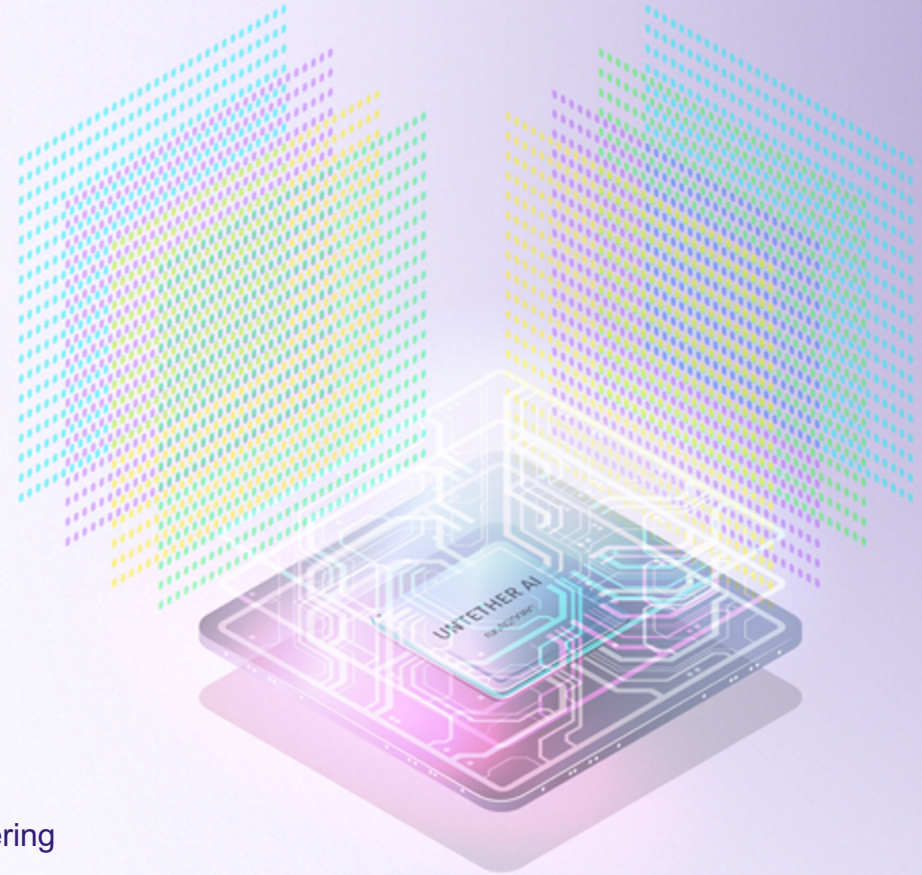
UNTETHER AI



Boqueria

Robert Beachler – VP of Product/Hardware Engineering

Dr. Martin Snelgrove – Co-founder and CTO

A Brief History of the Current AI Summer

2012 2014 2016 2018 2020 2022

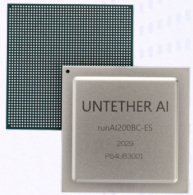
ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky, University of Toronto, kriz@cs.utoronto.ca
Ilya Sutskever, University of Toronto, ilya@cs.utoronto.ca
Geoffrey E. Hinton, University of Toronto, hinton@cs.utoronto.ca

Deepmind acquired by Google

AlphaGo beats Lee Sedol

Untether AI founded in Toronto



runAI200 introduced
First at-memory inference accelerator
500 INT8 TOPs
200MB SRAM
8 TOPs/W
TSMC 16nm

HOT CHIPS

Boqueria introduced
2 PetaFlops FP8
238MB SRAM
30 TFLOPs/W
TSMC 7nm

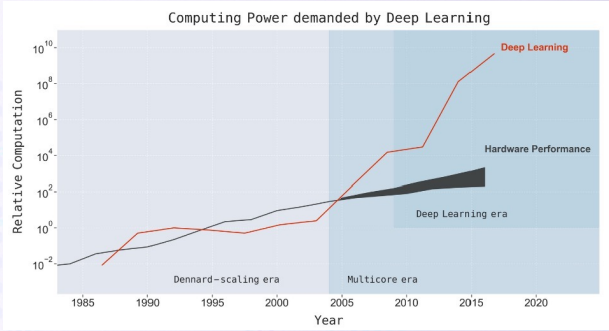
Copyright © 2022 UNTETHER AI Corp.

AI Inference Presents 3 Key Challenges to Chip Makers

Increasing computational and power requirements

Scalability and flexibility for changing NN landscape

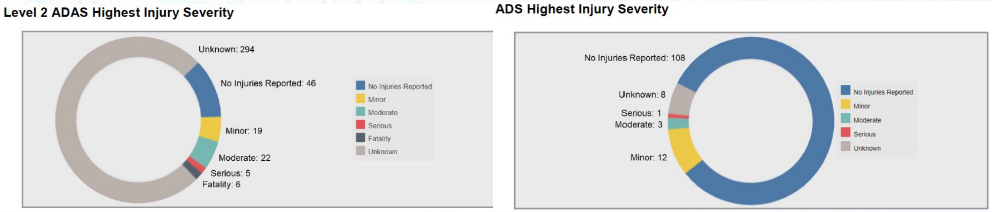
Accuracy loss costs $millions and risks lives



The Computational Limits of Deep Learning
Neil C. Thompson1 , Kristjan Greenewald2, Keeheon Lee3, Gabriel F. Manson

Model Category	Model Name	Model Size (Mparams)
Recommendation	Less complex	70,000
	More complex	>100,000
Computer Vision	ResNeXt101-32x4-48	44
	RegNetY	700
	FBNetV3 based model	28.6
Video Understanding	ResNeXt3D based	58
NLP	XLRM-R	558

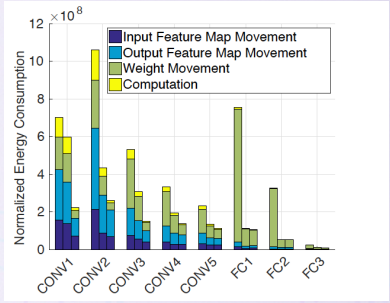
First-Generation Inference Accelerator Deployment at Facebook



NHSTA report, June 2022 for July 2021 to May 2022

UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

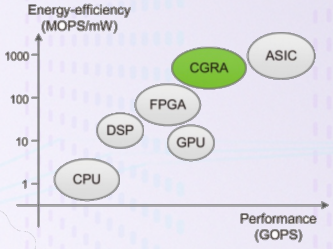
Architecting an AI Inference Accelerator



Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning
Tien-Ju Yang, Yu-Hsin Chen, Vivienne Sze, Massachusetts Institute of Technology

Power-efficient throughput is required to meet NN compute demand

- Data movement is the costliest part of inference – 90% of energy consumption
- Data movement is different between training and inference
- Optimizing compute architecture to minimizing distance travelled results in inference-specific AI accelerators



A Survey of Coarse-Grained Reconfigurable Architecture and Design: Taxonomy, Challenges, and Applications
LEIBO LIU, JIANFENG ZHU, ZHAOSHI LI, et. Al.

Proper level of granularity to create a scalable compute architecture

- Right balance between coarse-grained and fine-grained approach
- Don't over-fit for a particular application/NN

Utilize the most efficient datatype for a given application and accuracy requirements

- A mixture of datatypes provides the best results

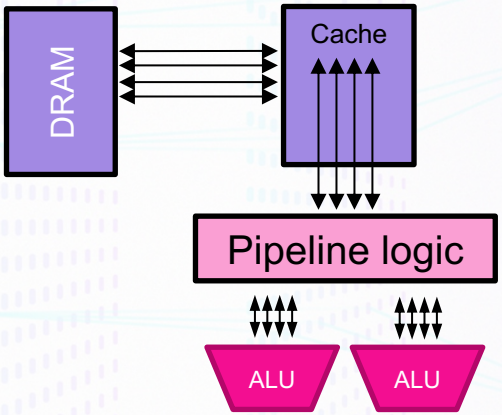
Datatype	F1
FP32	1.000
BF16	1.001
FP8	0.996

BERT-Base SQuAD1.1 accuracy, no retraining, Untether AI

Copyright © 2022 UNTETHER AI Corp.

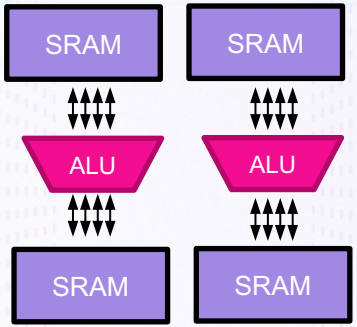
At-Memory Compute Is the Sweet Spot for AI Acceleration

Near Memory/ Von Neumann Architectures



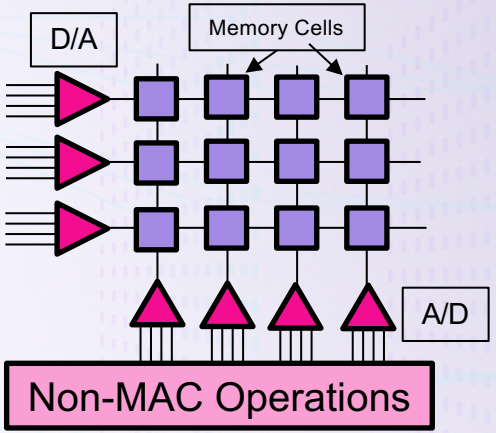
- Long, narrow busses
- Deep/shared cache

At-Memory Computation



- Short, massively parallel direct connections
- Dedicated, optimized memory for efficiency and bandwidth

In-Memory Computation



- Multi-value memory cell
- Analog techniques used for multiply-accumulate
- A/D and D/A support circuitry
- Digital processors for non-MAC operations

UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

Boqueria : A 2 PFLOPs, 30 TFLOPs/W At-Memory Inference Accelerator with 1,458 RISC-V Cores

729 Dual RISC-V memory banks provide unmatched performance

- 2,015 FP8 TFLOPs, 1,008 BF16 TFLOPs*
- 1.35 GHz, TSMC 7nm

At-memory compute provides energy efficiency and massive bandwidth

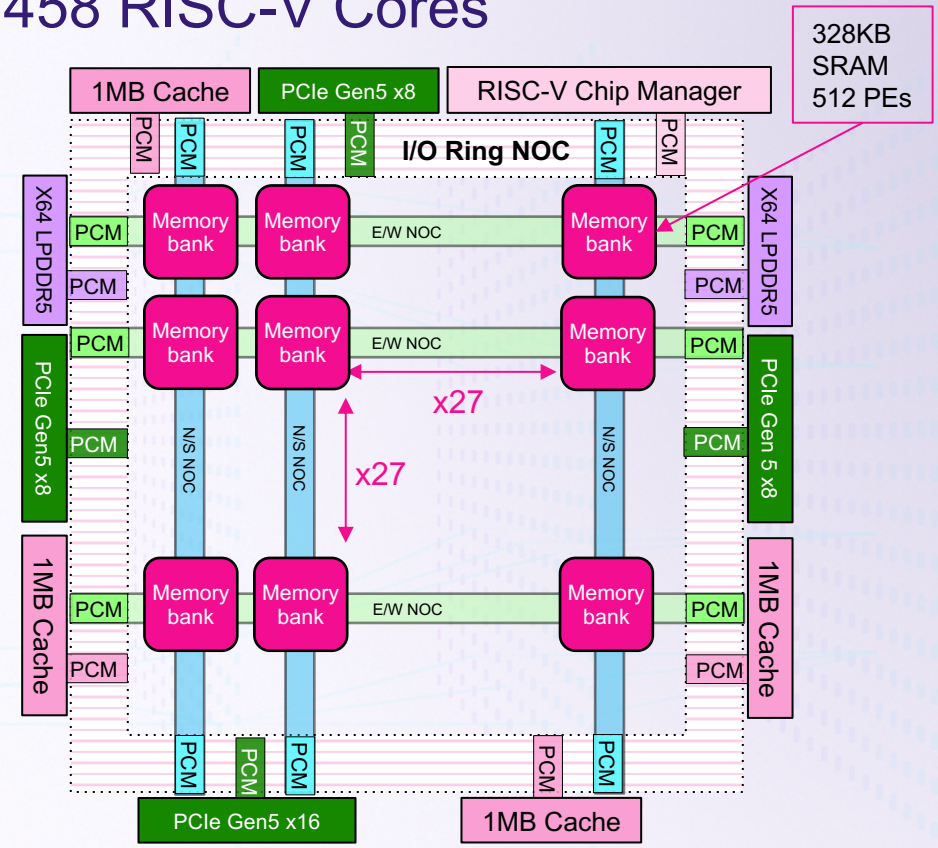
- 30 TFLOPs/W
- 238MB on-chip SRAM
- ~1 PB/s SRAM bandwidth

Scalability

- External memory support
  - 32GB LPDDR5 across two x64 ports
  - 819 Gb/s DRAM bandwidth
- PCI-Express Host and chip-to-chip connectivity

Accuracy

- Multiplicity of datatypes - INT4 to BF16



UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

All data based on full datapath parasitic extraction and simulation including leakage and dynamic power
*BF16 coefficients, FP8 activations

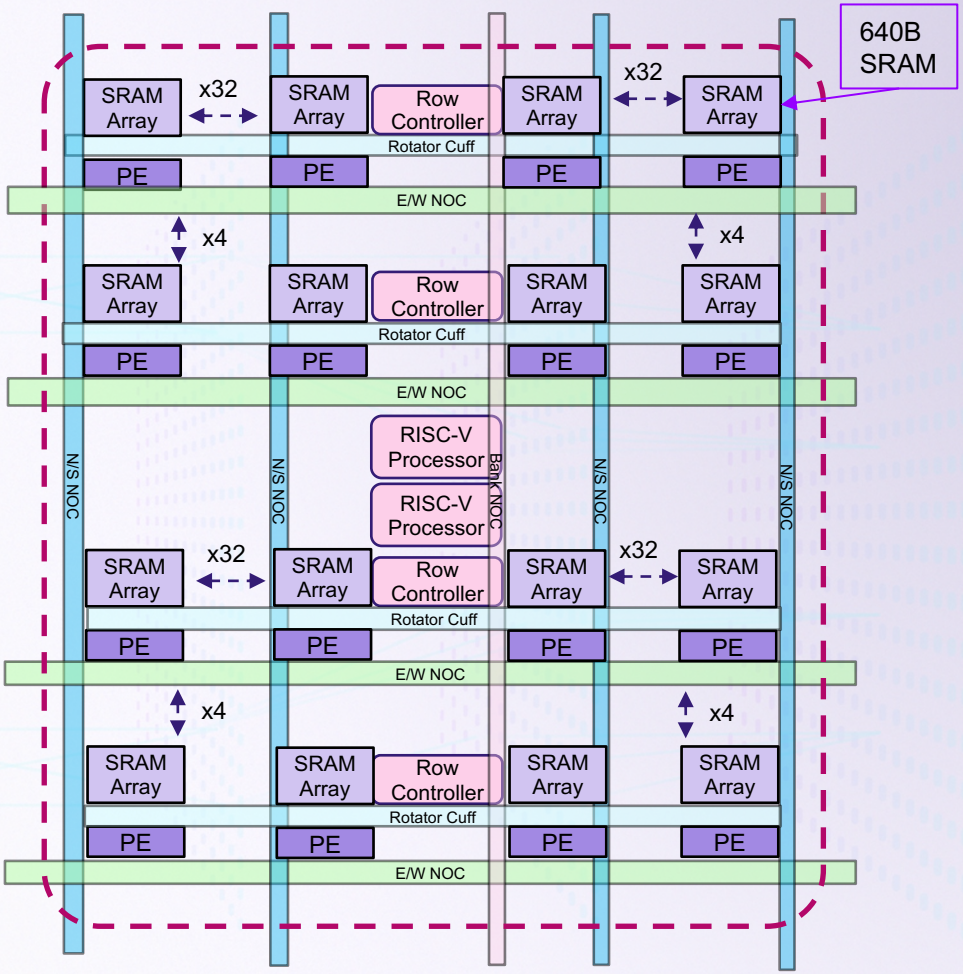
The Memory Bank

Dual multi-threaded RISC-V for programming flexibility

- Each RISC-V manages 4 row controllers
- Row controllers operate independently
  - Command/control 64 SIMD PEs for GEMV calculations, scalable to GEMM

Extreme connectivity – custom pipelined communication

- Rotator cuff moves activations between nearest neighbor PEs, conserving energy
- 8 E/W NOCs, each with 7GB/s bandwidth in each direction (56GB/s total in each direction)
- 1 N/S NOCs, with 70GB/s bandwidth in each direction
- Bank NOC allows communication between RISC-V Processors



UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

At-Memory Compute – Putting Compute Where the Data resides

Low-Power SRAM Array

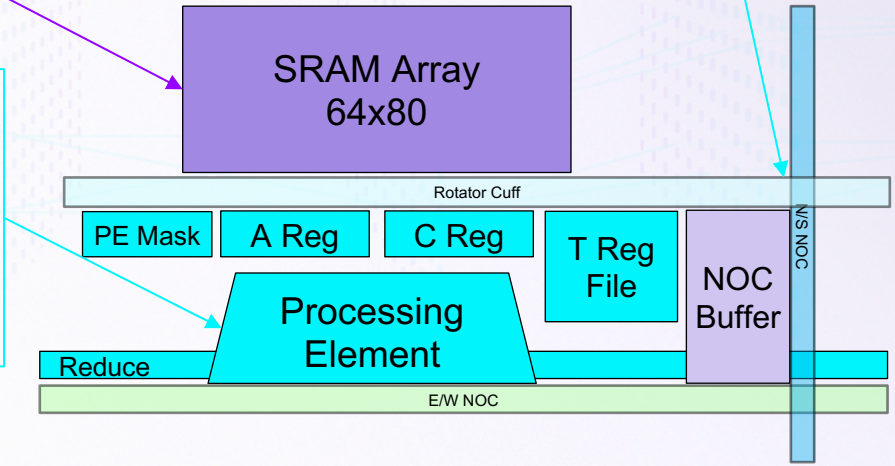
- Coefficient and data storage
- 0.4V datapath operation
- Custom drivers to minimize power when reading/writing to memory

Energy efficient data transfer

- Rotator cuff accesses activations from nearest neighbors
- E/W and N/S pipelined NOCs for spanning rows/columns of memory banks

Processing Element

- INT4, INT8, FP8p/r, BF16 Support
- Zero detect for power savings
- Structured sparsity support
- Dedicated reduce circuitry for functions such as SoftMax and LayerNorm



UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

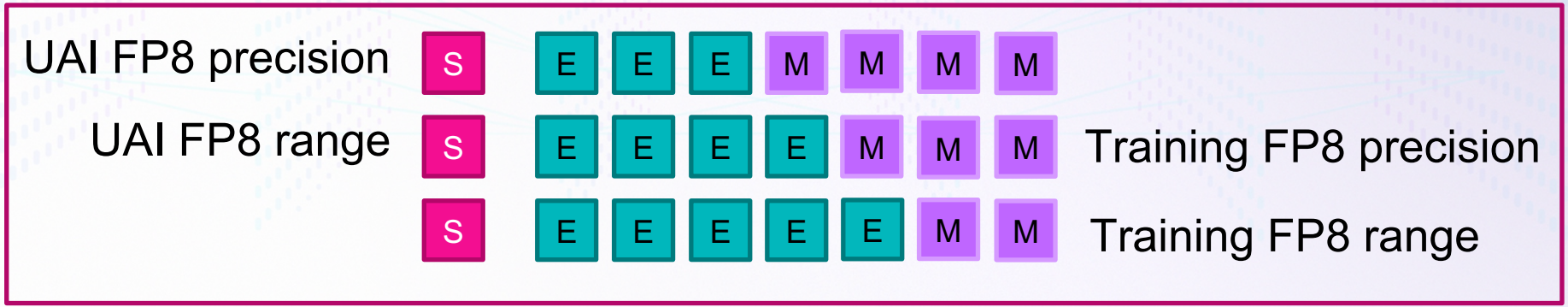
Untether AI FP8 – Designed for Inference Acceleration

AI Inference requires precision – but full floating point is expensive to achieve needed accuracy targets

Untether AI researched various datatype and found FP8 provided the best balance of precision, throughput, and energy efficiency

- But requires both range (FP8r) and precision (FP8p)
- Each additional mantissa doubles the precision, and reduces mean squared error by (1/2)^2

FP8 is 2x times more energy/die size efficient than if designed for native INT8



UNTETHER AI Copyright © 2022 UNTETHER AI Corp

Accuracy Results – Relative Degradation

ResNet-50 ImageNet accuracy

Datatype	Top-1 Accuracy
FP32	1.000
BF16	1.000
INT8	0.992
FP8	0.991

BF16 for ultimate accuracy

FP8 to quadruple throughput

Accuracy degradation between FP8 and INT8 is negligible

FP8 quadruples throughput

BERT-Base SQuAD1.1 accuracy

Datatype	Exact Match	F1
FP32	1.000	1.000
BF16	1.000	1.001
FP8	0.988	.996

Copyright © 2022 UNTETHER AI Corp

All data based on Untether AI quantization techniques and accuracy testing

Multi-threaded Custom RISC-V Processor – Adapted to AI Inference

Standard RV32EMC Instruction Set

- Embedded
- Multiplication/division
- Compressed instructions
- ***UAI added 20+ custom instructions specific to at-memory compute and inference acceleration***

Each Processor

- 6KB Memory
- 32-bit ALU
- 32-bit multiplier
- x16 register file with 4-way context switching to enable 4 threads

Example Custom Instructions

Instruction	Description
pe_move	Copy PE register to PE register
pe_recv	Receives packets from socket
pe_send	Send packets to socket
pe_load	Load PE registers from CRAM
pe_store	Store PE registers to CRAM
pe_macc	Perform MACC
pe_gemv	Perform GEMV
pe_multimove	Copy into multiple PE registers
pe_broadcast	Broadcast and copy to PE register
pe_rotate	Rotate for count cycles
pe_convert	Convert d and output to PE register
pe_set_fmt	Set number format
pe_row_reduce	Run row reduce function
pe_reduce	Run PE reduce function
pe_k_stack	Push and pop k stack
rwc_nop	nop for count cycles
rwc_sleep	Sleep with wakeup mask
eq_set_cfg	Set row extension configuration
eq_save	Save cuff registers to EQ
rwc_magic	Magic function to output RWC commands

Copyright © 2022 UNTETHER AI Corp.

Copyright © 2022 UNTETHER AI Corp.

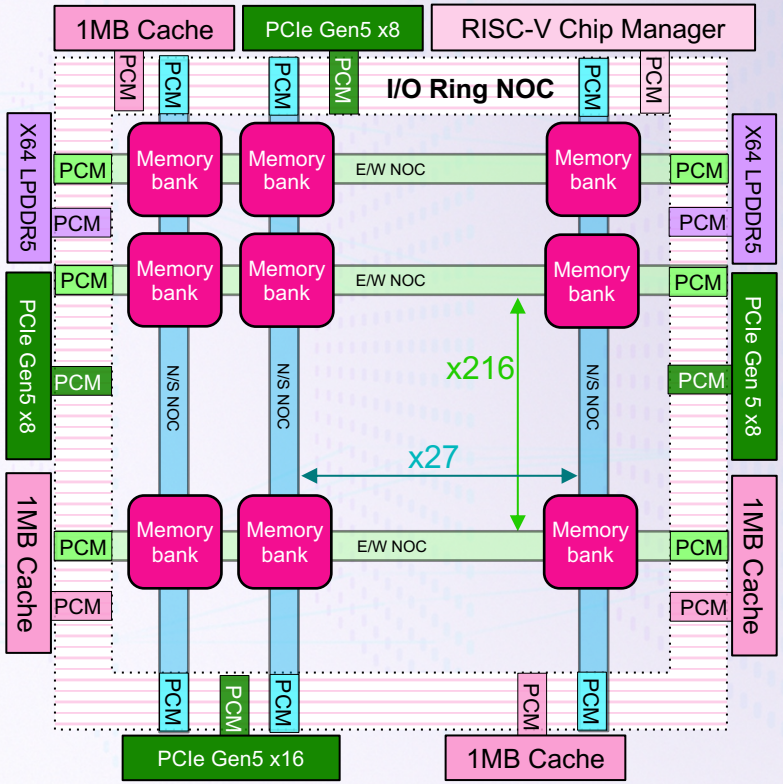
High-bandwidth I/O and Connectivity

High-Bandwidth I/O NOC
- 141GB/s in clockwise direction
- 141GB/s in counter-clockwise direction

Efficient, high throughput data NOCs
- 1.5 TB/s East and West throughput
- 1.9 TB/s North and South throughput

Extremely scalable
- X16 PCIe Gen5 for host connectivity – 63GB/s
- 3 ports of PCIe Gen5 x8 for chip-to-chip and card-to-card connectivity – each 31.5GB/s

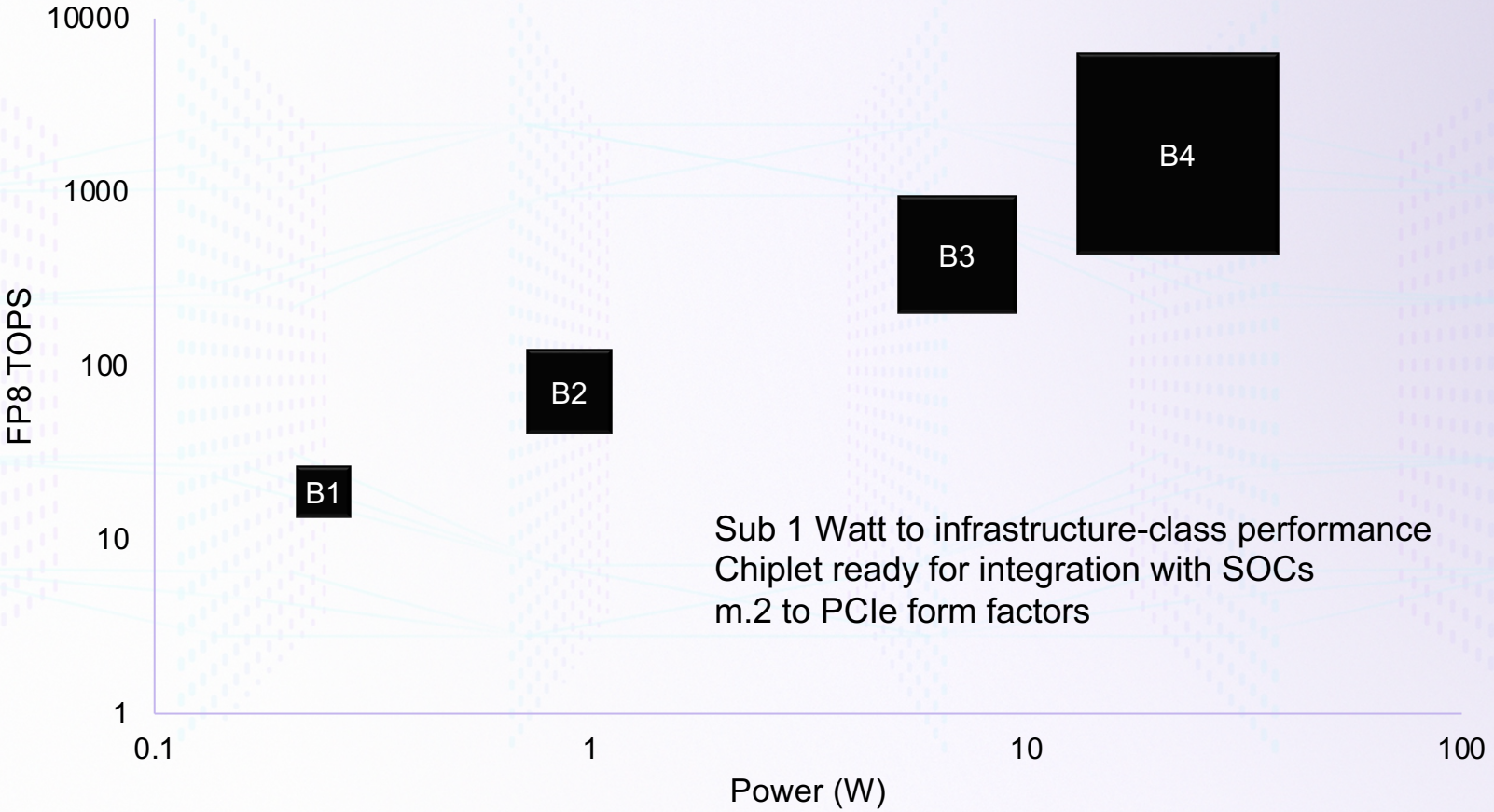
Multi-level Memory Architecture
- 238MB of At-memory SRAM for efficient compute
  - ~1 PB/s bandwidth
- 4MB of scratchpad for data manipulation
- 32 GB of external LPDDR5 with >100GB/s bandwidth



 Copyright © 2022 UNTETHER AI Corp.

Copyright © 2022 Untether AI Corp.

The Benefits of a Scalable Architecture



Sub 1 Watt to infrastructure-class performance
Chiplet ready for integration with SOCs
m.2 to PCIe form factors

 Copyright © 2022 UNTETHER AI Corp.

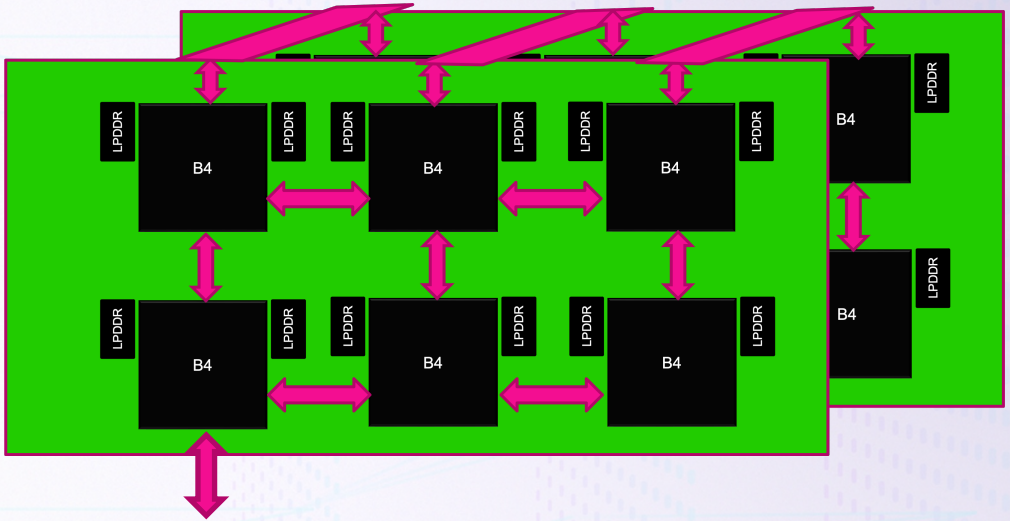
Scalability For Large Language Models

Maximum performance: 6-Chip PCIe Gen5 Card

- 1.4GB SRAM
- 12 PetaFLOPs of FP8 compute
- 192GB LPDDR5 DRAM

Scalability

- PCIe GEN5 x8 chip-to-chip and card-to-card communication



Copyright © 2022 UNTETHER AI Corp

imAIgine SDK Tool Flow

User Interface

Automated Optimization

constraints

Customer Algorithm

Standard Frameworks

Pre-Trained model

Quantization

Automated Quantization within Framework

Compiler

IR Creation
Layer Fusion
Kernel mapping
LLVM Compile

Physical Allocation

Kernel Merge
Multi-chip partitioning
Placement optimization
Weight/activation transforms
Bank communication merge

Analysis

Allocation Visualization
Utilization
Congestion
Cycle-accurate Simulation

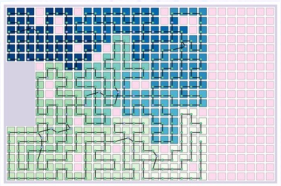
Runtime

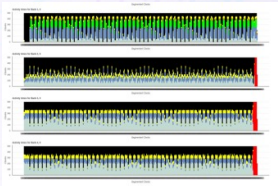
Easily Integrated communication

Health Monitoring software

Model Garden

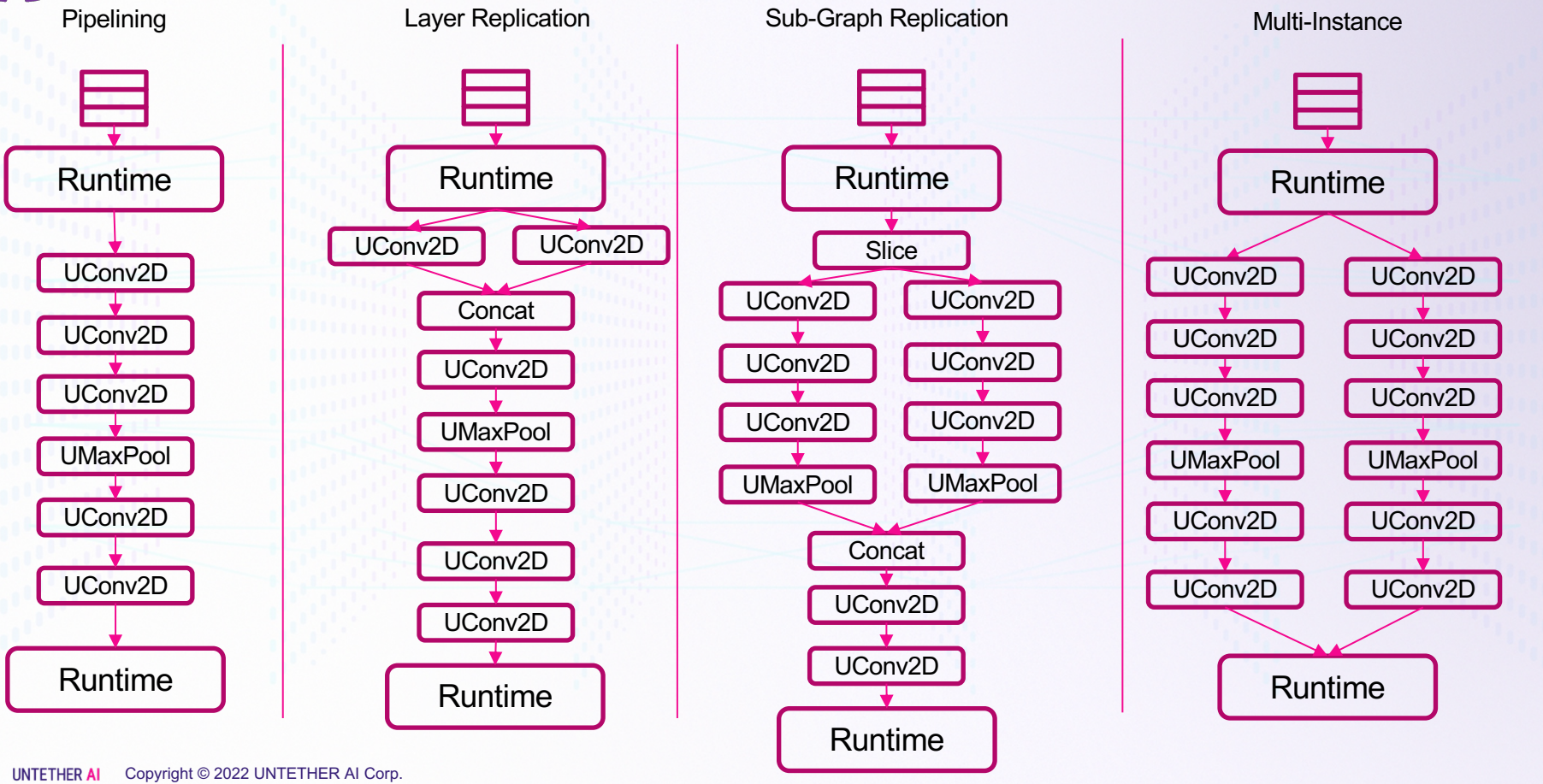
UAI Kernel Library
Custom Kernels
(C / C++ / ASM)





UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

imAIgine SDK: Spatial Compilation Optimizations

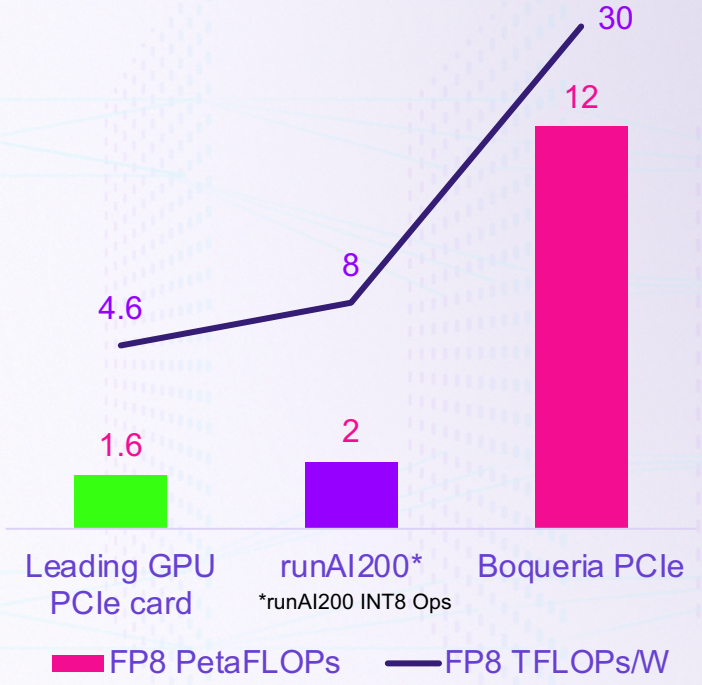


UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

The Efficiency of At-memory Compute

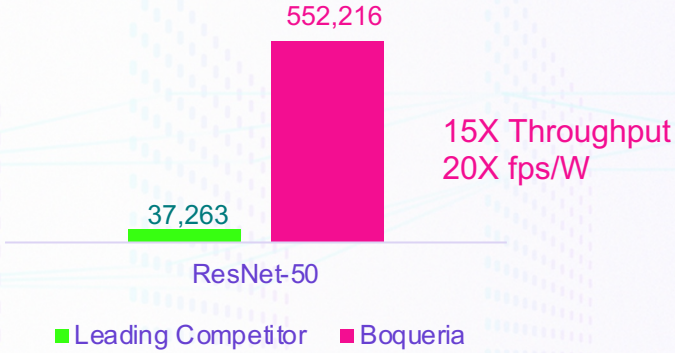
Boqueria compared to leading GPU

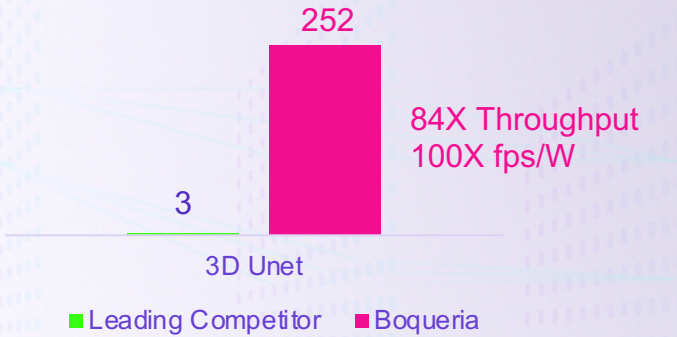
- 5x greater performance (PFLOPs/card)
- 7X greater efficiency (TFLOPs/W)

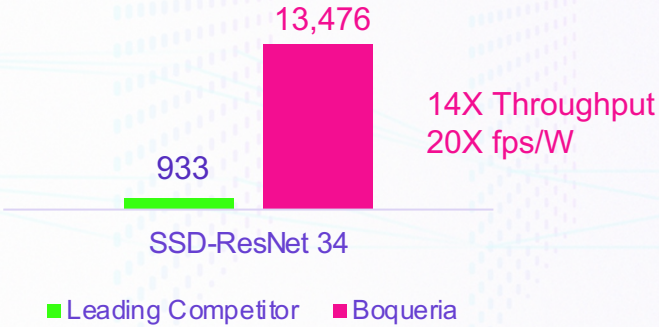


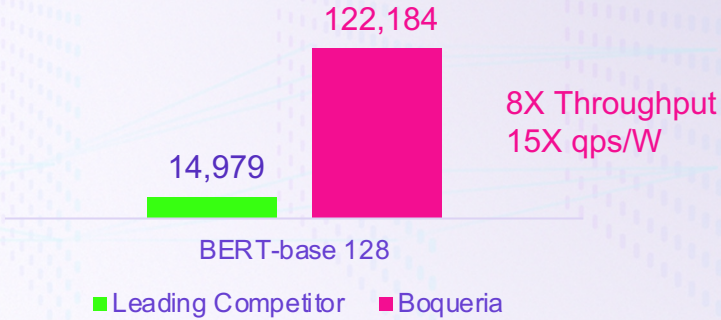
UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

Energy Efficiency Translates to Throughput









Competitor information based on MLPerf 2.0 best performance normalized to single PCIe-card and TDP except BERT-base 128 based on published benchmark by vendor and TDP

Boqueria information based on kernel code cycle counts and simulations

UNTETHER AI Copyright © 2022 UNTETHER AI Corp.

Boqueria – A 2 PFLOPs, 30 TFLOPs/W At-Memory Inference Accelerator

Unrivaled throughput and efficiency

- Up to 100X efficiency, 80X throughput for a variety of neural network models

Scalable from small to large models

- Spatial architecture and chip/card interconnects enables optimal performance/power

Flexible to adapt to latest NN architectures

- Equally adept at vision and NLP networks

Efficient and accurate datatypes

- FP8p/r for accuracy and throughput, BF16 for utmost accuracy

Copyright © 2022 UNTETHER AI Corp.