rebellions_

Start Lean, Scale Green

Flexible Gen AI Inference via Chiplet to Rack Solutions

AI Hardware & Edge AI Summit 2024

Jinwook Oh, Ph.D. | Co-Founder & CTO

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

About Rebellions

The world's finest talents

MIT, Harvard University, Stanford, NVIDIA, IBM Watson, ARM, Google, Meta, intel, Samsung Electronics, Apple, SK hynix

110+ tech elites with 35+ PhDs from local powerhouses to global big tech

The advanced technology and leadership

100+ KR IPs
80+ US IPs
X
ION™
ATOM™
REBEL™

MLPerf

STAC BENCHMARK COUNCIL

Verified core intellectual properties with 3 TOs and 2 international benchmarks

KR's most-funded startup

kt cloud
kakao MAU: 40M+
TEMASEK
korelya capital
DG Daiwa Ventures
WAED by aramco

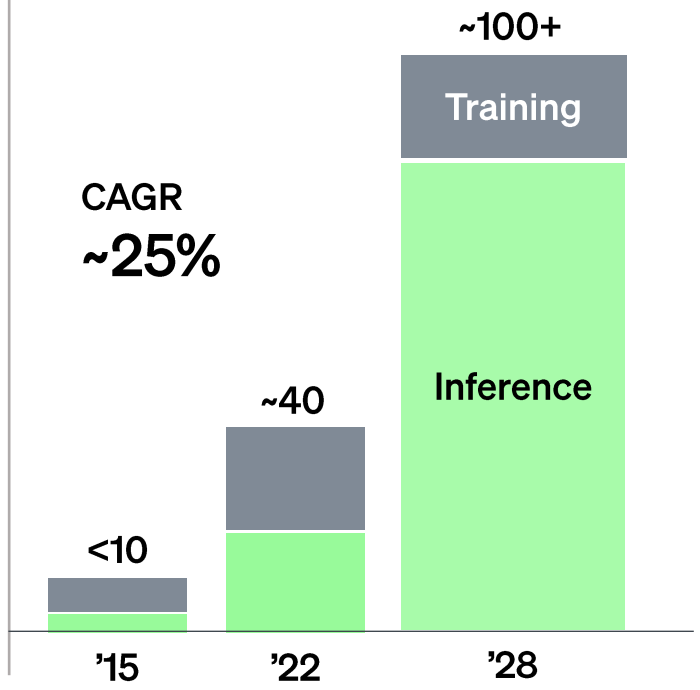
Top 10 AI chipmakers pulling the most VC cash ($225M USD investment in total)

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Target Market & Target Application

AI chip market size ($B)



Model size →

Gen AI model seg	<10B	<30B	<100B	100+B
	On-device	Task specific	Vertical killer	Ubiquitous
		Image, video, music, voice with small language model	Mid language model fine tuned with proprietary data	Large language model for general purpose (multi-modal)
Foundation model[1)]	Compressed task specific model Quantization, pruning/sparsity	OpenAI: DALL-E, CODEX, WHISPER Meta: Llama-7B, Llama-13B stability.ai: Stable Diffusion Microsoft: PHI-3 Google: T5, BERT	Meta: Llama-70B ANTHROPIC: Claude Google: Chinchilla mosaicML: MPT	OpenAI: GPT-3, GPT-4 Google: Bard, PaLM amazon: Titan Microsoft: Kosmos Baidu: ERNIE HUAWEI: Pangu
Target customer		Canva, Poe, Daz3D, GitHub Copilot	perplexity, NOMURA, LexisNexis, Jasper	Copilot, Morgan Stanley, Quora, Dropbox, duolingo

"Two-thirds of human occupational workloads are exposed to automation by Gen AI"
(Goldman Sachs Research)

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Scalable Infrastructure for Efficient Inference Solution

The amount of energy saved exponentially increases at the rack level, leading to significant cost savings

AI Accelerator
Energy-efficient & performant AI chip

Card-level
Number of chips per card

Server-level
Number of cards per server

Rack-level
Number of servers per floor area

Unlocking Exponential Energy Savings for A More Sustainable Future

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

AI Accelerator Roadmap

2022	2023	2024	2025	2026
ION™ (TSMC 7nm)	ATOM™ (Samsung 5nm)	REBEL™ (Samsung 4nm)	REBEL™ –IO (TBA)	REBEL™ –CPU (TBA)
Single Neural Engine	**Inference SoC for AI-as-a-Service (AIaaS)**	**Multi-chiplet LLM Accelerator**	**Scalable chiplet solutions up to datacenter scale**	**ML and CPU integrated AI accelerator**
CGRA-like ML Compute 4 TFLOPS (FP16) 16/32 TOPS (INT8/4) 4MB On-chip SRAM	ML SoC w/ Flexible Cores 32 TFLOPS (FP16) 128/256 TOPS (INT8/4) PCIe Gen5, GDDR6 64MB On-chip SRAM	ML Chiplet (1 to N) Support large on-chip/off-chip mem. BW High ML compute density	Heterogenous integration Scalable platform core to rack Low-latency all-to-all connection	Energy-efficient server design 10x efficient server and rack solutions Reliable and scalable compute

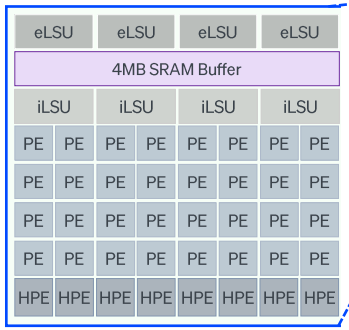
This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

AI Accelerator Roadmap

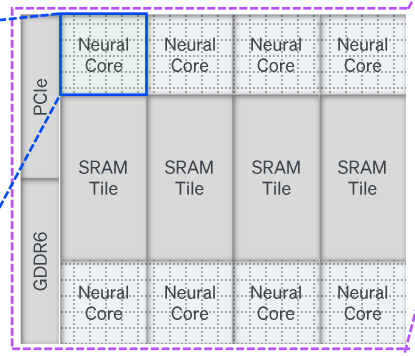
2022

- ION™ (TSMC 7nm)



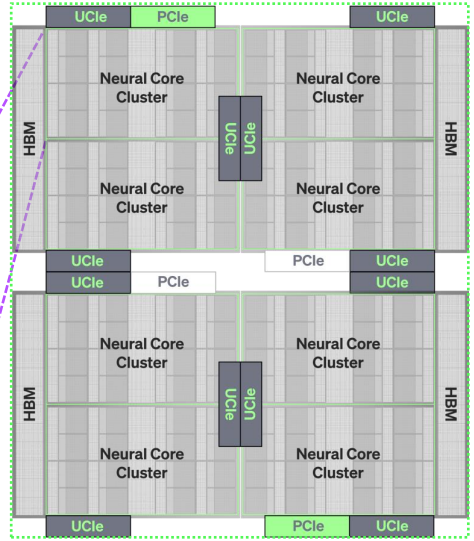
2023

- ATOM™ (Samsung 5nm)



2024

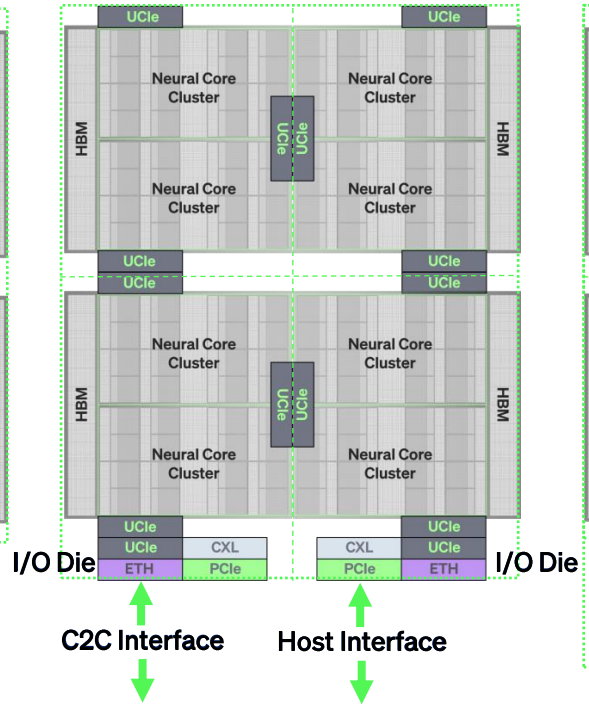
- REBEL™ (Samsung 4nm)



2025

- REBEL™ –IO (TBA)

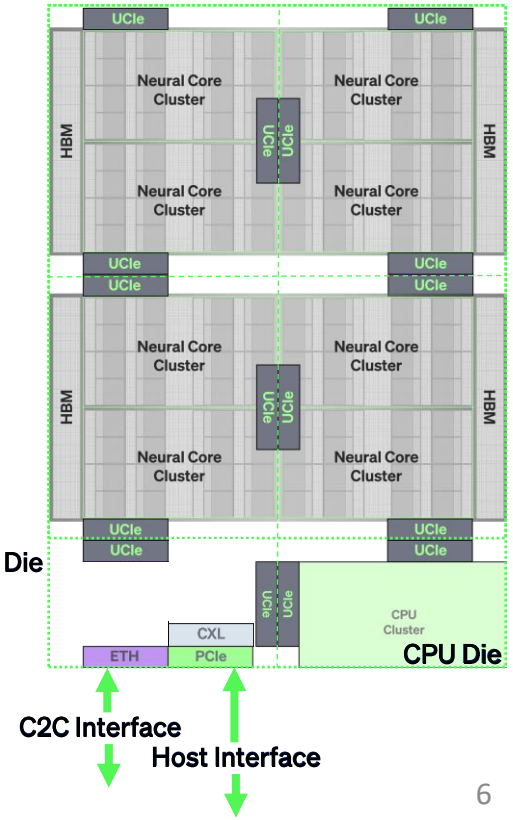
*Open to change



2026

- REBEL™ –CPU (TBA)

*Open to change



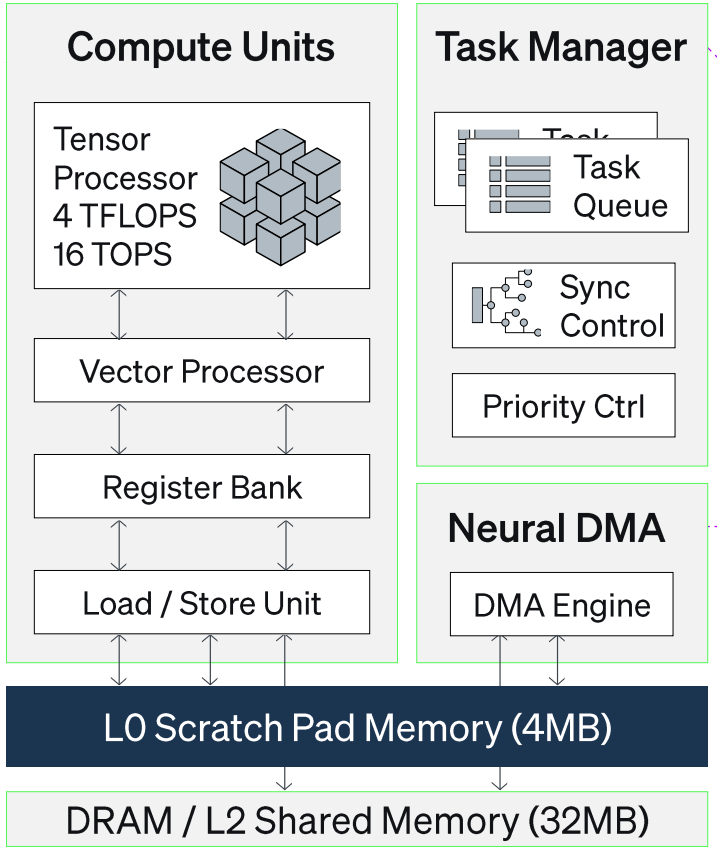
This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

Single Neural Engine ION™

Compute Granule of Rebellions Hardware

Compute Units

- Heterogenous SIMD/MIMD elements
  - 4 TFLOPS (FP16), 16 TOPS (INT8)
  - 4 MB Scratch-Pad for Working Memory
- Programmable dataflow architecture
  - Instruction Level Parallelism/Sync.



Task Manager

- Multi-task Queues
- Task Level Parallelism/Sync.

Neural DMA

- Prioritized Multiple DMA Operation

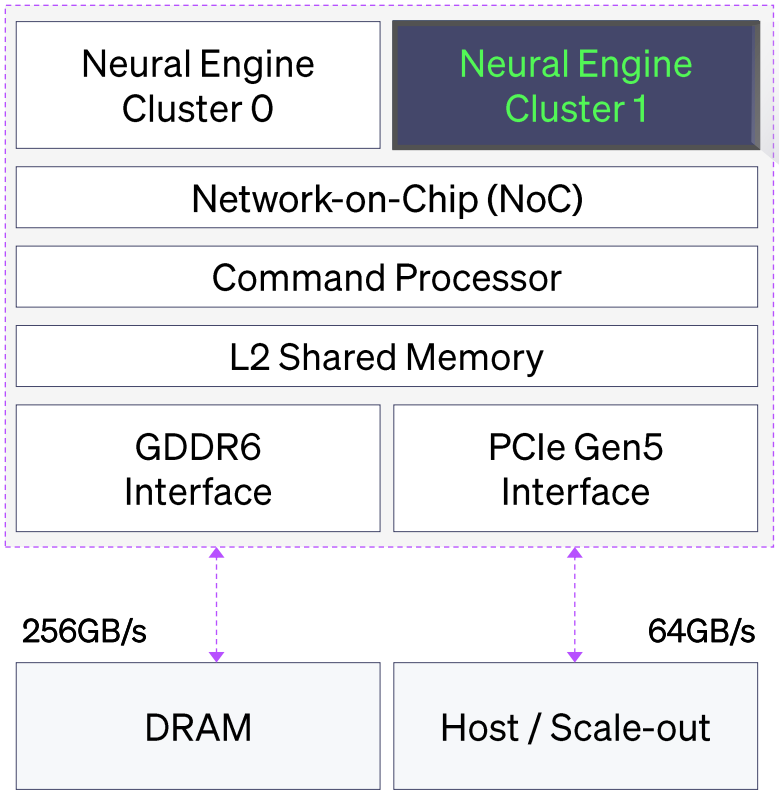
Flexible and Energy-efficient AI Core with High Compute Utilization

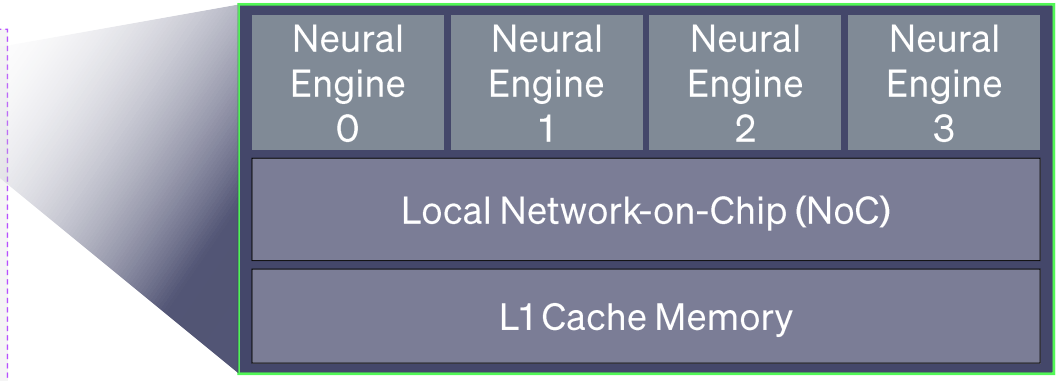
This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Neural Engine Cluster ATOM™

ML SoC for AI-as-a-Service Applications



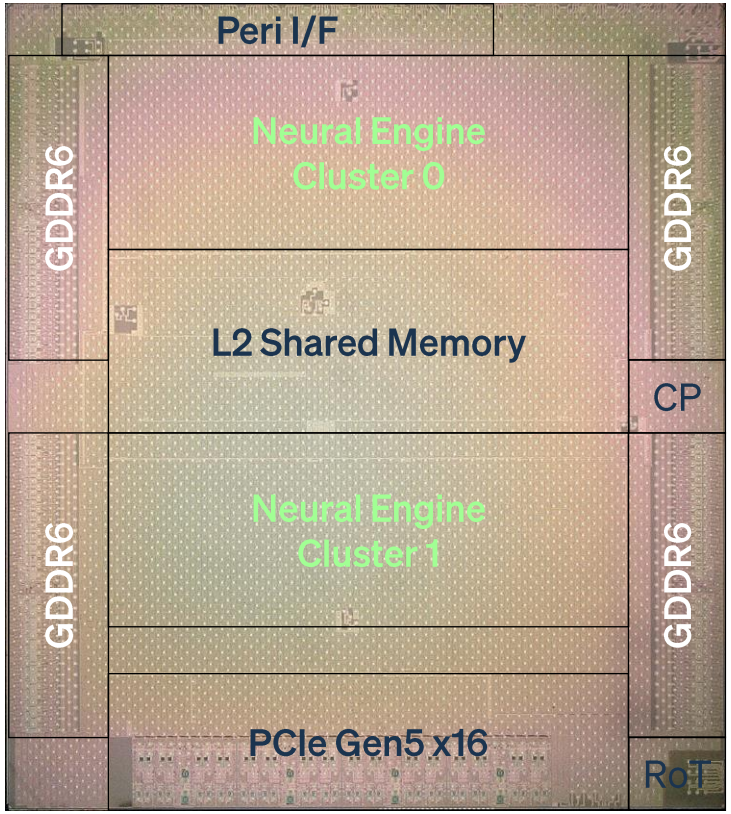


- **Multi-core SoC with 8 Neural Engine Cores ganged into 2 Clusters**
  - 32 TFLOPS, 128 TOPS compute
  - Fine control by dedicated sync bus
- **Multi-card LLM acceleration**
  - Geared with PCIe Gen5 and GDDR6

Energy-efficient AI Inference for GenAI Applications

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

Neural Engine Cluster ATOM™ Chip Implementation Results



ATOM™ Specifications

Single Chip	
FP16	32 TFLOPS
INT8 / INT4	128 TOPS / 256 TOP
On-chip SRAM	64MB
Max Thermal Design Power (TDP)	60W – 130W (Configurable)
External Memory	GDDR6 16GB, 256GB/s
Host / Chip-to-Chip Interface	PCIe Gen5, 64GB/s
Multi-Instance NPU	HW isolation up to 16 Independent tasks

Energy-efficient AI Inference for GenAI Applications

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

Rebellions Scalable Design

Card to Server and Rack Solutions

rebellions_

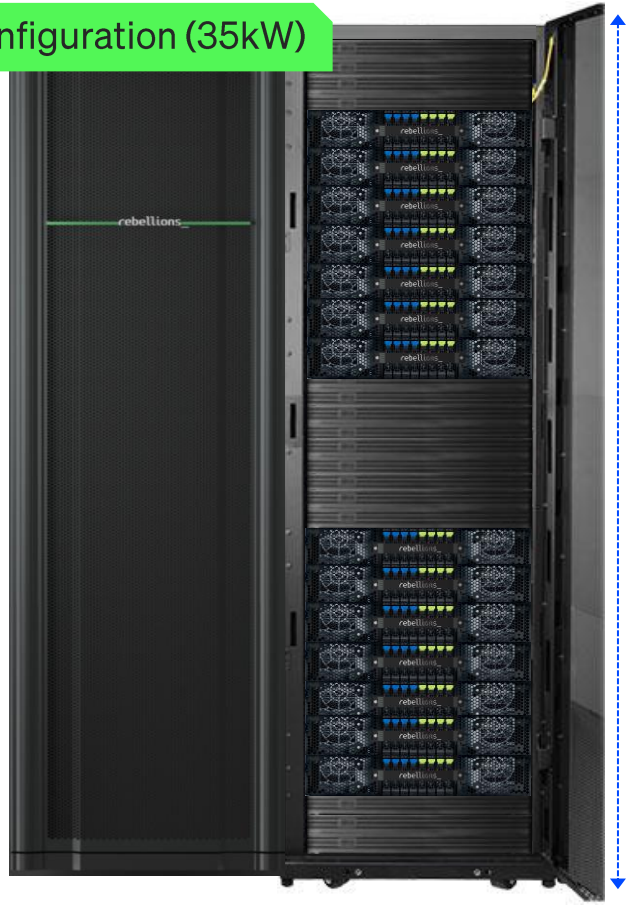
This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Rebellions Scalable Design

Card to Server and Rack Solutions

Configuration (35kW)



High-density servers (2U)

Specification

Feature	Description
Chassis	42U (Height:2100 mm, Width:600 mm, Depth:1200 mm^3)
Server	Up to 14 servers, each with 16 ATOM cards (Optional) 1 additional redundant server
Network	2x High performance 10 GbE data switch 1x1 GbE management switch
Storage	3x nodes storage cluster (100TB)
Power	4x redundant PDUs (2N redundancy)
OS	Ubuntu Linux RHEL & Rocky Linux
Cloud Platform	Kubernetes, OpenStack

Service

- Custom RAG
- Agentic Task
- Normal Chat

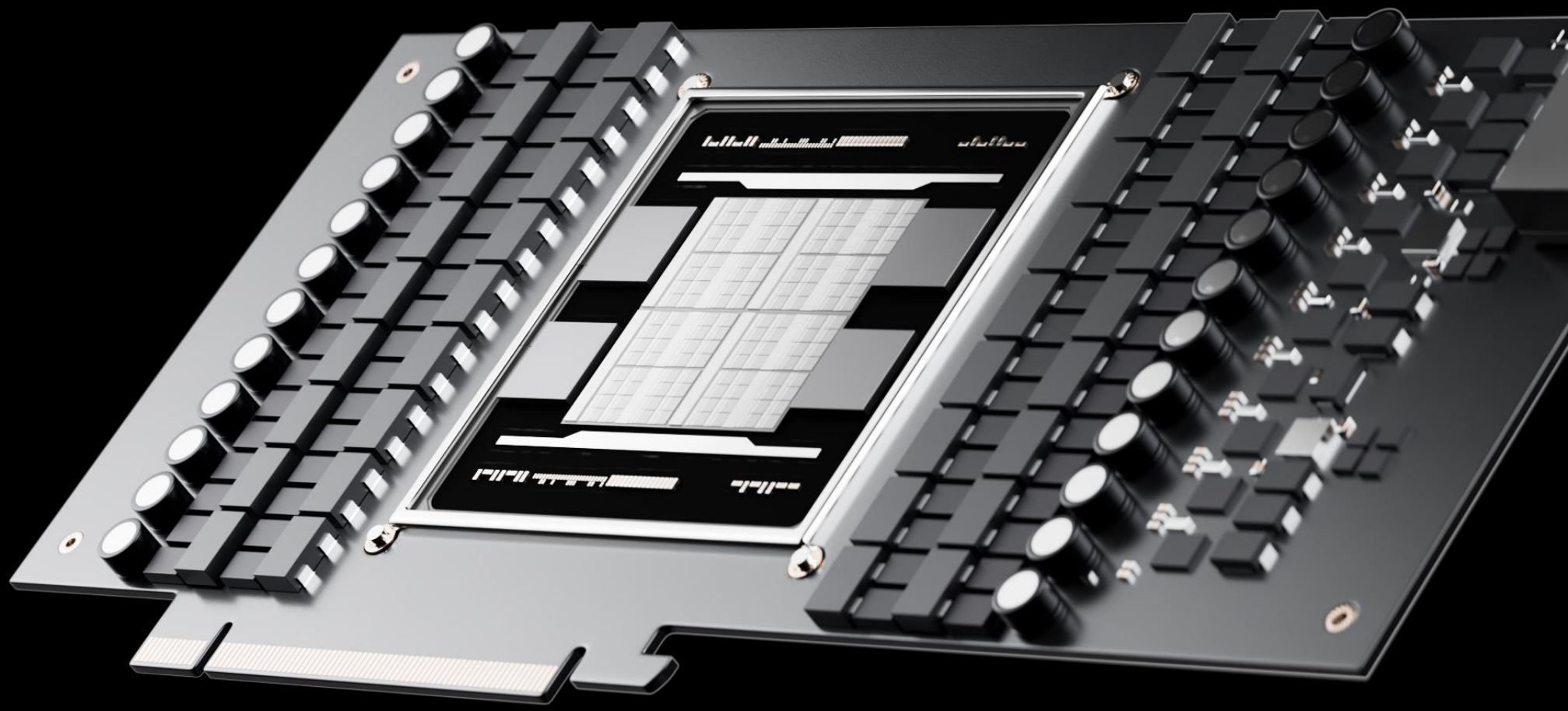
PyTorch 2.0
Hugging Face
vLLM
NVIDIA TRITON INFERENCE SERVER
LlamaIndex
LiteLLM

Efficient Integration of ATOM Cards for High Throughput LLM Service

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

REBEL™

Experience Powerful, Effortless AI Compute



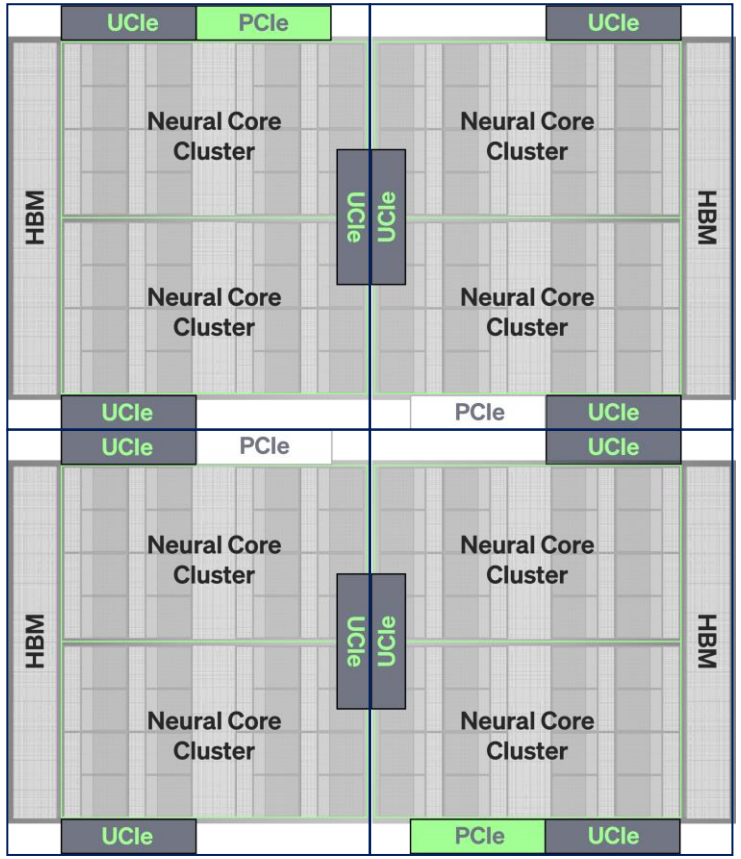
rebellions_

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Chiplet-based GenAI Accelerator REBEL™

2-3X Higher TFLOPS/Watt vs GPUs for Enterprise and OEM Customer



Performance

- Hight Density Compute Engine
  - 1 Peta (FP16), 2 Peta (FP8) compute @300W
- High Bandwidth Memory Architecture
  - 144GB, 4.8 TB/s HBM3E modules

Scalability

- World's first UCIe-ADV based NPU Accelerator
  - Less than 10% area and 5% power
  - 4TB/s BW, 5.2Tbps/mm BW density
- Seamless Data Transaction btw Chiplets
- Large/scalable VA space

Datacenter-proof Features

- Confidential computing
- Extensive RAS supports for SLA

1 Peta FLOPS (FP16) , 2 Peta FLOPS(FP8) Support under 300 Watt Power

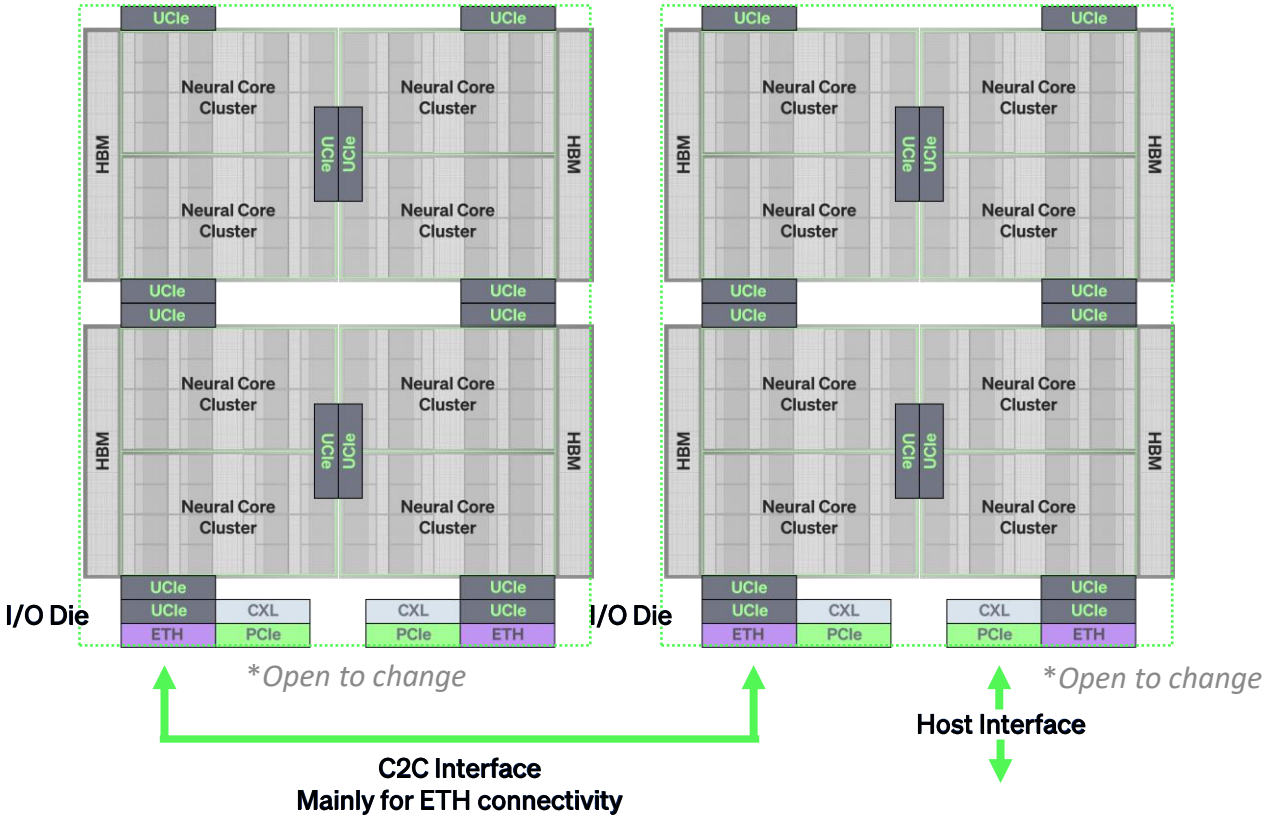
This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

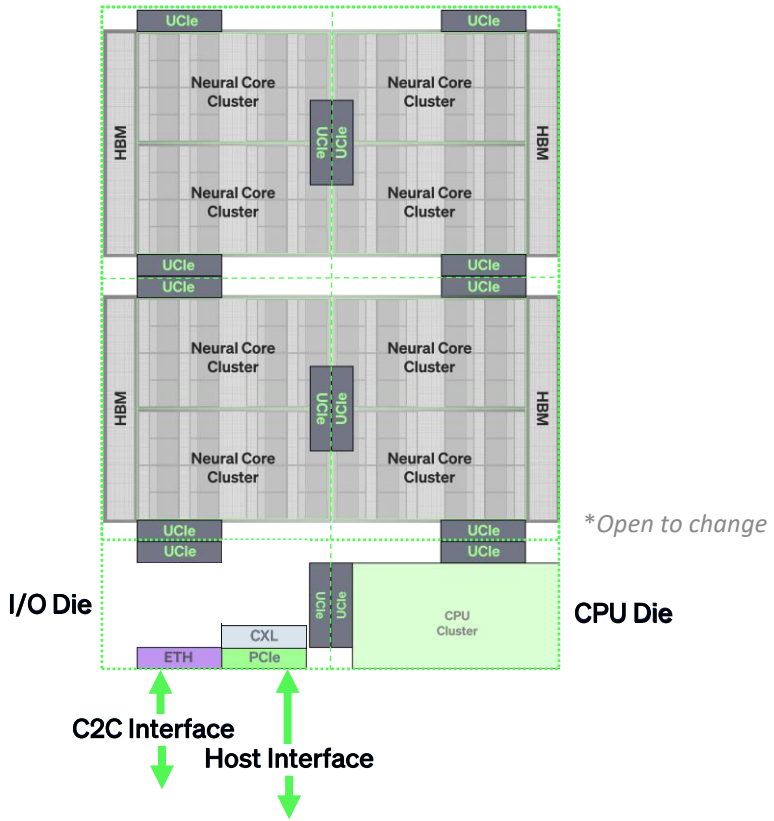
Heterogenous Chiplet Solutions

Solutions for Massive Inference and CPU-centric Ops

REBEL™ x REBEL™-IO
Large-scale Distributed Inference



REBEL™ x REBEL™-CPU
Energy-efficient AI Server



This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

Software Completes Hardware

ML Frameworks TF, PyTorch, mxnet, ONNX,...

RBLN compiler stack

Adaptation Layer (python plug-in)
quantization, graph modification, ...

Compute Library
RBLN_dnn, RBLN_blas,
...

Frontend Compiler
graph IR transform, graph-opt, ...

Backend Compiler
scheduling (comp, mem), code gen, ...

Runtime Driver
runtime module, UMD, KMD, ...

RBLN Backends

ASIC | FPGA | C-model

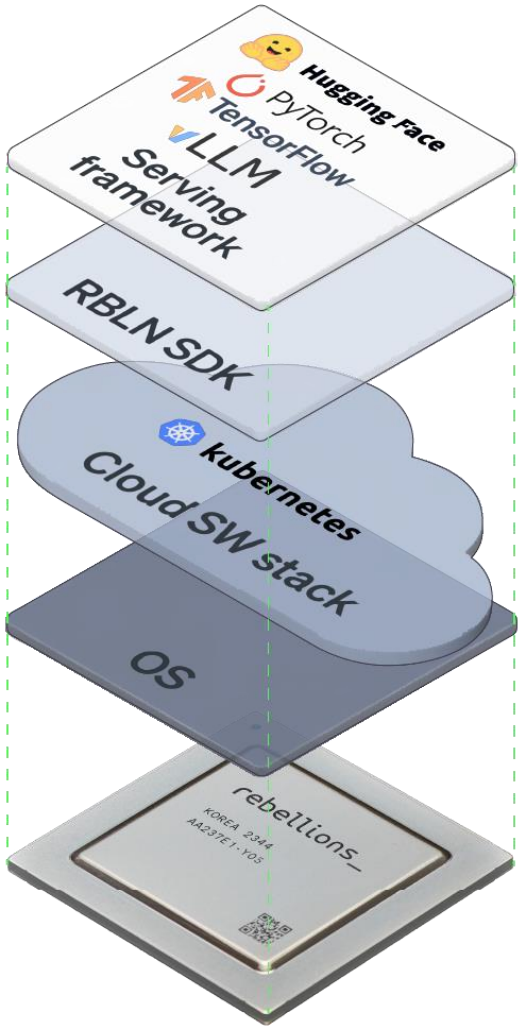
- Seamless Integration into ML Frameworks
- Custom IR (*i.e.* RBLN IR) based Optimization and Scheduling
- Runtime Supports for HW-level Virtualization
- Profiling and Debugging Capabilities

- **Key Highlights of Graph Compiler**
  - Tensor/Layer/Data/Model Parallelization
  - Data Transformation and Communication
  - Performance
    - Overlapping Compute and Transfers
    - On-chip Memory, NoC, Ext. I/O Util. Boost

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Software Completes Solutions



Serving framework

- **Service framework :**
  NVIDIA Triton with Python/C++ backend
  LLM serving framework : continuous batching with vllm-rbln backend
  Customized vllm to ensure peak performance for ATOM

RBLN SDK

- Provides developer-friendly runtime APIs in both Python and C/C++
- Optimum-rbln to natively support models from transformers library

- **Model Zoo :**
  Accommodates majority of models; already providing support through KT Cloud
  - PyTorch Model Zoo
  - TensorFlow Model Zoo
  - Huggingface Optimum- compatible model suite

Rebellions Cloud SW Stack

- Kubernetes and Openstack based cloud SW stack
- SRIOV-based virtualization / monitoring tools

OS

- Ubuntu Linux 20.04 LTS / 22.04 LTS
- RHEL 9.3 and etc. (Rocky Linux)

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Ease-of Use: ATOM™ Data Framework Use Case

Building Performant RAG Applications for Production

@CPU

@ATOM

LlamaIndex Application Code

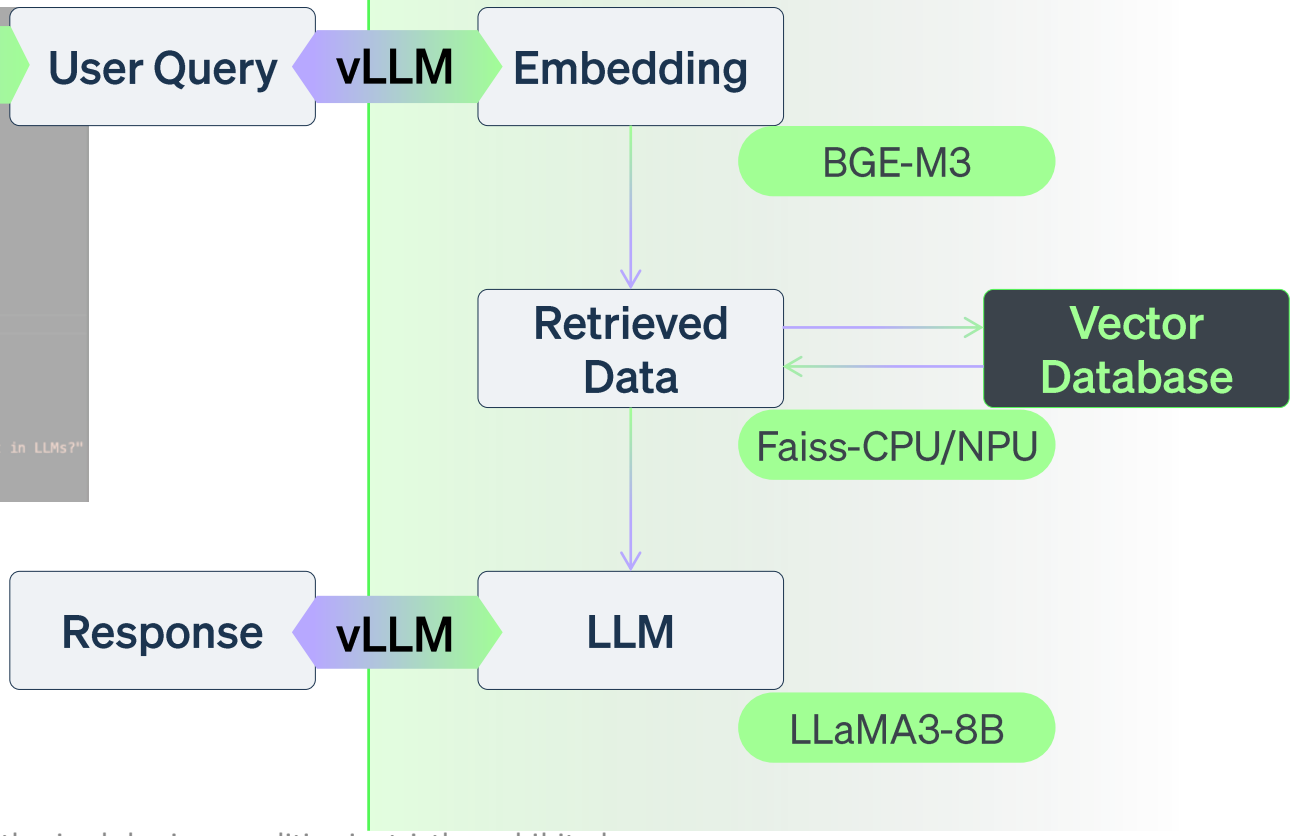
```python
# 1. Set up the RBLNHuggingFaceEmbedding class with the required model to
Settings.embed_model = RBLNHuggingFaceEmbedding(model_name="BAAI/bge-m3")

# 2. Set up the vllm-rbln OpenAI-like endpoint to use with llamaindex core
from llama_index.llms.openai_like import OpenAILike
Settings.llm = OpenAILike(
        model="meta-llama/Meta-Llama-3.1-8B-Instruct",
        api_base="http://localhost:8000/v1",
        api_key="EMPTY"
    )

# 3. Load vector store and create storage context
vector_store = FaissVectorStore.from_persist_dir(persist_dir="./llama_arxiv_papers")
storage_context = StorageContext.from_defaults(vector_store=vector_store, persist_dir="./llama_arxiv_papers")

# 4. Load index onto the memory and make it a query engine
index = load_index_from_storage(storage_context=storage_context)
query_engine = index.as_query_engine()

# 5. Pass the input message to the query engine, and get back the response
question = "According to recent Meta publications, what are some typical fine-tuning approach to ensure alignment in LLMs?"
response = query_engine.query(question)
response_str = str(response)
```



External API for Application Dev

OpenAI API

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Business Partners

Rebellions Successfully Commercialized NPUs and Our Partnership Spans Across The Public and Private Sectors

Recent addition to the network

CSP & Telco

SK telecom

kt cloud

NHN Cloud

Global Telco AI Alliance

SK telecom, T, e&, Singtel, SoftBank

Enterprise

aramco

kakao, kt

IGLOO

KONAN TECHNOLOGY

Gen AI for the Public Sector

과학기술정보통신부 Ministry of Science and ICT

산업통상자원부 Ministry of Trade,Industry and Energy

대한민국 국방부 Ministry of National Defense

KRIT 국방기술진흥연구소 Korea Research Institute for defense Technology planning and advancement

육군군수사령부 ARMY LOGISTICS COMMAND

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

rebellions_

Rebellions Technical Highlights for LLM Serving

Technical Present and Futures

Chip

- Chiplet-based Scalable Solutions
- HBM3E
- High Density Neural Engine

→ **Scalable Chiplet Solutions**

SW

- Pytorch 2.x , vLLM
- Triton-like Low-level API Support for Flash-Attention
- K8S and VM based Cloud Customer Software Stack

→ **Open Source & Ecosystem**

System

- Rebellions Scalable Design
- Monitoring and Profiling Tool for Higher SLA
- RAS Support

→ **Energy-efficient Datacenter Solutions**

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.

Redefine AI Compute

Thank you

rebellions_

This presentation is intended solely for the audience of this session only. Unauthorized sharing or editing is strictly prohibited.