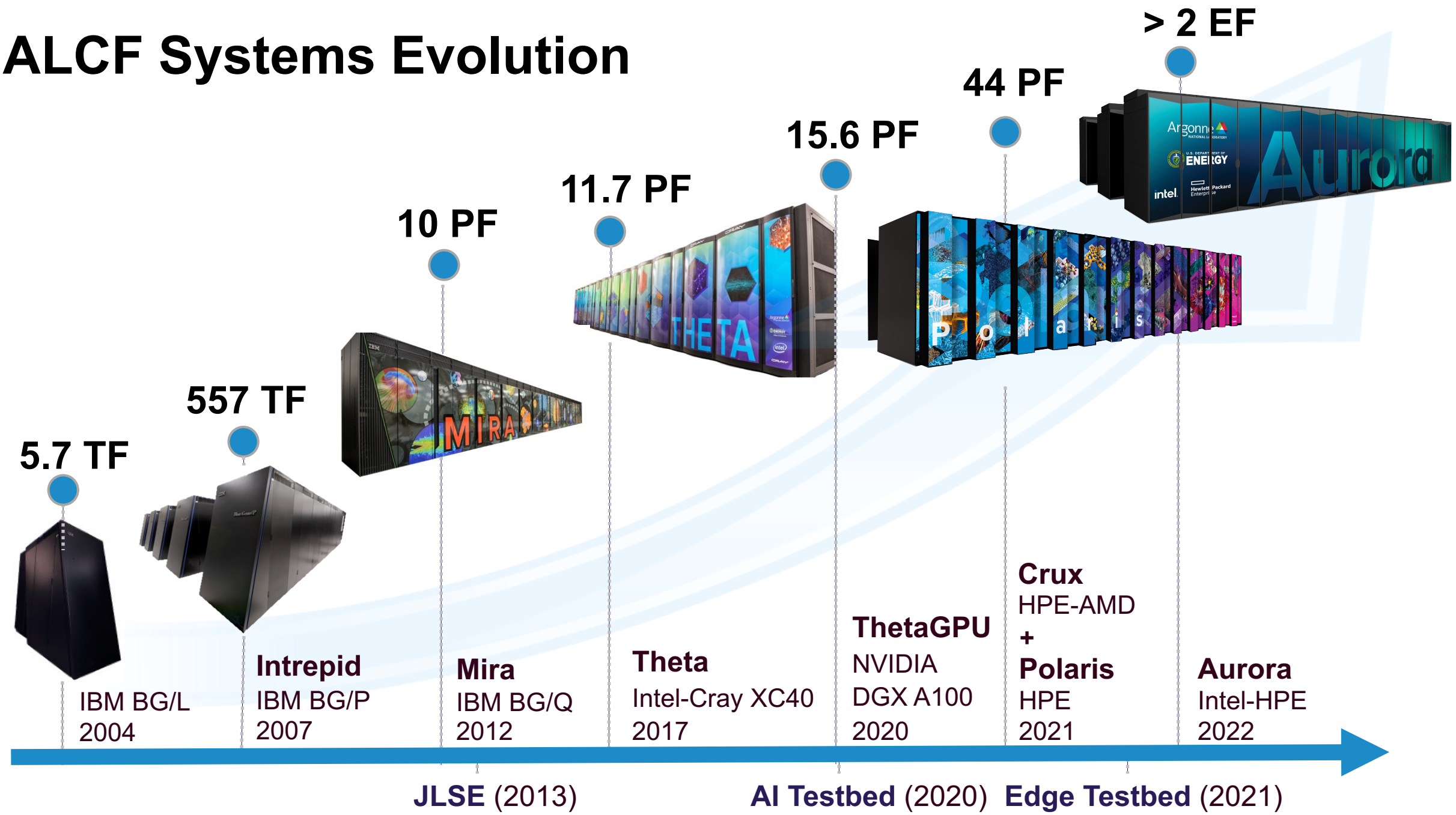
Artificial Intelligence Accelerators For Science

Venkatram Vishwanath
Argonne Leadership Computing Facility
venkat@anl.gov

November 8, 2022

ALCF Systems Evolution



AI Chip Landscape

V0.7 Dec., 2019

S.T.

▪ MLPerf results available ▪ AI-Benchmark results available

Tech Giants/System

- Google: ▪ TPU ▪ Edge TPU
- Microsoft: ▪ Brainwave
- facebook
- aws: ▪ Inferentia
- Apple: ▪ A13
- IBM
- Alibaba Group 阿里巴巴集团: ▪ Hanguang800 ▪ TG6100N
- HUAWEI HISILICON: ▪ Ascend ▪ Kirin ▪ Hi35xx
- Baidu 百度: ▪ Kunlun/ Honghu
- Tesla: ▪ FSD
- Tencent 腾讯
- Hewlett Packard Enterprise
- FUJITSU: ▪ DLU
- DELL
- inspur 浪潮
- Western Digital
- NOKIA
- LG

IC Vendor/Fabless

- intel: ▪ NNP-I ▪ NNP-T/Myriad X/ EyeQ/Arria FPGA
- SAMSUNG: ▪ Exynos 9825
- NVIDIA: ▪ Volta/Turing /T4/Xavier/NVDLA
- QUALCOMM: ▪ Snapdragon 855 ▪ Cloud AI 100
- AMD: ▪ EPYC
- XILINX: ▪ VERSAL
- MEDIATEK: ▪ Dimensity
- UNISOC: ▪ Tiger T710
- MARVELL
- Rockchip 瑞芯微电子: ▪ RK3399Pro
- Ambarella: ▪ CV22S/25S
- NationalChip: ▪ GX8010

Automated Driving

- NXP
- Texas Instruments
- RENESAS
- TOSHIBA
- ST

Startup in China

- Cambricon 寒武纪科技: ▪ MLU100/270/220
- 地平线 Horizon Robotics: ▪ Journey
- BITMAIN: ▪ BM1682/1880
- intelliFusion 云天励飞: ▪ DeepEye1000
- 依图 YITU: ▪ QuestCore
- 耐能 NextVPU: ▪ N171
- Canaan: ▪ K210
- artosyn 芯驰微电子: ▪ AR9000
- 探境科技 INTENGINE: ▪ Voitist611
- TSING MICRO: ▪ TX101/210/510
- 亿智科技: ▪ TAi8010
- BLACK SESAME TECHNOLOGIES: ▪ HuaShan
- Enflame: ▪ DTU
- WITHMEM 忆芯科技: ▪ MemCore001

Smart Voice

- ChipIntelli: ▪ CI100X/110X
- ▪ CSK4002
- 云知声 Unisound
- AISPEECH 思必驰 专注人性化的智能语音

Startup Worldwide

- Cerebras: ▪ WSE
- WAVE COMPUTING
- Graphcore: ▪ GC2
- SambaNova SYSTEMS
- habana: ▪ Gaudi ▪ Goya
- HAILO: ▪ Hailo-8
- blaize: Xplorer X1000
- KALRAY: ▪ MPPA2-256
- groq
- Tachyum
- Esperanto TECHNOLOGIES
- PEZY Computing: ▪ PEZY-SC2
- Eta Compute: ▪ ECM3531
- GREENWAVES TECHNOLOGIES: ▪ GAP8

FPGA/eFPGA

- Achronix
- flexlogix
- EFINIX

Processing in Memory

- MYTHIC
- SYNTIANT
- AREANNA
- gyrfalcon technology: ▪ GAINBOARD/ Lightspeeur
- ANAFLASH

Optical Computing

- LIGHTELLIGENCE
- LIGHTMATTER
- Optalysys
- LUMINOUS

Neuromorphic

- brainchip
- aiCTX
- RAIN
- GML GrAI Matter Labs
- ABR

- Preferred Networks: ▪ MN-Core
- INNOGRIT: ▪ Shasta/Rainier/Tacoma
- Kneron: ▪ KL520
- AISTORM
- NeuroBlade
- NOVUMIND
- Tenstorrent
- AIMOTIVE

More at https://basicmi.github.io/AI-Chip/

IP/Design Sevice

- arm
- SYNOPSYS
- Imagination
- CEVA
- cadence
- SiFive
- ARTERIS IP
- Moortec

Design service with In-house IP

- VeriSilicon
- BROADCOM
- GUC
- alchip
- FARADAY
- eSilicon

Compilers

- TensorFlow | XLA | MLIR
- GLOW
- tvm
- OctoML
- NVIDIA | TensorRT
- PlaidML
- nGraph
- ONNC
- Tiramisu Compiler
- The Tensor Algebra Compiler (taco)

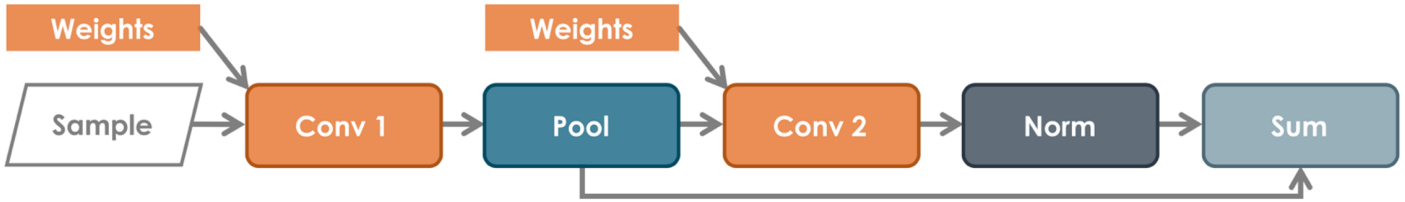
Benchmarks

- MLPerf
- AI - Benchmark
- AI Matrix.
- 中国人工智能产业发展联盟 Artificial Intelligence Industry Alliance
- DAWNBench
- MLMark An EEMBC Benchmark

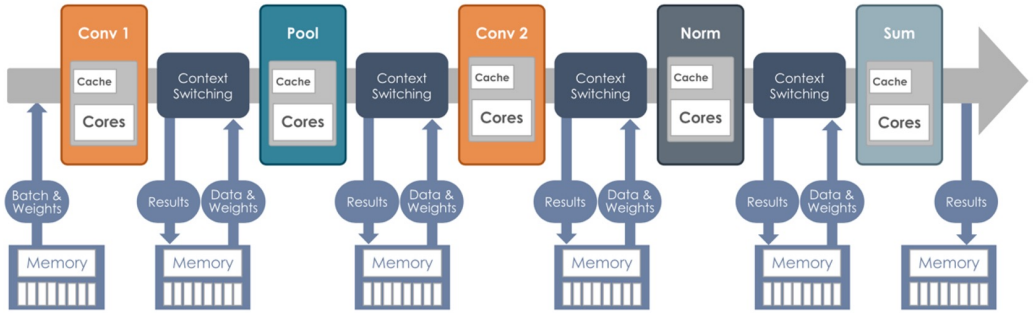
All information contained within this infographic is gathered from the internet and periodically updated, no guarantee is given that the information provided is correct, complete, and up-to-date.

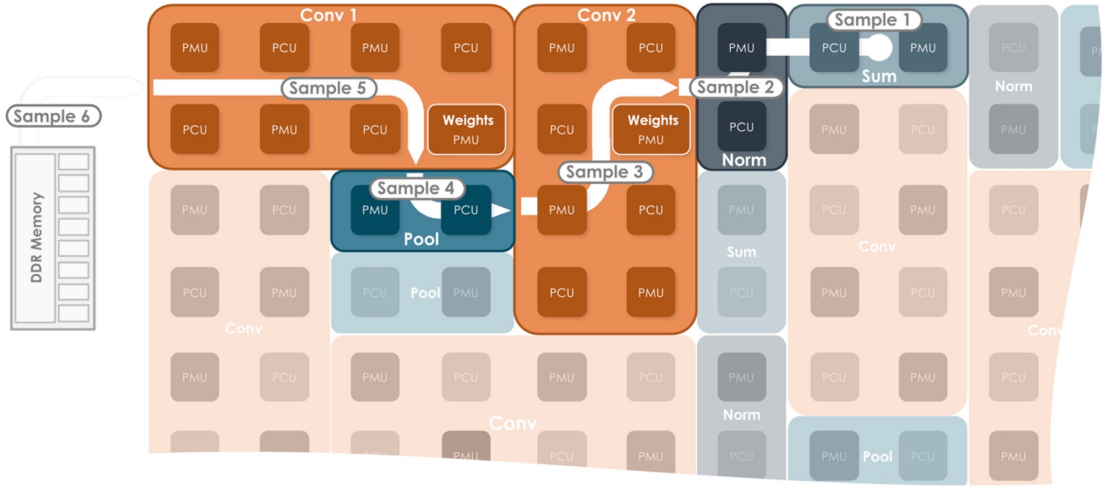
Source: https://github.com/basicmi/AI-Chip

Dataflow Architectures



Simple Convolution Graph

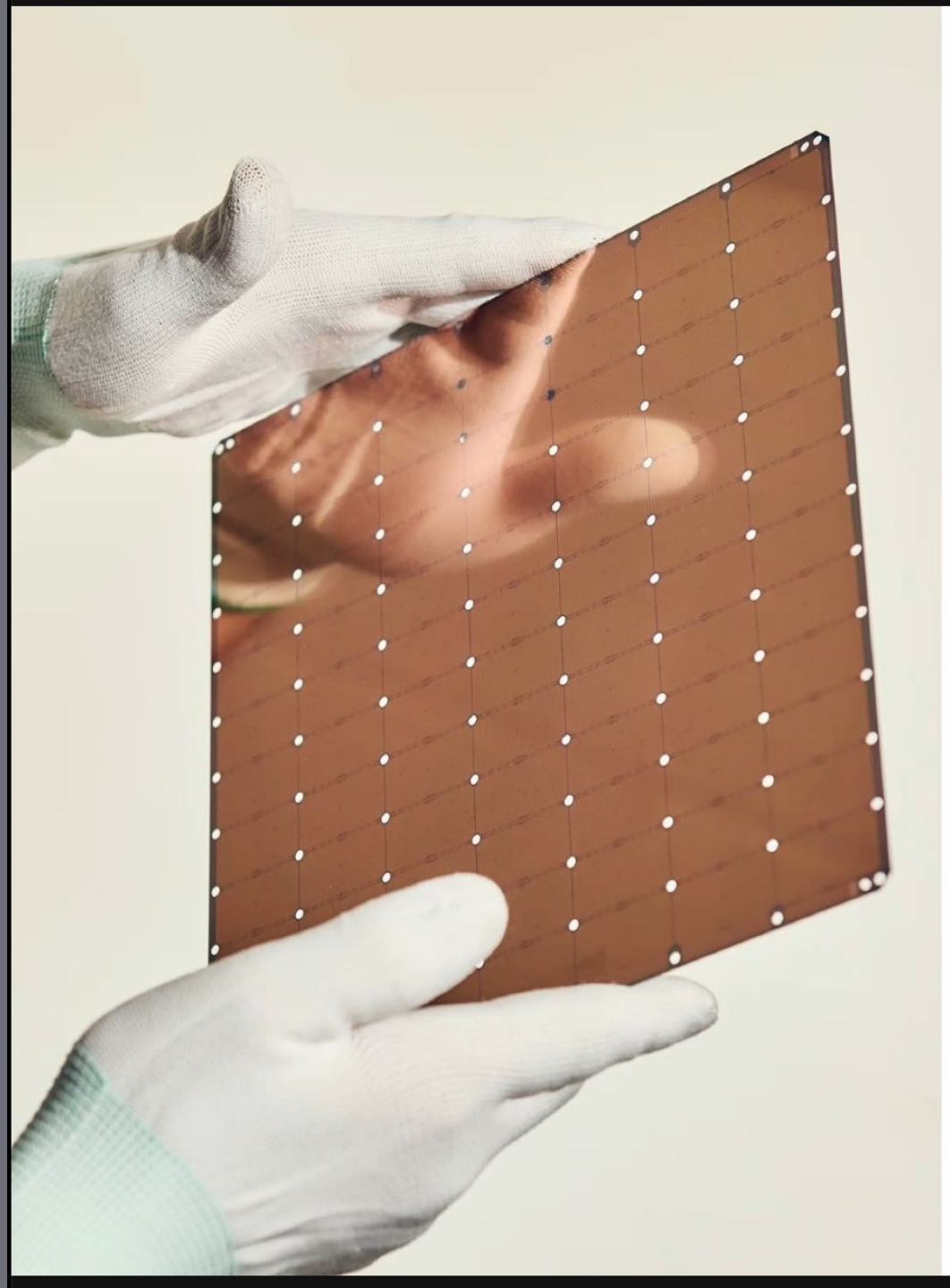




GPU accelerators: Each kernel is launched onto the device and bottlenecks include memory bandwidth and kernel-launch latencies

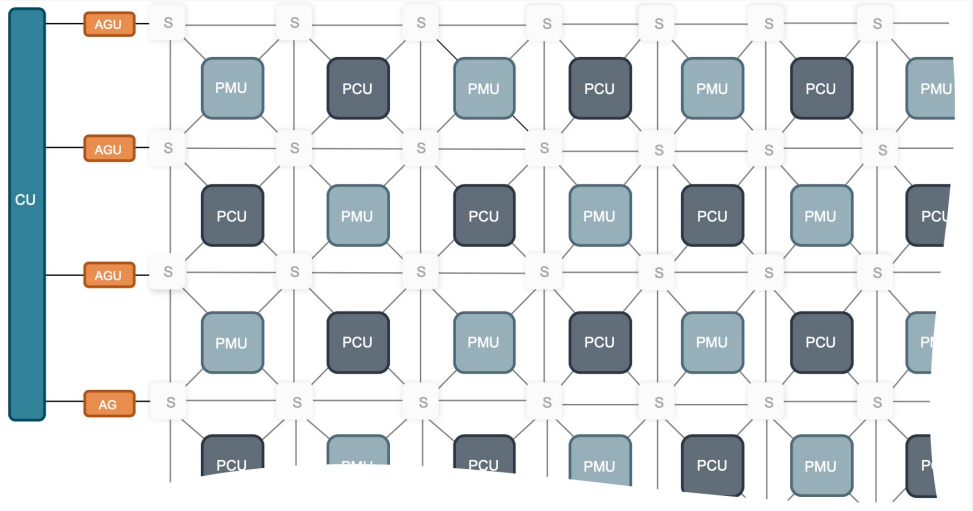
Dataflow: Kernels are spatially mapped onto the accelerator and data flows on-chip between them reducing memory traffic

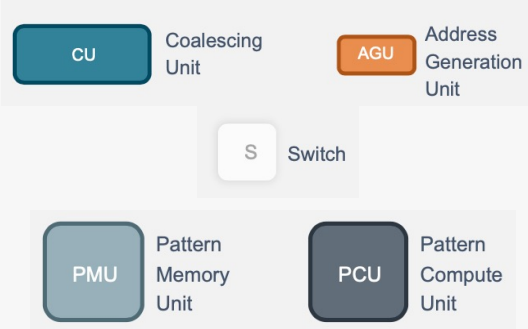
Image Courtesy: Sumti Jairath, SambaNova



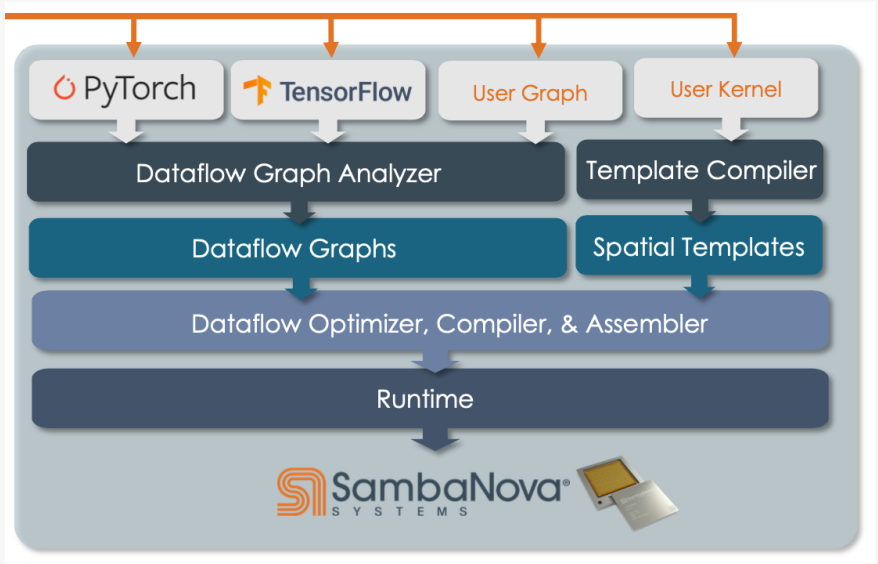
Cerebras Wafer Scale Engine

	Gen 1	Gen 2
Fabrication process	**16nm**	**7nm**
Silicon Area	**46cm²**	**46cm²**
Transistors	**1 T**	**2 T**
AI-optimized cores	**400k**	**850k**
Memory (on-chip)	**18GB**	**40GB**
Memory bandwidth	**9PB/s**	**19PB/s**
Fabric bandwidth	**100Pb/s**	**212Pb/s**





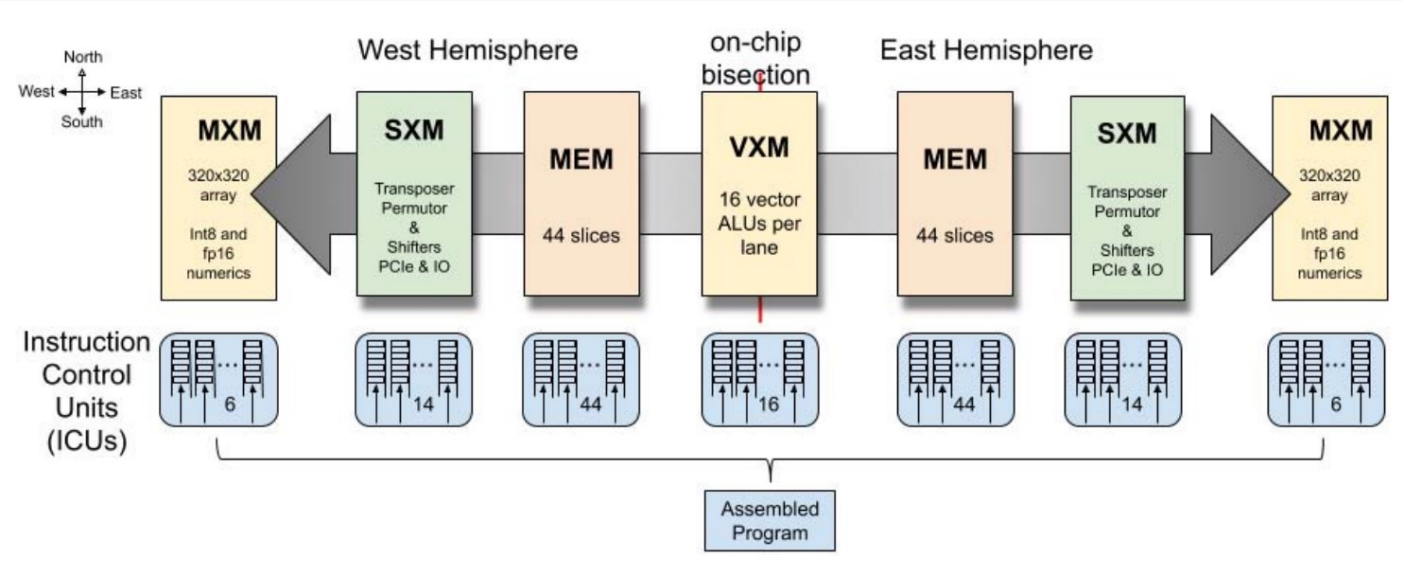
Simplified Reconfigurable Dataflow Unit (RDU) architecture



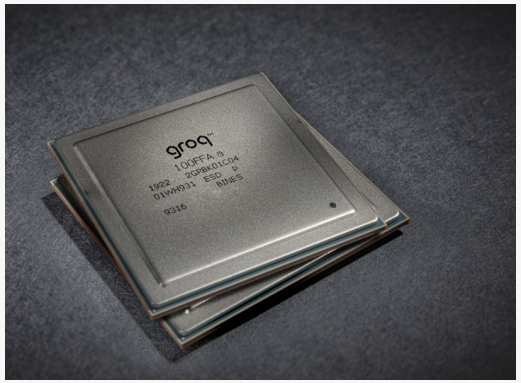
SambaFlow Software Stack

- 300+ MB of on chip memory
- 512GB DDR4 Memory
- 40B transistors, 7nm TSMC
- 300+ TFLOPS of estimated performance
- Support for Sambaflow and PyTorch
- Support for training and inference

Source : Sambanova Whitepaper



The organization and dataflow within a row in the on-chip network.



Tensor Streaming Processor

- 220MB of on-chip memory
- 14nm process, >26.8B transistors
- Estimated performance of
  200+TFlops FP16
  1 PetaOps in int8
- 80TB/s on-die memory bandwidth
- 300W of Power consumption
- Support for GroqAPI and ONNX.
- Support for Inference only
- reduces instruction-decoding overhead, and handles integer and floating-point data

Source : Groq Whitepaper

	Cerebras CS2	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	14nm	7nm
System Size	2 Nodes	2 nodes (8 cards per node)	4 nodes (8 cards per node)	4 nodes (16 cards per node)	2 nodes (8 cards per node)	Several systems
Estimated Performance of a card (TFlops)	>5780 (FP16)	>300 (BF16)	>188 (FP16)	>250 (FP16)	>150 (FP16)	312 (FP16), 156 (FP32)
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

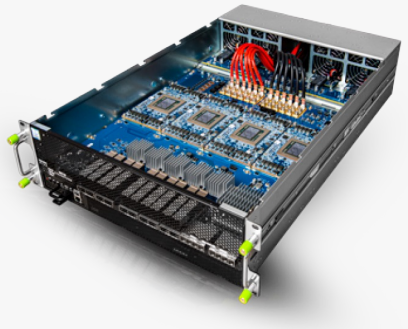
ALCF AI Testbeds

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

Cerebras (CS-2)

SambaNova

Graphcore

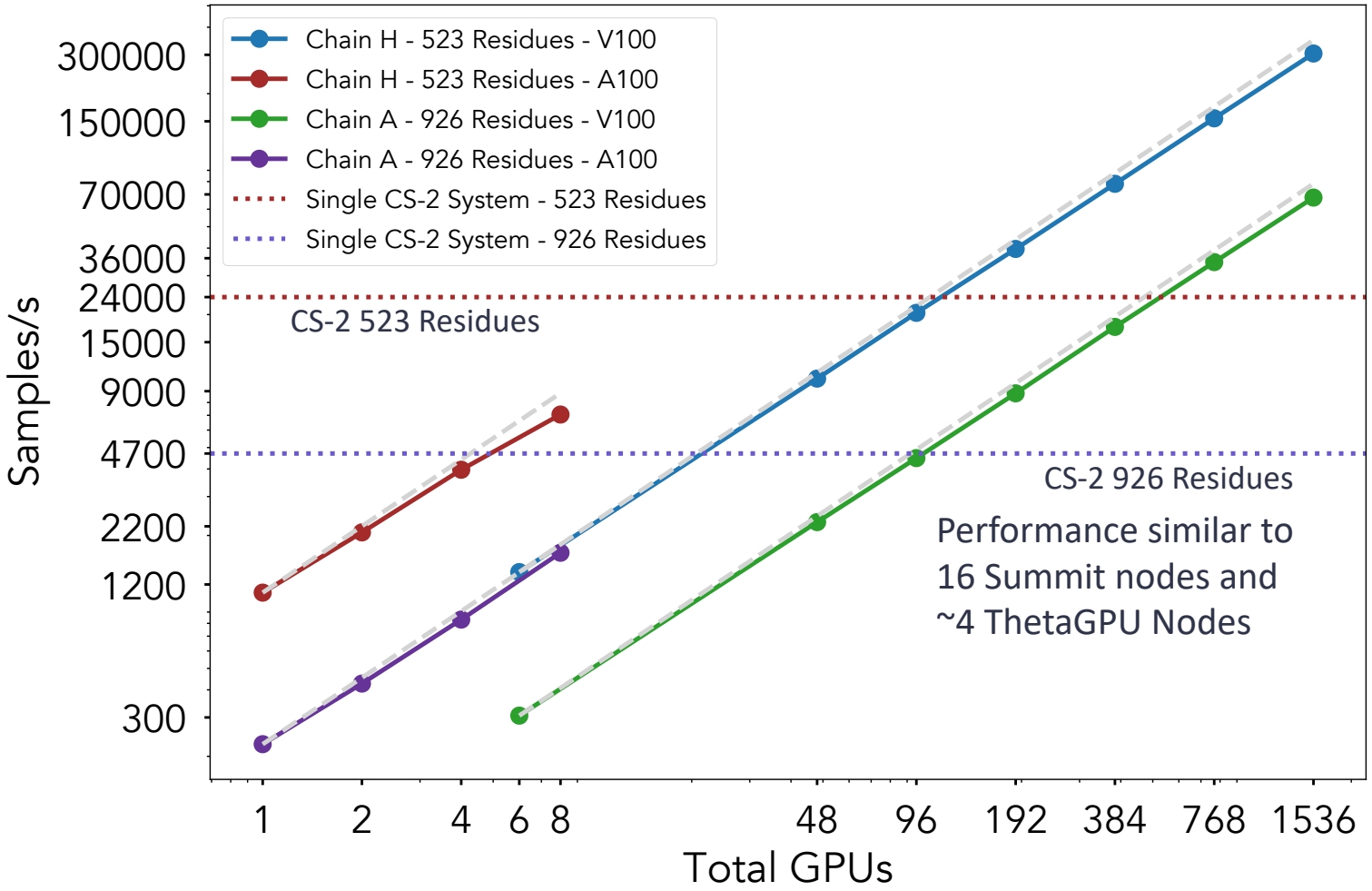


Habana

Groq

- Infrastructure of next-generation machines with hardware accelerators customized for artificial intelligence (AI) applications.
- Provide a platform to evaluate usability and performance of machine learning based HPC applications running on these accelerators.
- The goal is to better understand how to integrate AI accelerators with ALCF's existing and upcoming supercomputers to accelerate science insights

COVID-19 CVAE Training on Summit and Cerebras CS-2

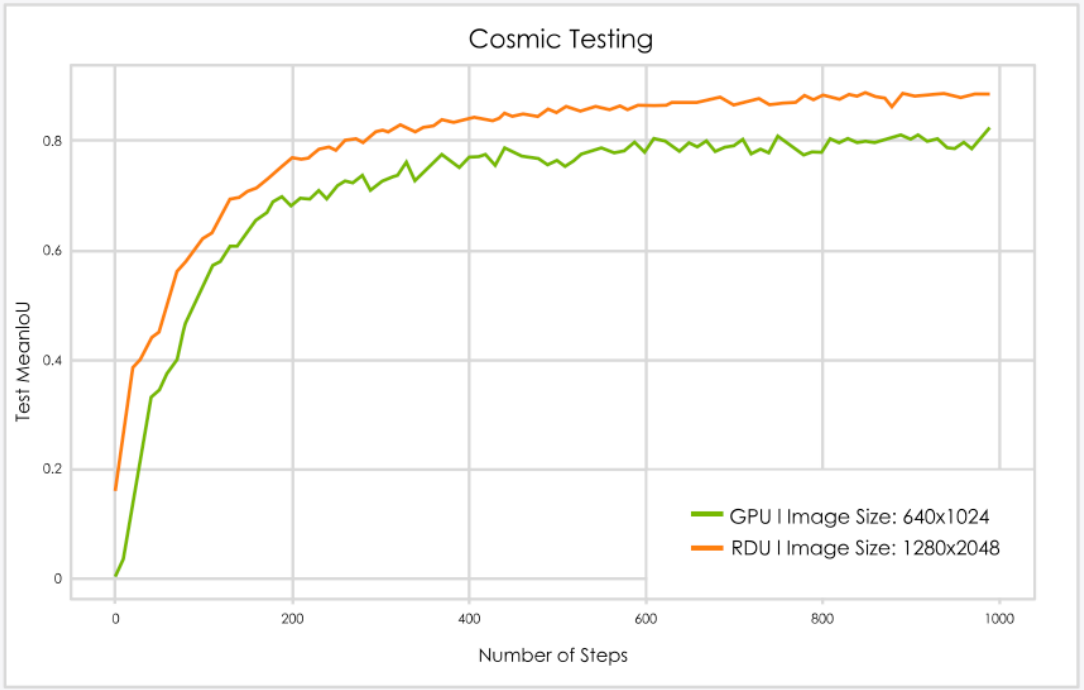


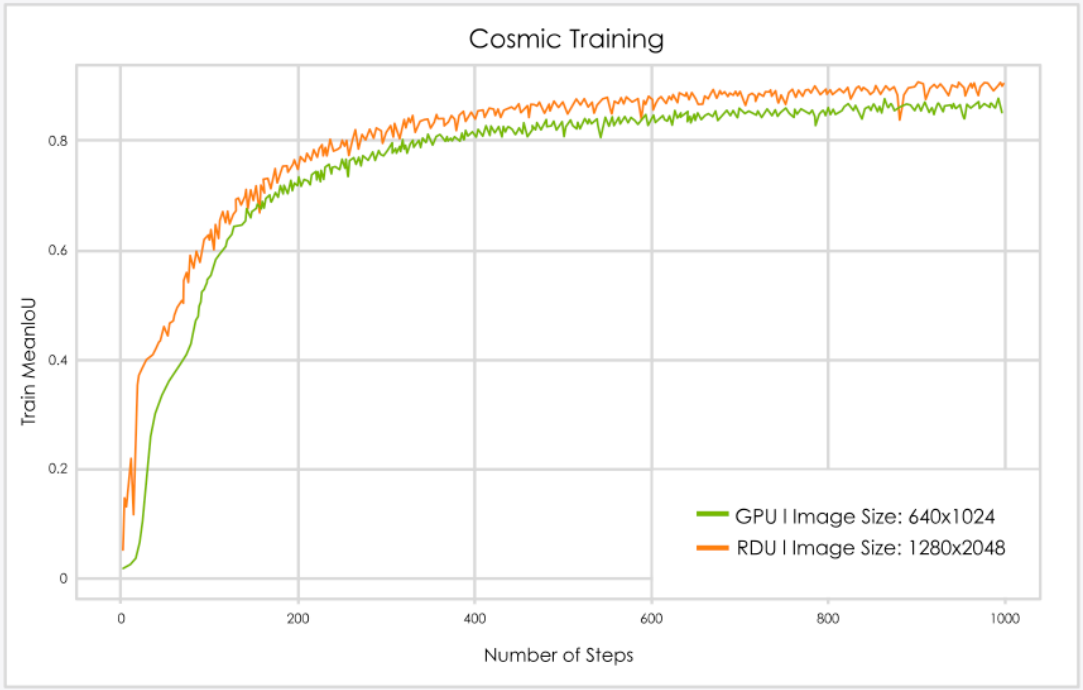
- Single CS-2 delivers performance of over 100 GPUs on CVAE
- Results are for **out-of-the-box performance** based on model config not optimized for CS-2.

Performance	**523 X 523**	**926 X 926**
Throughput (samples/sec)		
1x CS-2 System	24,000	4700
1x V100 GPU	228	51
1x A100 GPU	~1100	~150
Speedup (CS2 vs. GPU)		
1 x V100 GPU	113x	101x
1 x A100 GPU	~22X	~32X

Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action, SC21 COVID19 Gordon Bell Finalist, In IJHPCA 2022
https://www.biorxiv.org/content/10.1101/2021.10.09.463779v1.full.pdf

COSMIC TAGGER ON SAMBANOVA DATASCALE

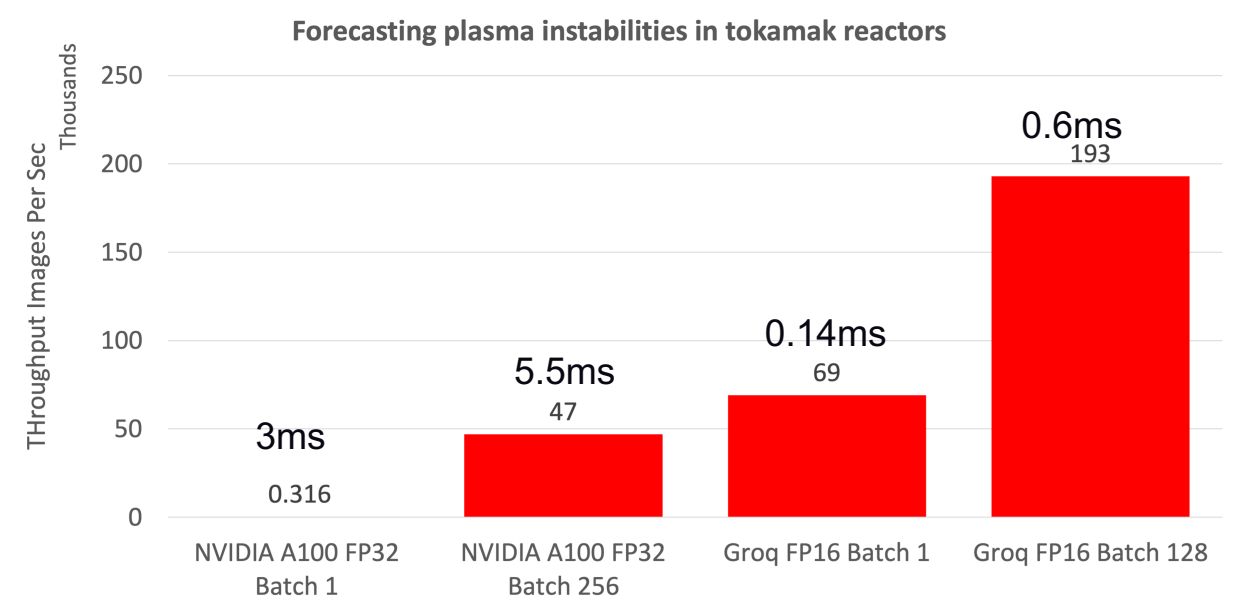




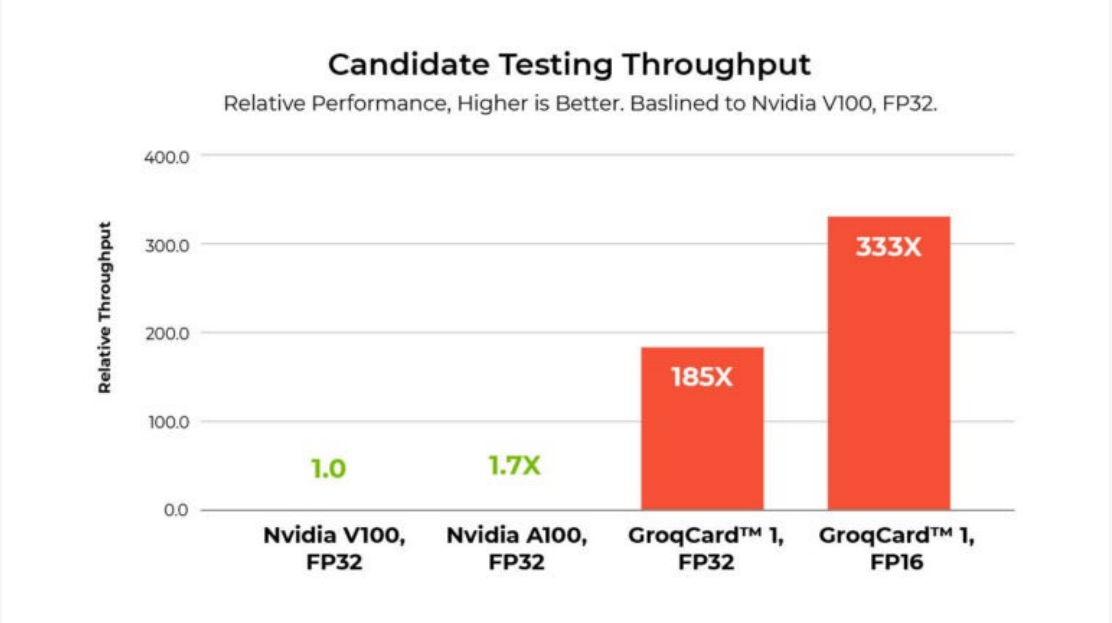
SambaNova RDUs able to accommodate larger image sizes and achieve higher accuracy

M. Emani et al., "Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture," in Computing in Science & Engineering, vol. 23, no. 2, pp. 114-119, 1 March-April 2021, doi: 10.1109/MCSE.2021.3057203.

Early Experience with Inference on Groq



Forecasting Plasma Instability in Tokamak



COVID19 Candidate drug molecule screening

Promising results using GroqChip for science Inference use-cases with respect to latency and throughput in comparison to GPUs

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.

- Murali Emani, 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

venkat@anl.gov