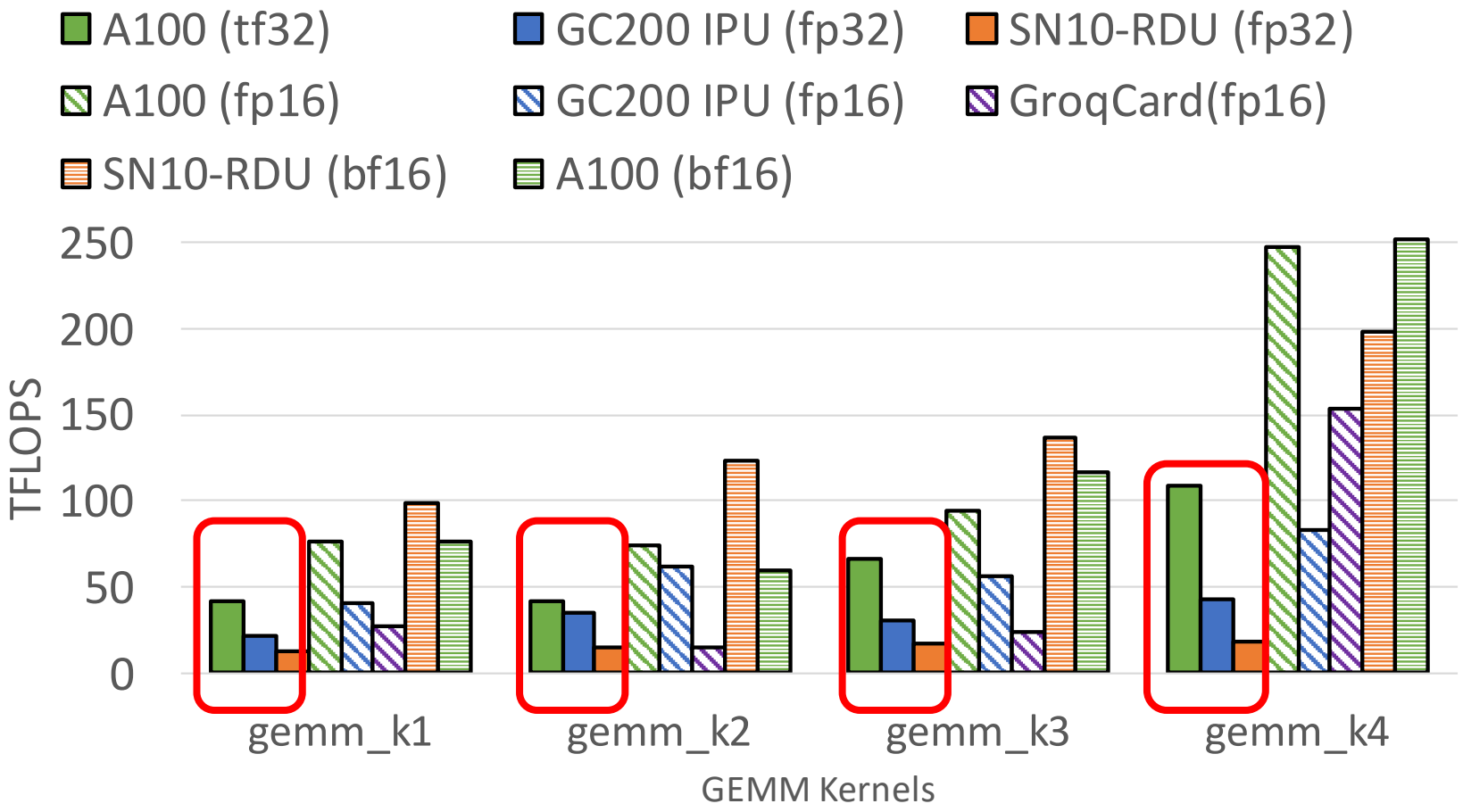
DL primitives - GEMM



Name	M	N	K
gemm_k1	64	1760	1760
gemm_k2	2560	64	2560
gemm_k3	1760	128	1760
gemm_k4	2560	2560	2560

Kernels chosen from DeepBench

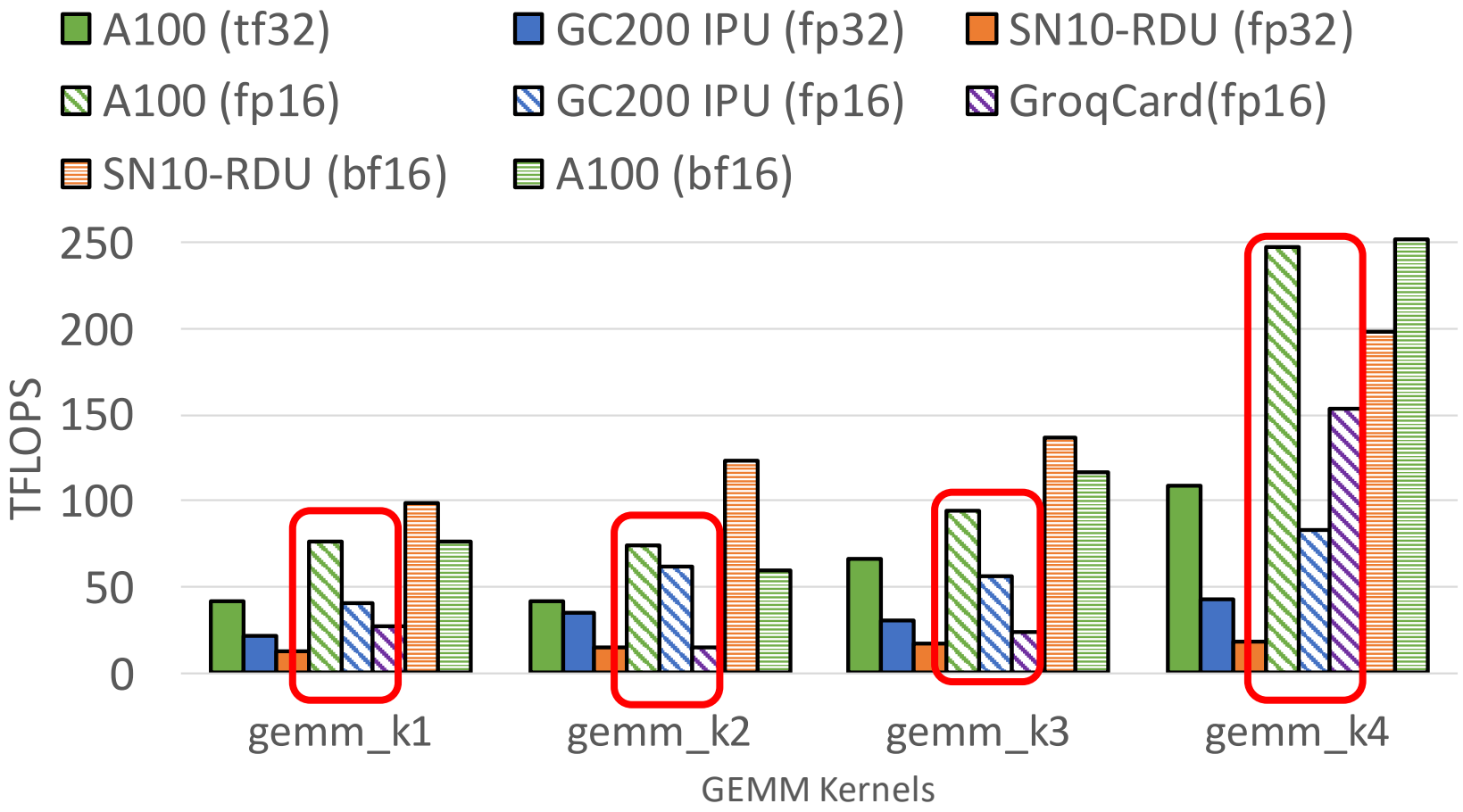
DL primitives - GEMM



Name	M	N	K
gemm_k1	64	1760	1760
gemm_k2	2560	64	2560
gemm_k3	1760	128	1760
gemm_k4	2560	2560	2560

- A100 reported highest FLOPS for full precision

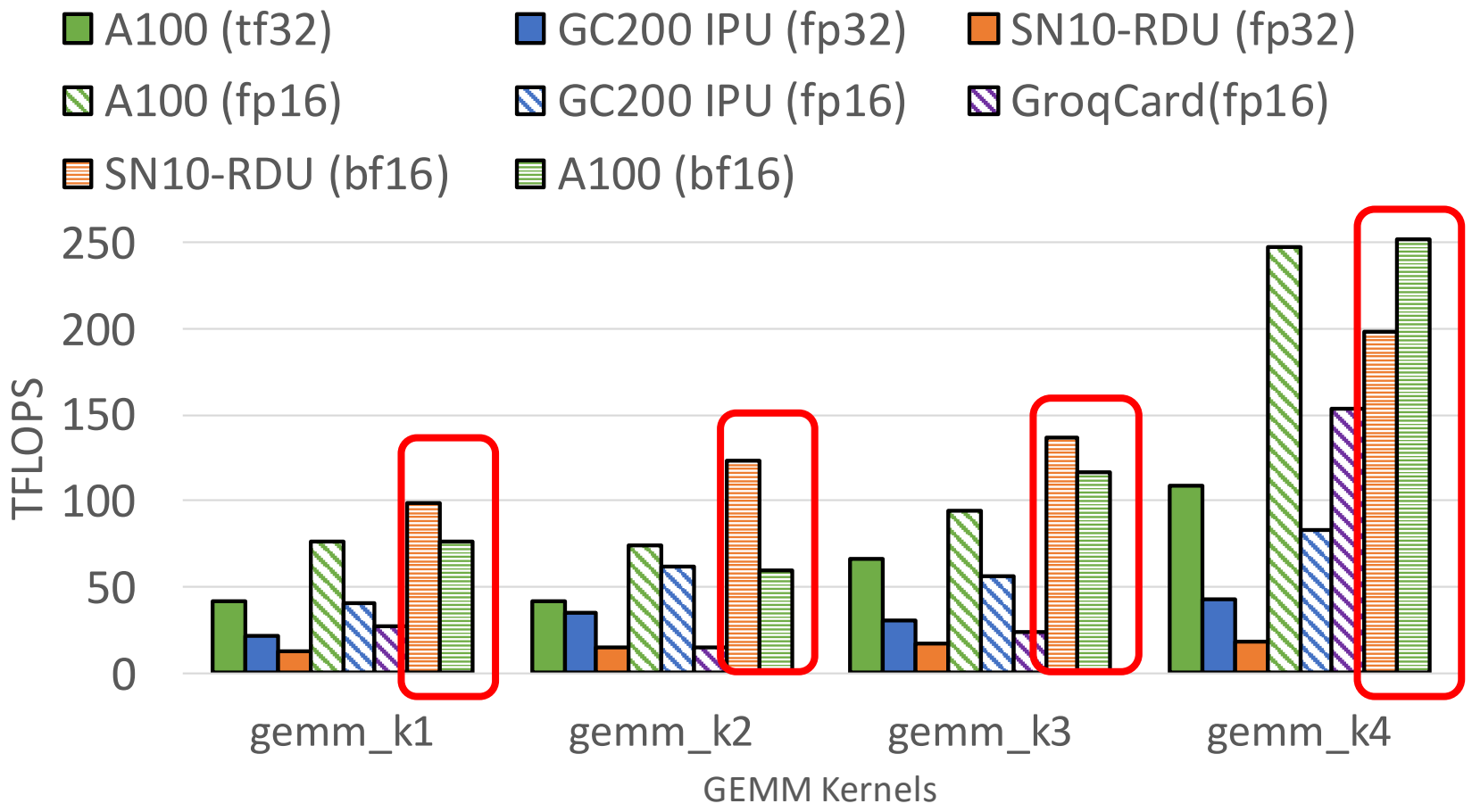
DL primitives - GEMM



Name	M	N	K
gemm_k1	64	1760	1760
gemm_k2	2560	64	2560
gemm_k3	1760	128	1760
gemm_k4	2560	2560	2560

- A100 reported highest FLOPS for half precision (fp16)

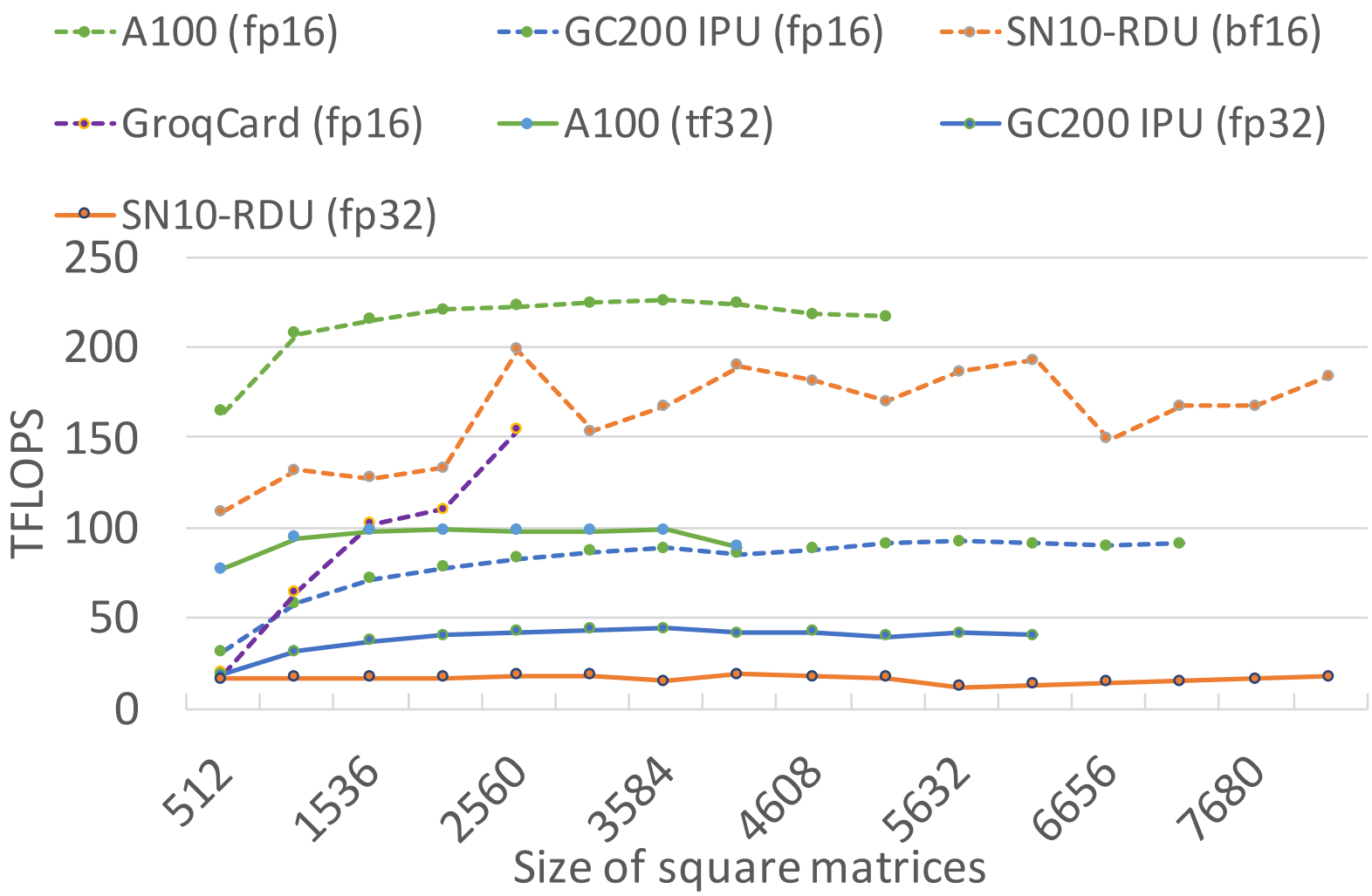
DL primitives - GEMM



Name	M	N	K
gemm_k1	64	1760	1760
gemm_k2	2560	64	2560
gemm_k3	1760	128	1760
gemm_k4	2560	2560	2560

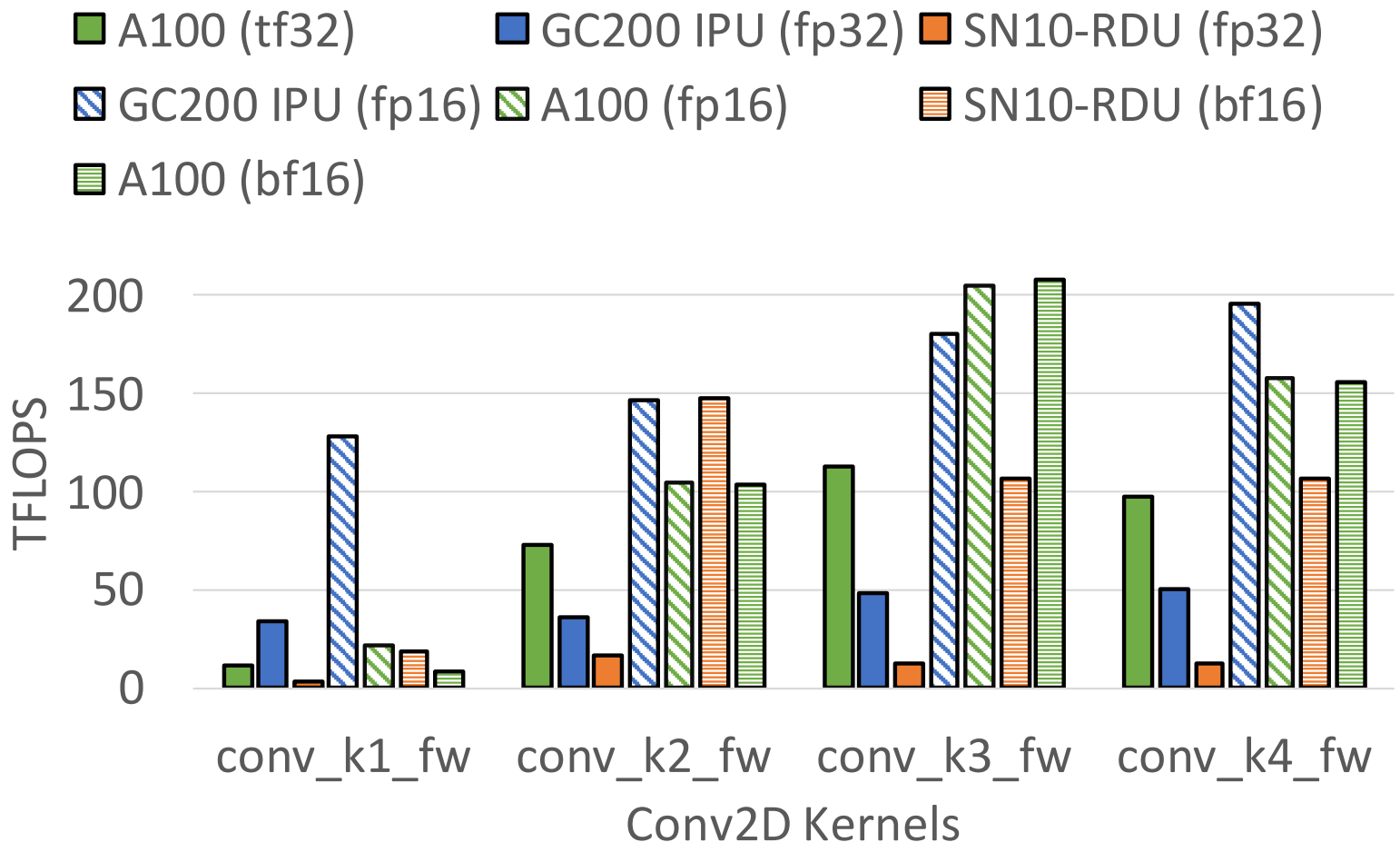
- SN reported highest TFLOPS for bf16 except for large matrix sizes

DL primitives - GEMM Scaling



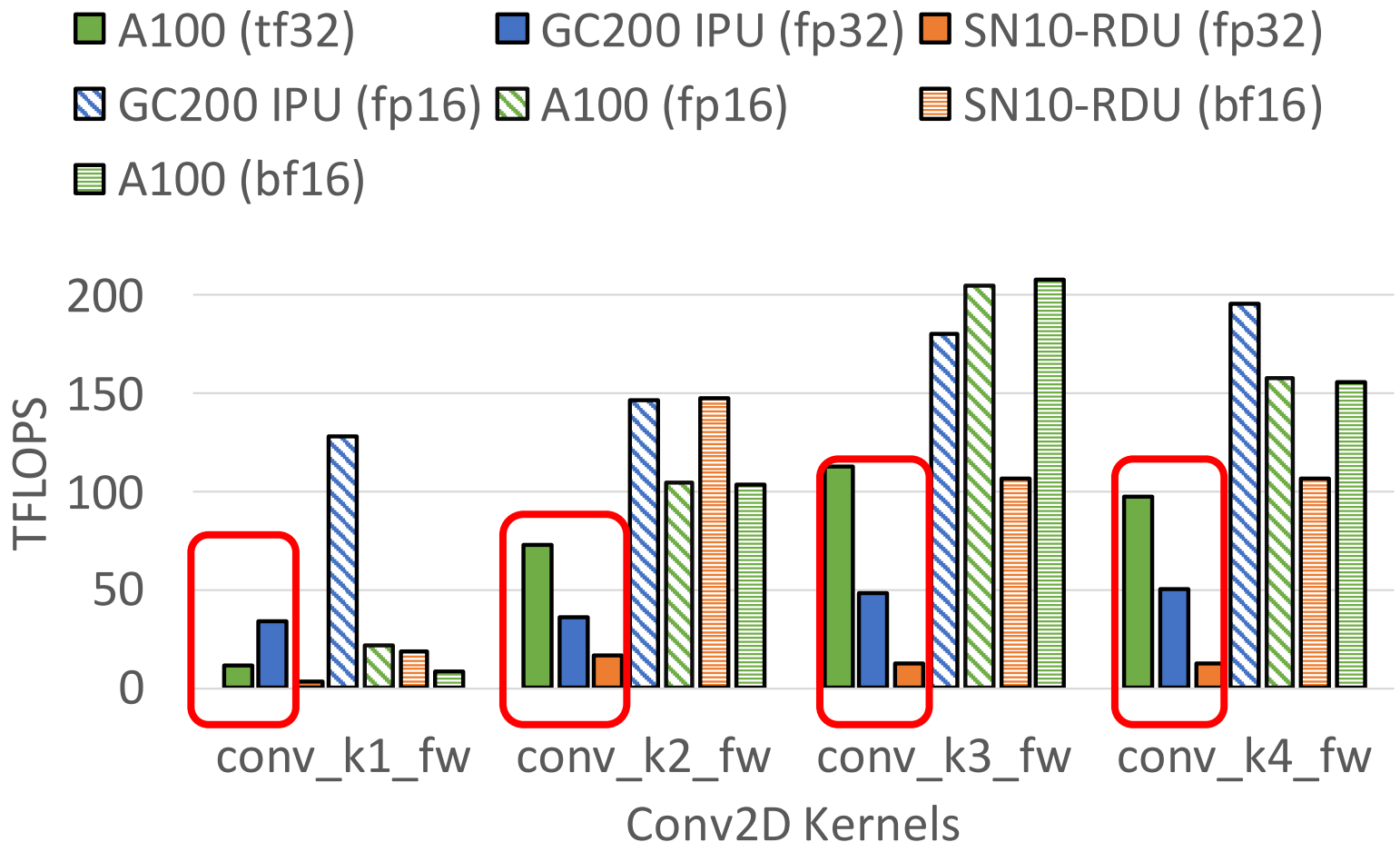
- Increase matrix sizes to saturate on-chip memory
- SN can run larger matrix sizes due to highest memory capacity

DL Primitives – Conv2D (Training)



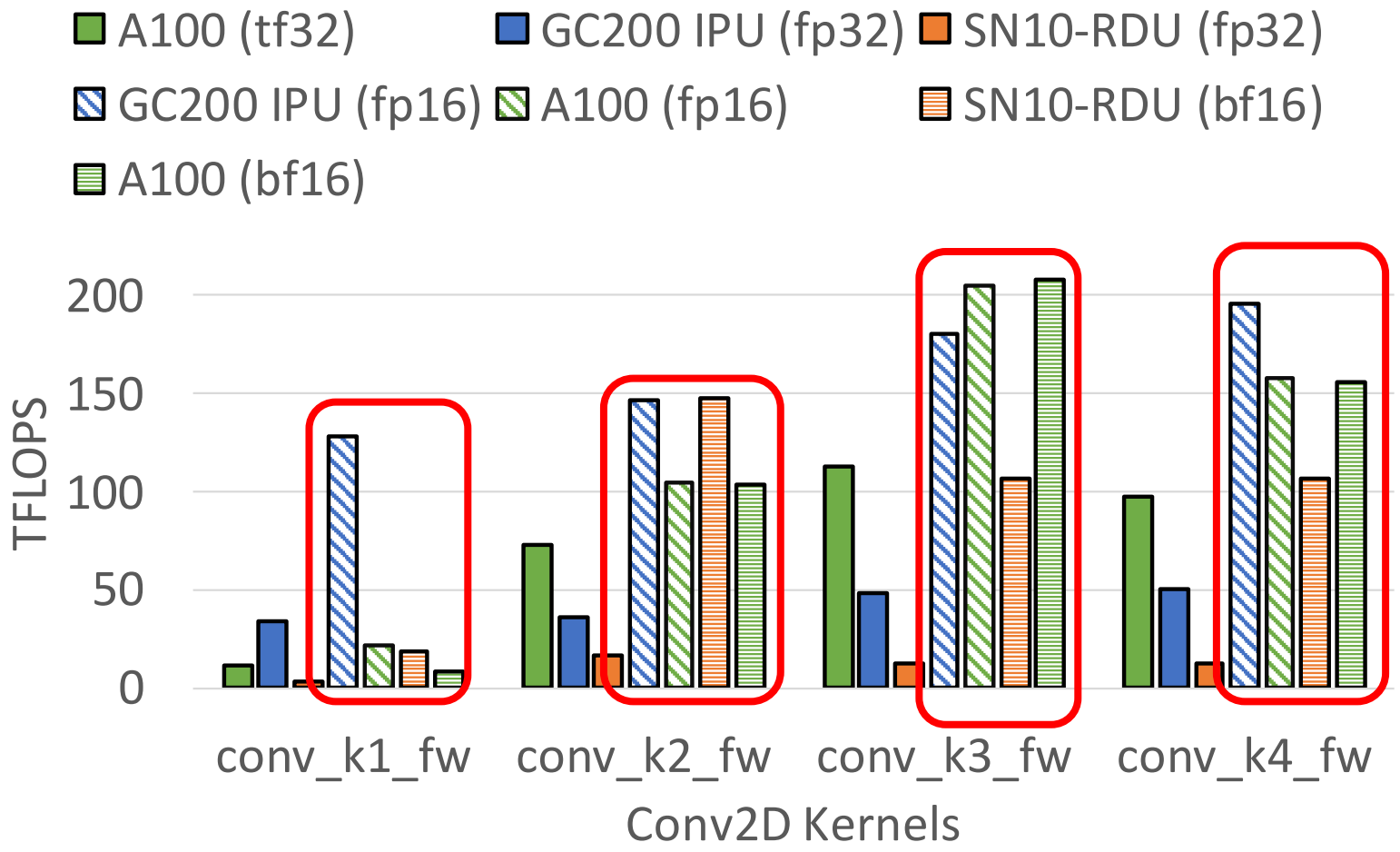
Kernels conv_k1_fw and conv_k2_fw are memory-bound, whereas conv_k3_fw and conv_k4_fw are compute-bound

DL Primitives – Conv2D (Training)



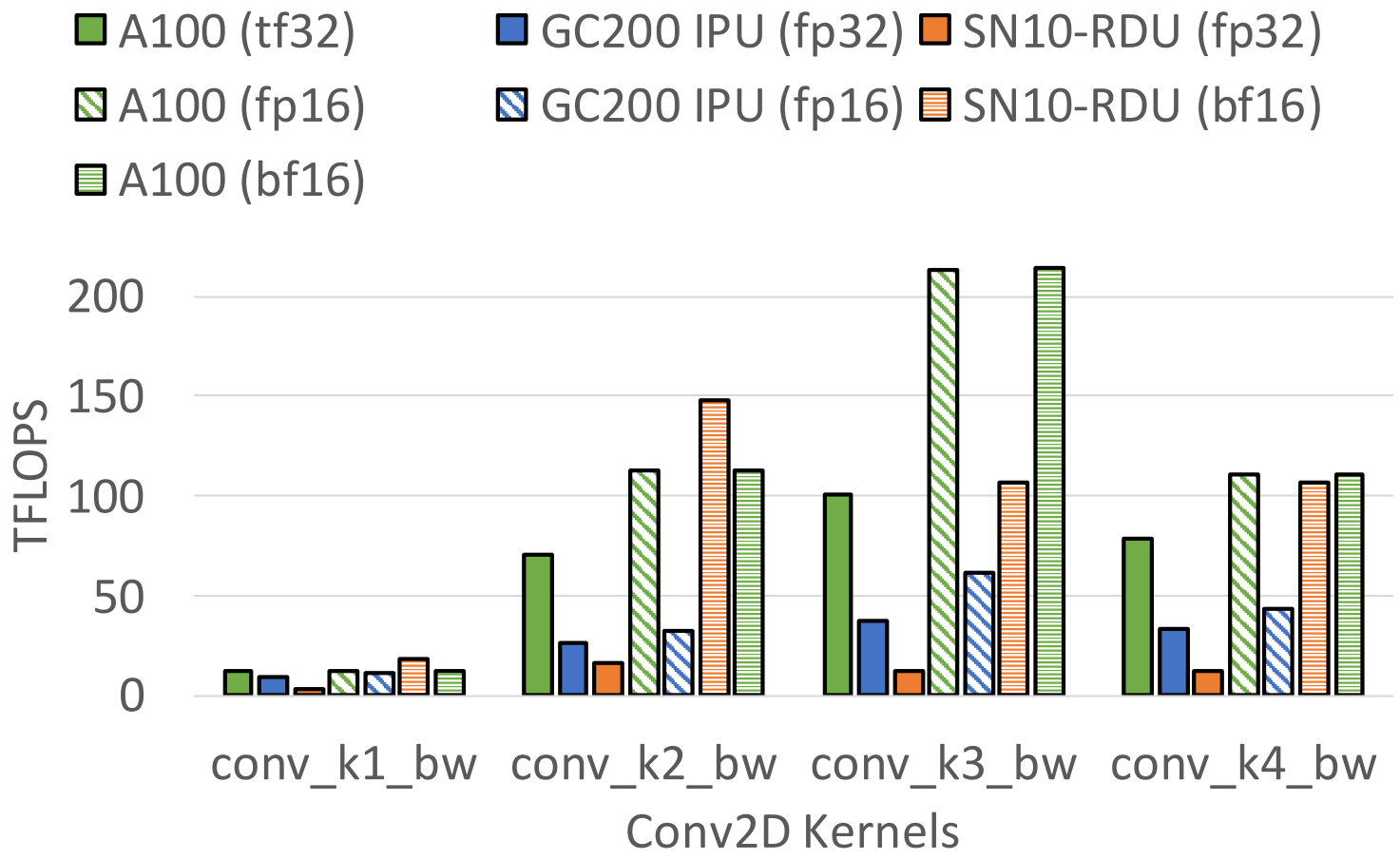
- A100 accelerates compute-intensive convolution operations better
- GC200 IPU is more sensitive to the data format in the Conv2D kernel

DL Primitives – Conv2D (Training)



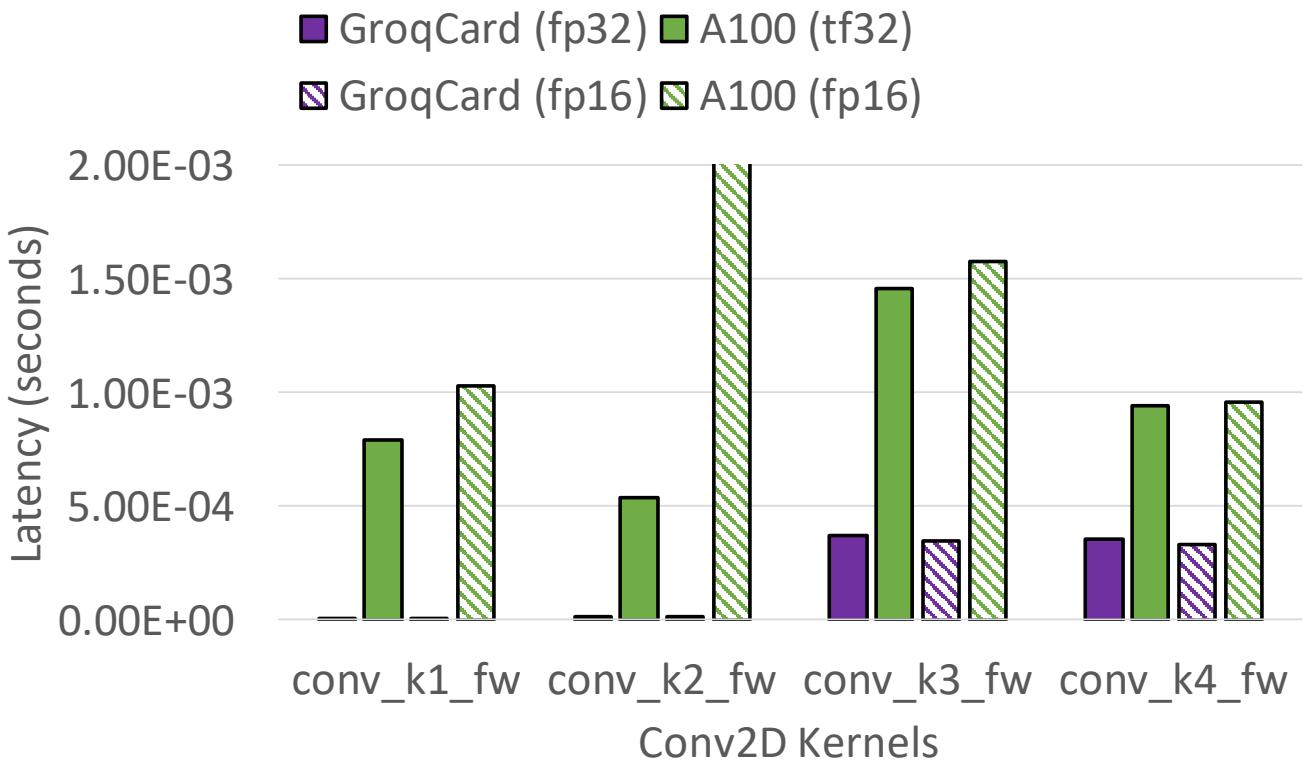
- A100 accelerates compute-intensive convolution operations better
- GC200 IPU fares better with half-precision run

DL Primitives – Conv2D (Training)



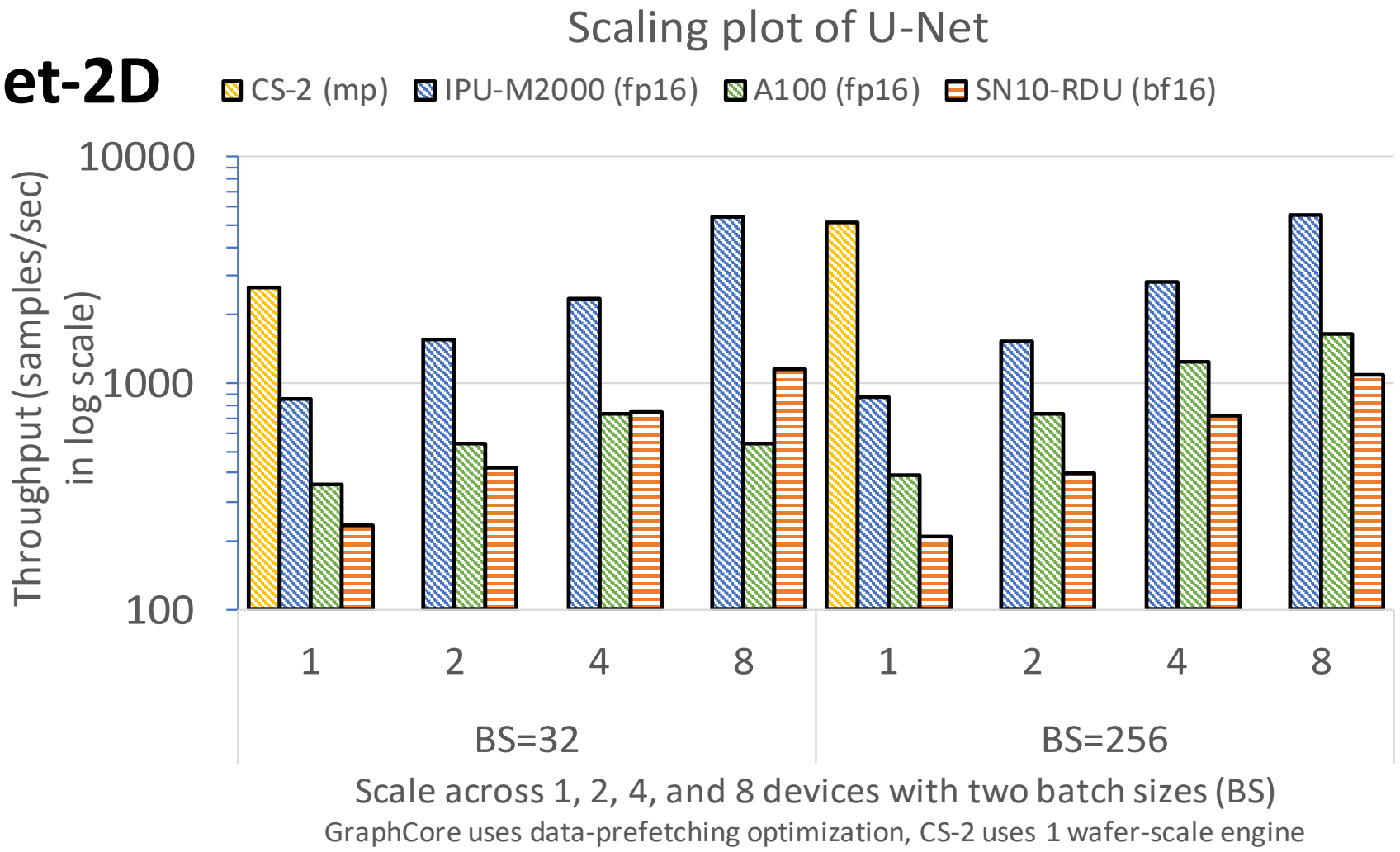
For the backward pass in training mode, A100 performs best on conv_k2_bw and conv_k3_bw, SN10 RDU performs best on conv_k1_bw and conv_k4_bw kernels.

DL Primitives – Conv2D (Inference)



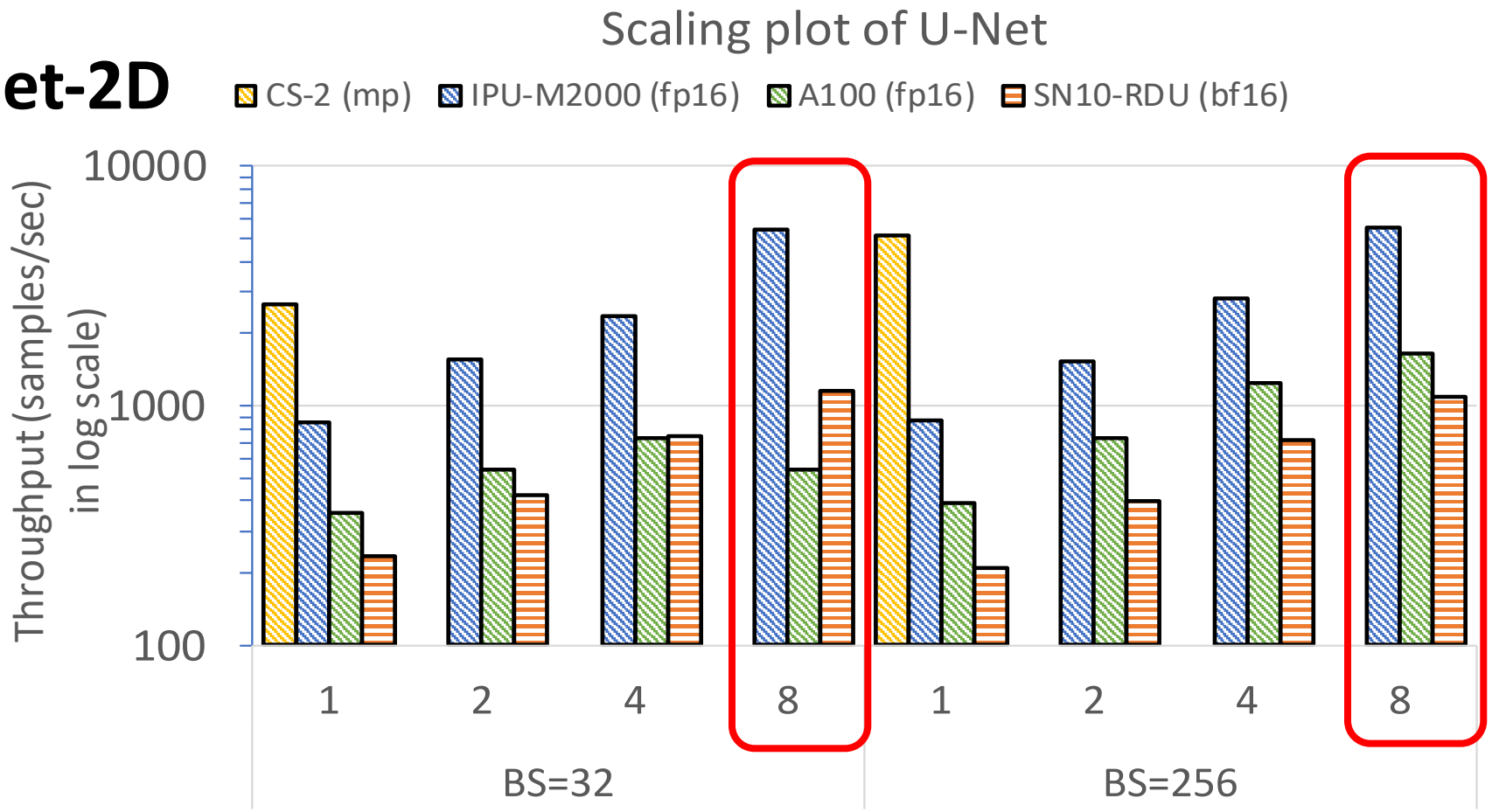
- GroqCard reported at least 2.8x, upto two orders lower latency than A100
- Dedicated MXM planes for matrix multiplications and the VXM for bitwise multiplications
- Dataflow pipelines avoid write-backs to memory and allow for optimized performance.

UNet-2D



- 256x256 image size BrainMRI image dataset
- All evaluated AI systems can run U-Net with much larger image sizes
- A100, SN10-RDU - PyTorch,
- IPU-M2000 - TensorFlow
- CS-2 - TF Estimator

UNet-2D



Scale across 1, 2, 4, and 8 devices with two batch sizes (BS)

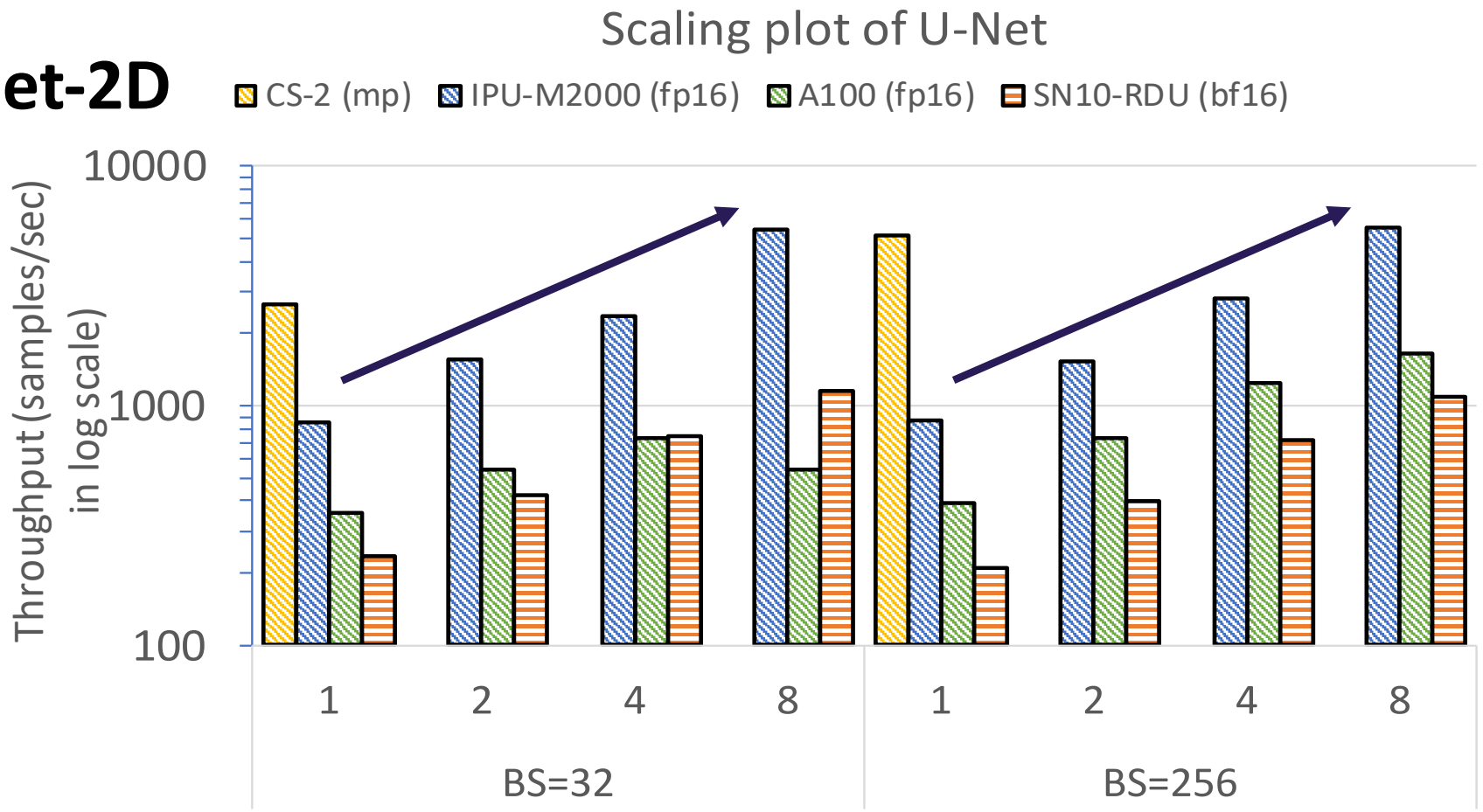
GraphCore uses data-prefetching optimization, CS-2 uses 1 wafer-scale engine

Throughput improvement Over 8 A100s

Batch size	8 SN10-RDUs	1 CS-2	8 GC 200 IPUs
32	2.1x	4.9x	10x
256	0.7x*	3.1x	3.3x

*2.1x in latest sw release

UNet-2D



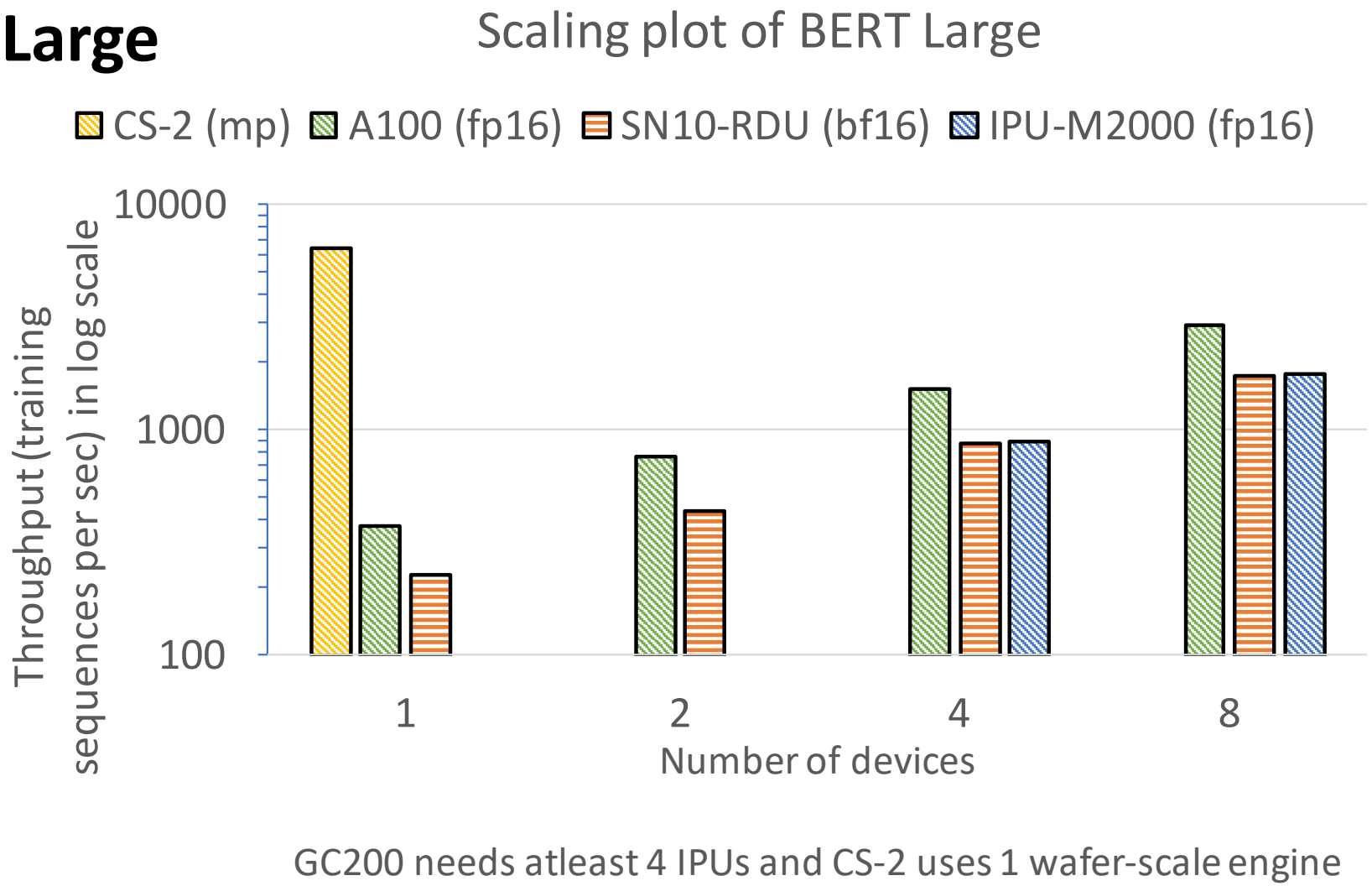
Scale across 1, 2, 4, and 8 devices with two batch sizes (BS)

GraphCore uses data-prefetching optimization, CS-2 uses 1 wafer-scale engine

Scaling efficiencies

Batch size	A100	SN10-RDUs	GC200 IPUs
32	18.8%	42%	79.6%
256	52%	28%	79.5%

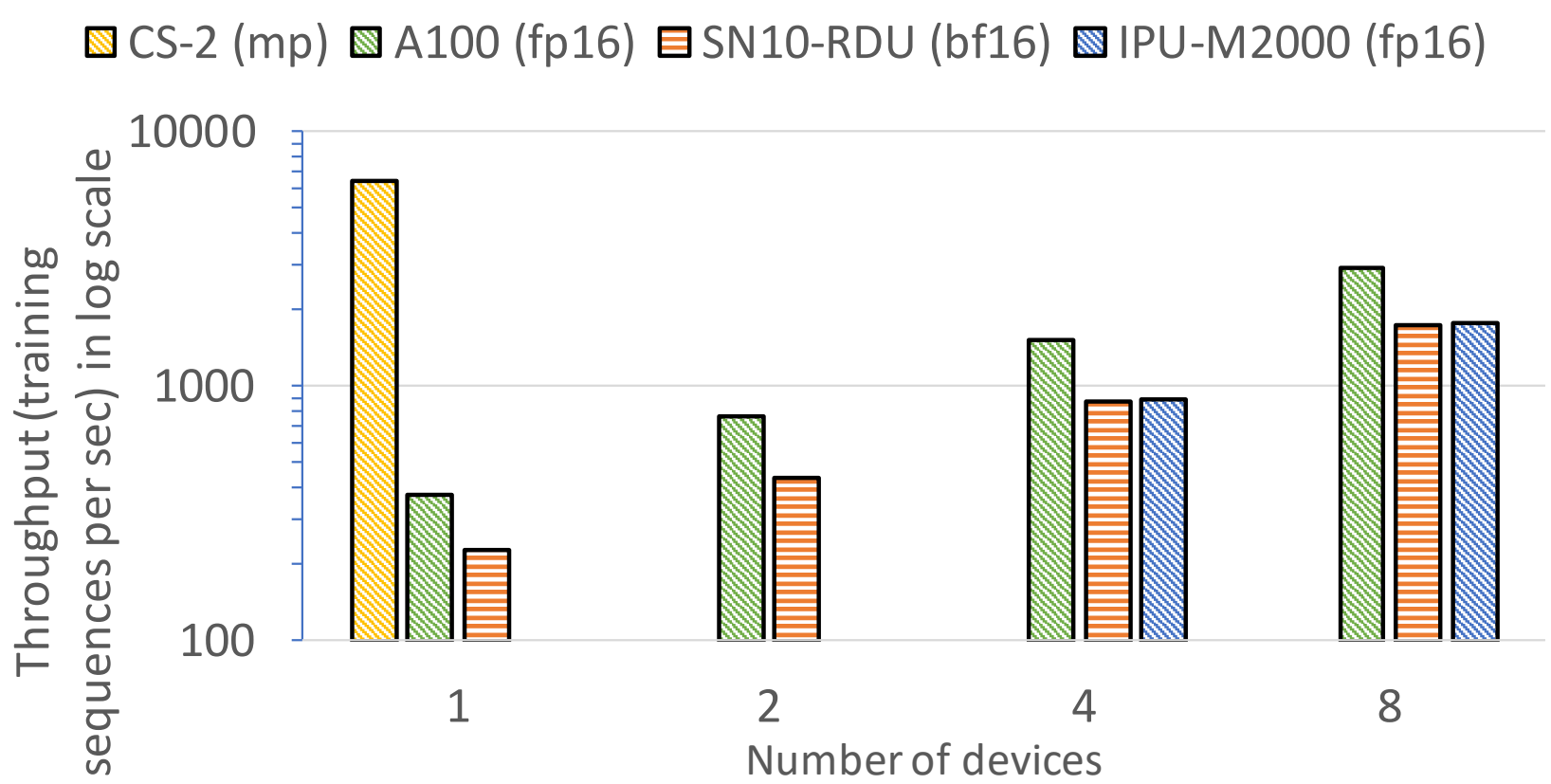
BERT Large



pretraining phase of the BERT-Large model with Wikipedia and BookCorpus datasets.

global batch size = 256, maxim sequence length (MSL) = 128

BERT Large

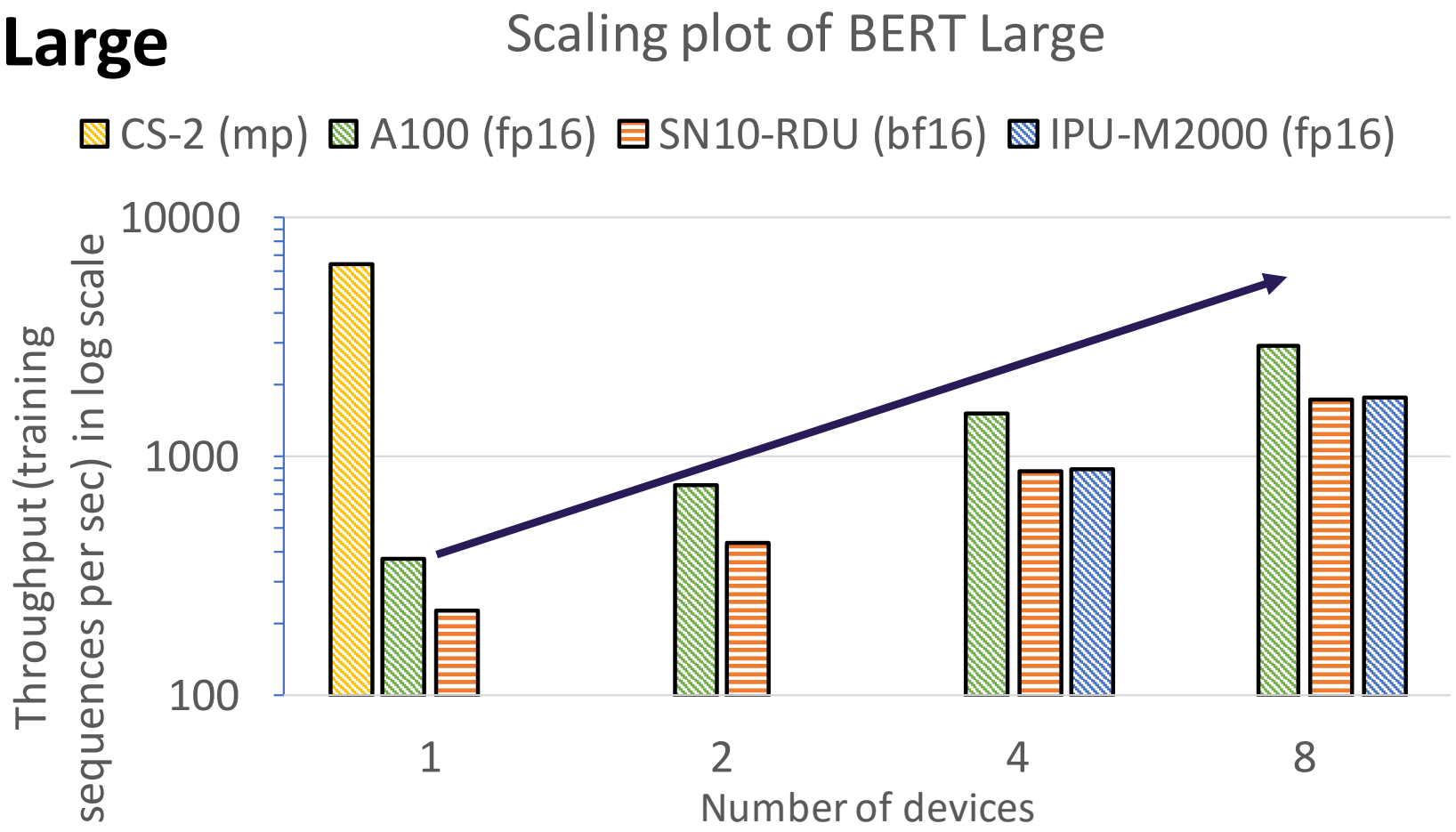


GC200 needs atleast 4 IPUs and CS-2 uses 1 wafer-scale engine

Throughput improvement Over 8 A100s

Batch size	8 SN10-RDUs	1 CS-2	8 GC 200 IPUs
256	0.67x	2.37x	0.61x

BERT Large



GC200 needs atleast 4 IPUs and CS-2 uses 1 wafer-scale engine

Scaling efficiencies

Batch size	A100	SN10-RDUs	GC200 IPUs
256	97%	93%	100%

BERT

For inference mode runs (DistilBERT)

Latency improvement Over A100

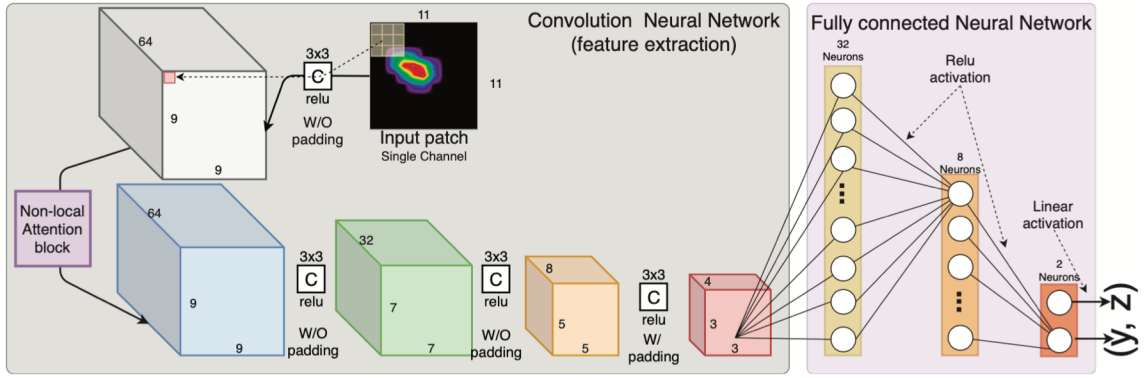
Batch size	GC200 IPU	GroqCard
1	9x	13x

Fast X-Ray Bragg Peak Analysis

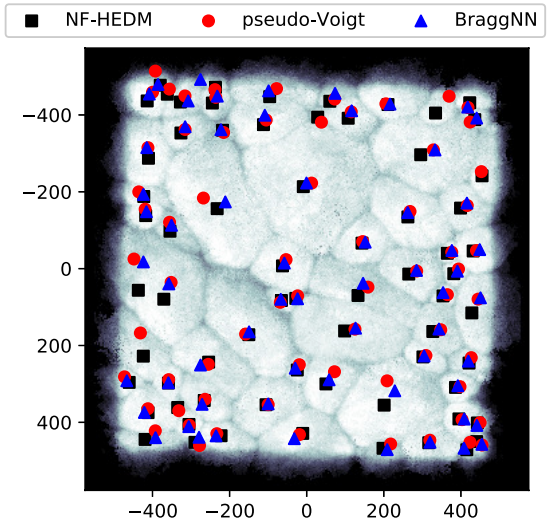
Goal: Enable rapid analysis and real-time feedback during an in-situ experiment with complex detector technologies

Proposed Approach: Deep learning-based method, BraggNN, for massive extraction of precise Bragg peak locations from far-field high energy diffraction microscopy data. BraggNN has achieved 200X improvement over conventional pseudo-Voight profiling

Challenges: Model training capability is limited by the hardware



Application of the BraggNN deep neural network to an input patch yields a peak center position (y, z). All convolutions are 2D of size 3 × 3, with rectifier as activation function. Each fully connected layer, except for the output layer, also has a rectifier activation function.



A comparison of BraggNN, pseudo-Voigt FF-HEDM and NF-HEDM. (a) Grain positions from NF-HEDM (black squares), pseudo-Voigt FF-HEDM (red circles) and BraggNN FF-HEDM (blue triangles) overlaid on NF-HEDM confidence map

Courtesy: Z. Liu et al. BraggNN: Fast X-ray Bragg Peak Analysis Using Deep Learning. International Union of Crystallography (IUCrJ), Vol. 9, No. 1, 2022

Fast X-Ray Bragg Peak Analysis

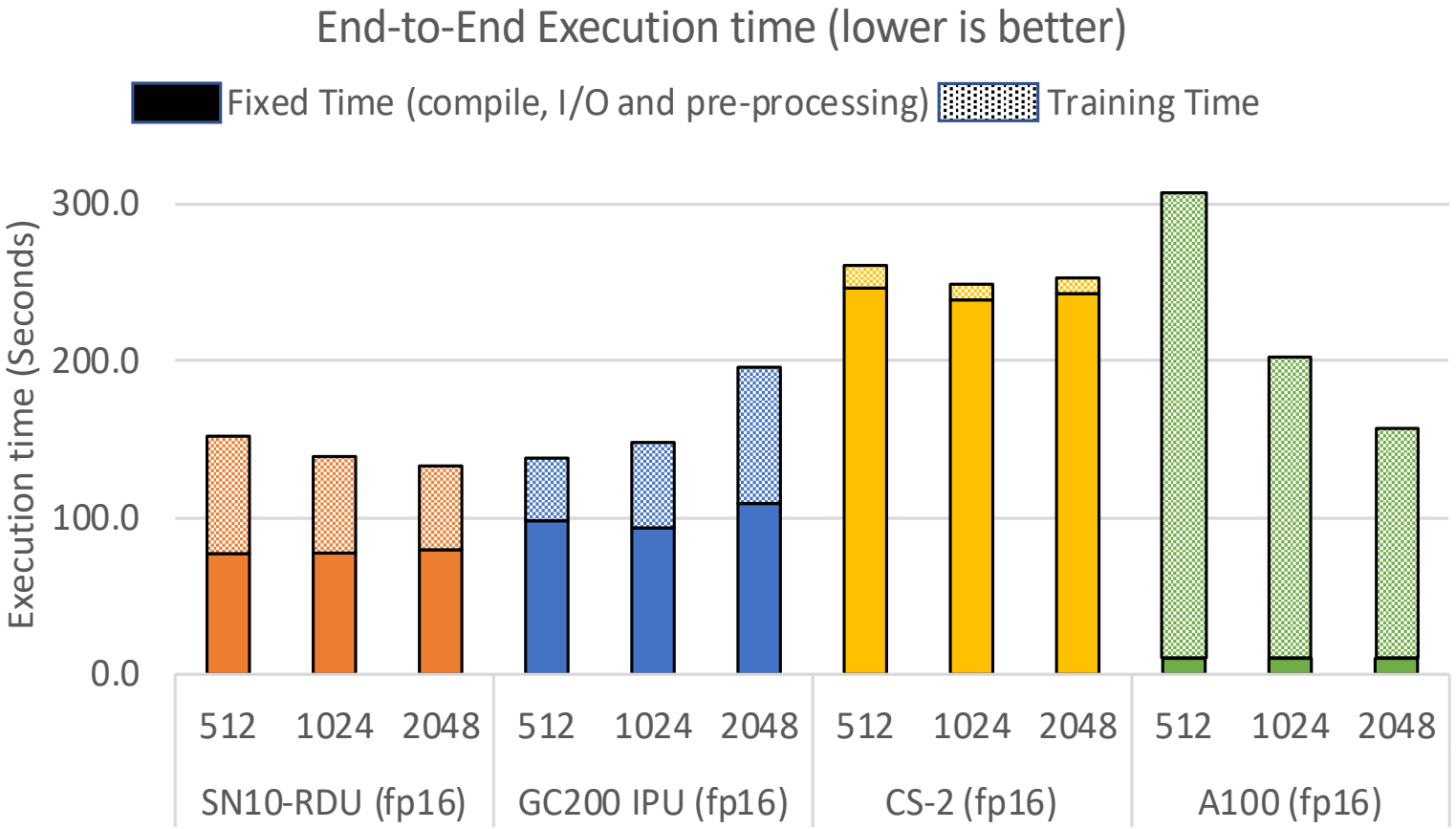


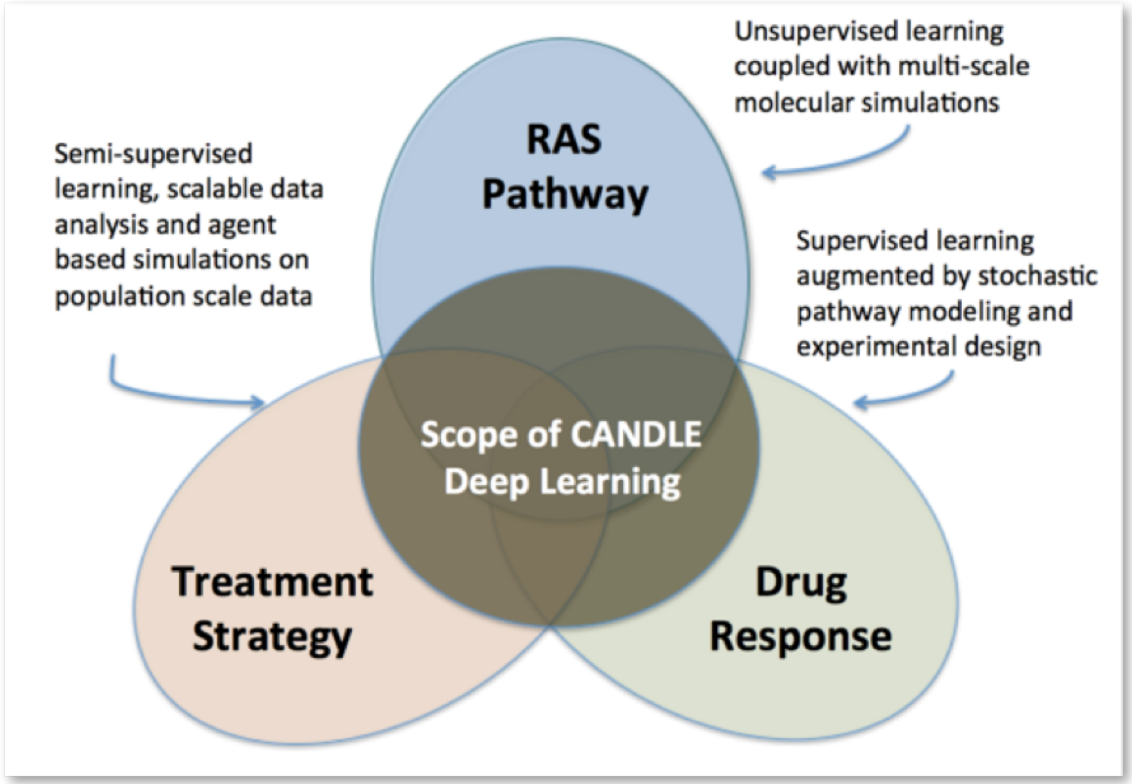
TABLE II: BraggNN Throughput (in order of 1k samples/sec) with various batch sizes (BS)

System	BS=512	BS=1024	BS=2048
CS-2 (FP16)	1365.4	2463	2787.9
GC200 IPU (FP16)	478.0	350.6	219.9
SN10 RDU (BF16)	369.7	449.8	518
A100 (FP16)	53.9	65.5	73.7

- SambaNova and Graphcore achieve lowest time to solution and achieve up to 1.55x and 1.46x speedup in comparison to Nvidia A100 respectively.
- Cerebras achieves up to 37.8x throughput improvement over A100.

Drug Discovery - Uno

- CANDLE: Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer
- Implement deep learning architectures that are relevant to problems in cancer.
- Focus on “Uno” application which aims to predict the drug response based on molecular features of tumor cells and drug descriptors.



Drug Discovery - Uno

- Model has small memory footprint, however, the large data set stresses the I/O

TABLE III: Uno Performance Evaluation with Full Dataset

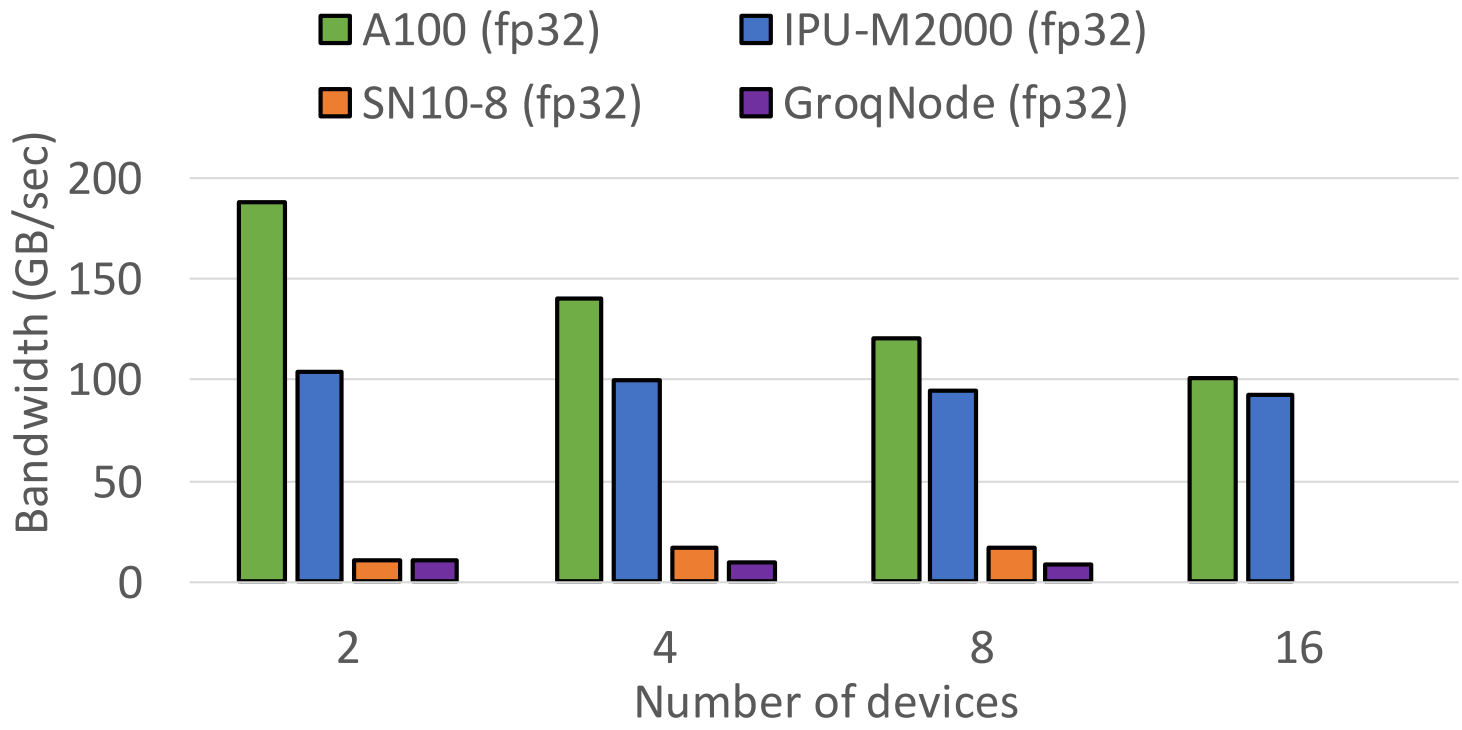
System	#Units	Batch size	Throughput (samples/sec)
CS-2 (mp)	1 CS2 WSE	2000	872258.7
GC200 IPU (FP16)	1 IPU	512	46123
SN10-8 (BF16)	2 RDUs	16	31958
A100 (TF32)	1 GPU	512	7567

Throughput improvement Over 1 A100s

SN10-8	IPU-M2000	CS-2
4.2x	6x	115x

- Evaluation with same hyper-parameters is work in progress

Collective Communication Bandwidth



DeepBench and OSU MPI Benchmarks used for the all_reduce communication evaluation and we scale the number of devices to 16. We use up to 8 devices for Groq and SambaNova

Nvidia DGX3 achieves higher All Reduce performance in comparison to other AI systems

Observations, Challenges and Insights

- Significant speedup achieved for a wide-gamut of scientific ML applications
  - Easier to deal with larger resolution data and to scale to multi-chip systems
- Room for improvement exists
  - Porting efforts and compilation times
  - Coverage of DL frameworks, support for performance analysis tools, debuggers
- Good progress made in integration of AI accelerators, in production, at a national user facility and significant more work is needed for effective coupling
- Training and Outreach is critical to educate users to effectively use AI systems
- Close collaboration with vendors is necessary to realize the vision of AI for science

Thank You

- This research was funded in part and used resources of the Argonne Leadership Computing Facility (ALCF), a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

- Venkatram Vishwanath, Michael Papka, William Arnold, Bruce Wilson, Varuni Sastry, Sid Raskar, Corey Adams, Rajeev Thakur, Anthony Avarca, Arvind Ramanathan, Alex Brace, Zhengchun Liu, Hyunseung (Harry) Yoo, Ryan Aydelott, Sid Raskar, Zhen Xie, Kyle Felker, Craig Stacey, Tom Brettin, Rick Stevens, and many others have contributed to this material.

- Our current AI testbed system vendors – Cerebras, Graphcore, Groq, Intel Habana and SambaNova. There are ongoing engagements with other vendors.

Please reach out for further details

Murali Emani, memani@anl.gov