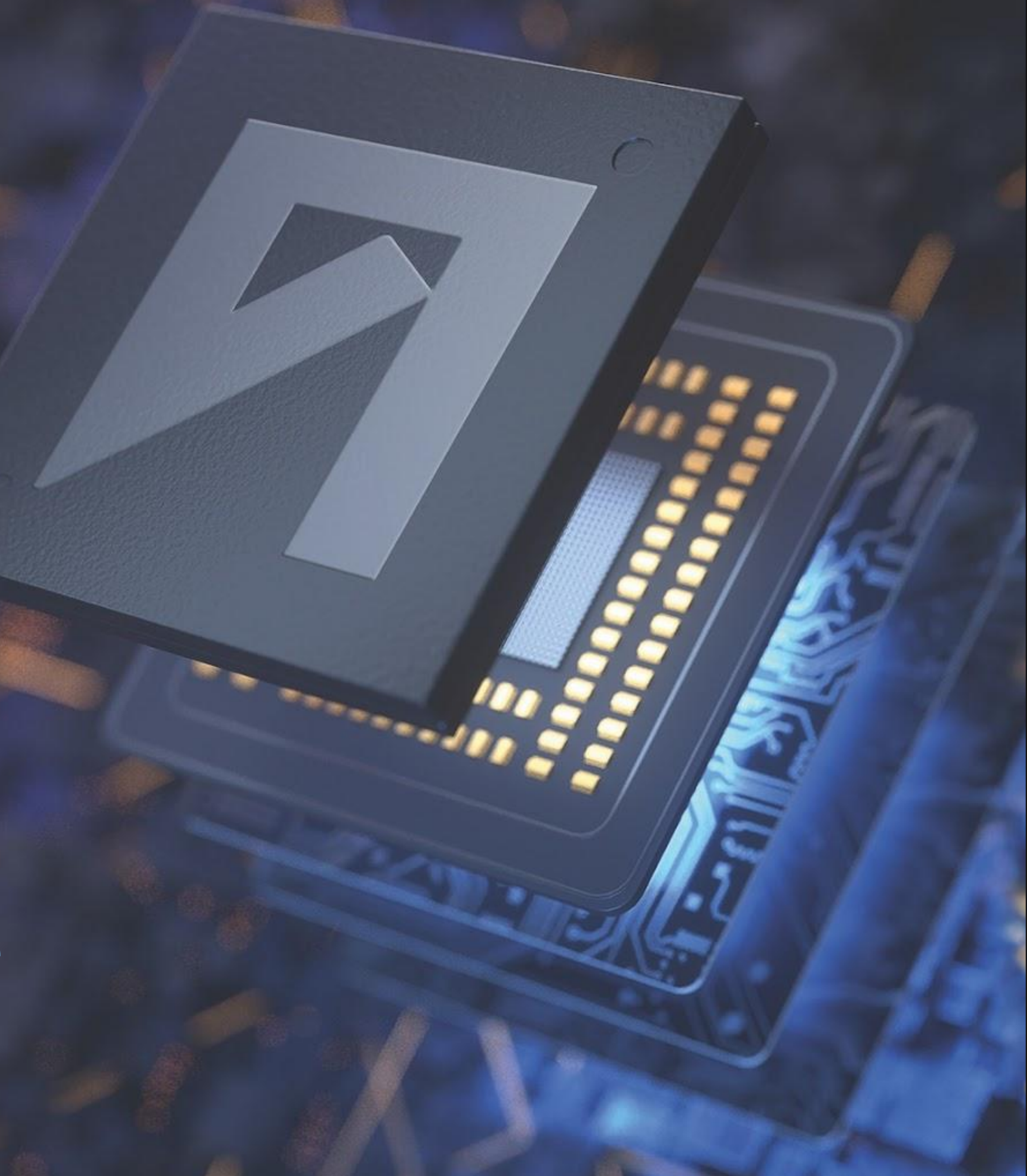
Breaking Performance Barriers

RISC-V Summit 2025. Paris, France

Graham Wilson – Product Group



Processor IP company (RISC-V)

Proven Performance Legacy

Team comprised of veterans from ThunderX2 ARM server chip development, Oracle, MIPS, Intel, Google

KLEINER PERKINS

Mayfield

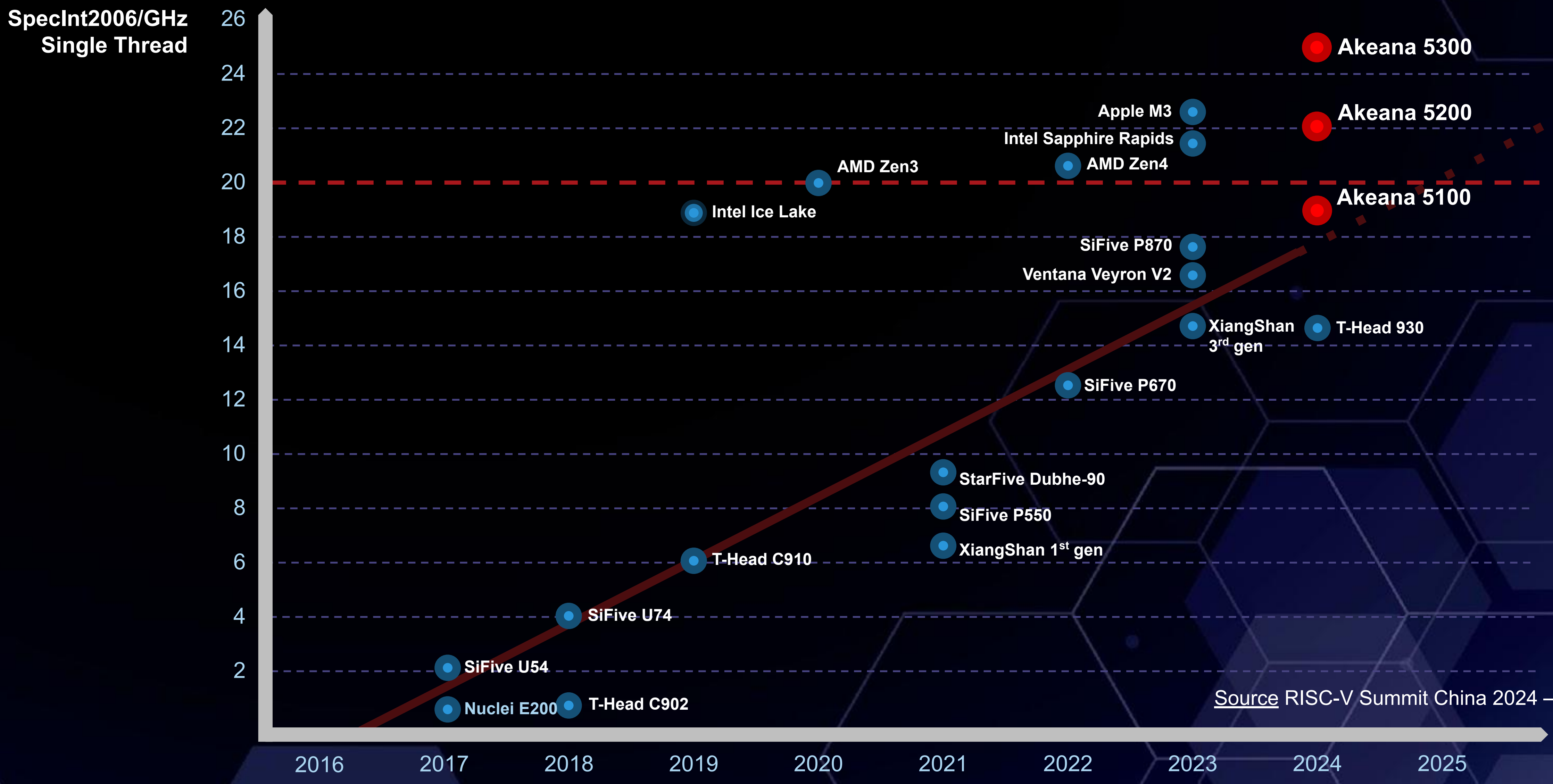
Fidelity INVESTMENTS

Akeana’s drive is to enable our SoC customers to success with leading edge performance processor system IP

Akeana Processor IP Product Lines

Akeana 100 Series **Micro-controller, Embedded**	32-bit, PMP. 32-bit physical addressing Single to Dual issue In-Order architecture 4-stage to 9-stage pipeline Private L1. Shared Coherent L2 Cache Local ICCM, DCCM	**Equivalent to ARM Cortex-M and Cortex-R**
Akeana 1000 Series **Consumer, Automotive**	64-bit, MMU. Up to 57-bit virtual addressing 2- to 4-wide issue In-Order architecture, with 5-stage to 9-stage pipeline 12-stage Out-of-Order architecture (2,3,4-issue) Private L1, L2 Caching. Shared Coherent L3 Cache Vector Extension (up to 2048 bit). AI Acceleration Multi-Threaded support (up to 4 threads)	**Equivalent to ARM Cortex-A**
Akeana 5000 Series **Mobile / Data Center, Ultra Performance**	64-bit, MMU. Up to 57-bit virtual addressing 6-wide to 10-wide issue Out-of-Order architecture 12-stage pipeline Private L1, L2 Caching. Shared Coherent L3 Cache Vector Extension (up to 512-bit). AI Acceleration Multi-Threaded support (up to 4 threads)	**Equivalent to ARM Neoverse N2, N3**

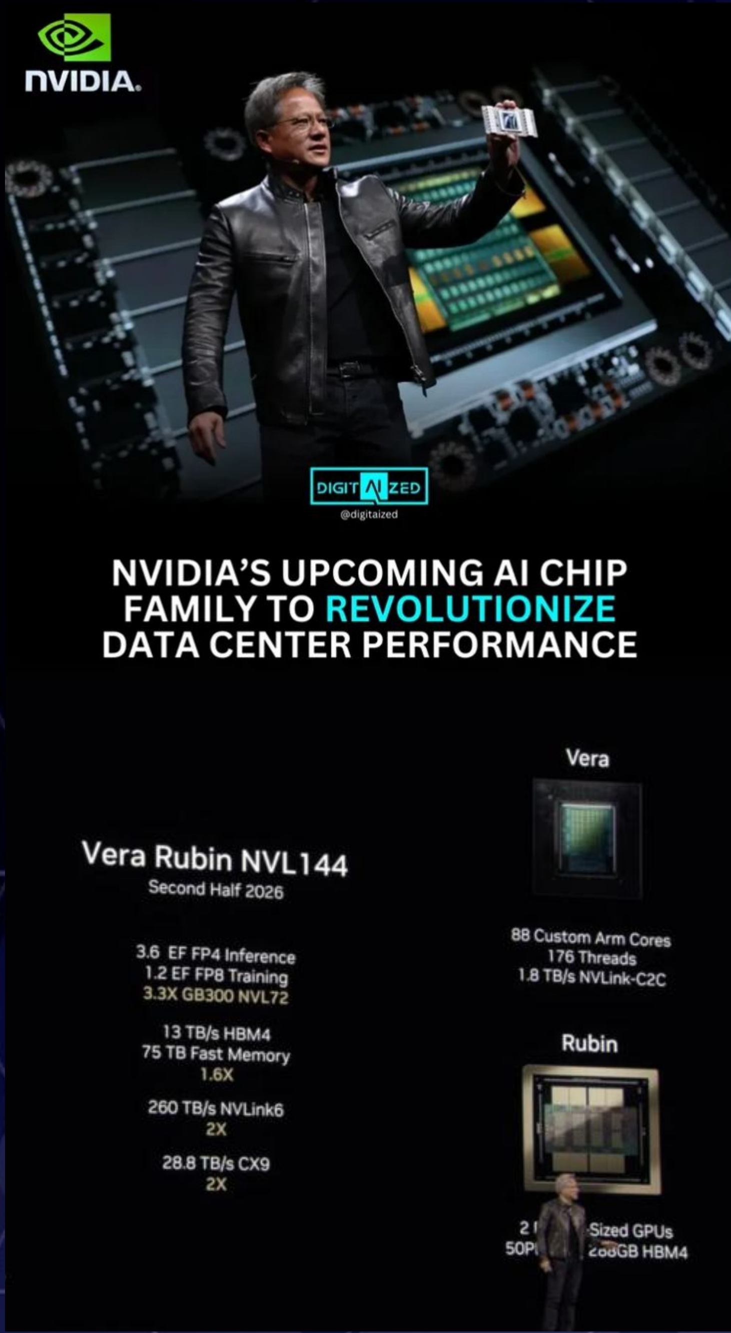
Akeana, Leader in Core Performance



Source RISC-V Summit China 2024 – Jianyi Meng

AI CPU Compute Evolving

- NVIDIA shifted towards a customizable CPU implementation to achieve required computation performance for AI / HPC
- To achieve required performance in CPU compute array system, Simultaneous Multi-Threading (SMT) has been used
    - NVIDIA has planted the SMT flag for AI CPU compute
- AI SoC developers are recognizing this push to higher performance in the CPU Compute array
- Akeana is the leading provider for Multi-core, Multi-Threaded, Coherent Infrastructure IP
    - Supported in Akeana 1000 series(In-Order), 5000 series(Out-of-Order), up to 4 threads



Performance Boost with SMT

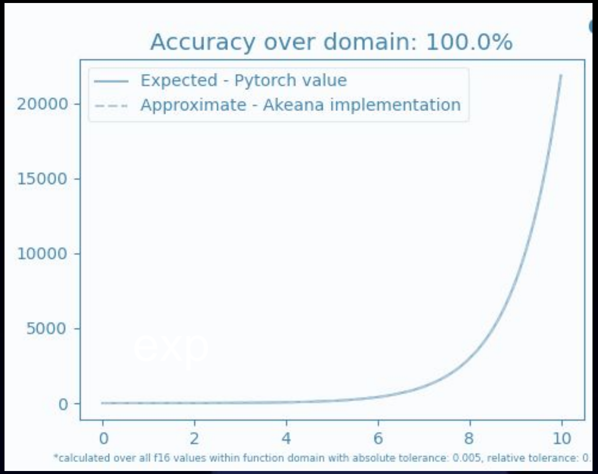
DataBase Processing		1 thread	2 threads	4 threads
	Performance Increase *	1.00	1.79x	2.25x

SpecINT Industry Standard Testbench		1 thread	2 threads	4 threads
	Performance Increase *	1.00	1.18x	1.28x

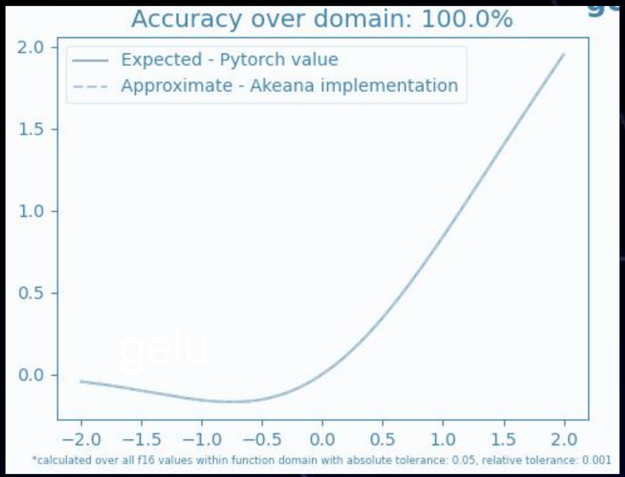
* Numbers based upon current results, subject to change

Performance Data Compute Enabled

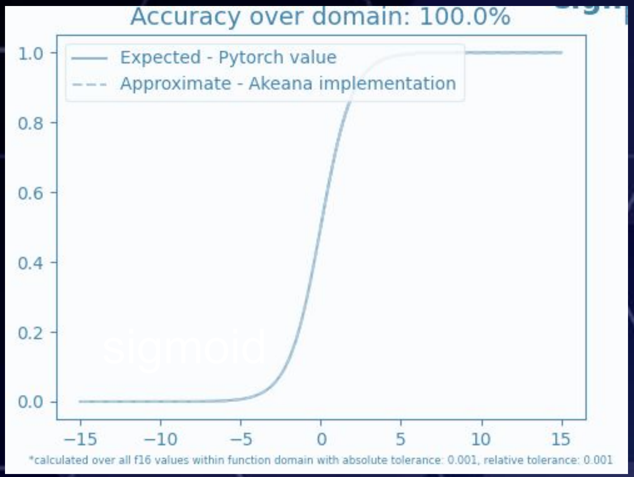
- Akeana AI Nonlinear acceleration instructions provide >10x performance increase (example FP16 datatype with sigmoid, gelu, tanh, exp functions)
  - Benefit of lower core count needed, reducing area and power consumed
  - Supported within Akeana AI Performance library functions
- Implemented with RISC-V Vector Extensions up to 2048-bit VLEN for further data compute acceleration



*calculated over all f16 values within function domain with absolute tolerance: 0.005, relative tolerance: 0.



*calculated over all f16 values within function domain with absolute tolerance: 0.05, relative tolerance: 0.001

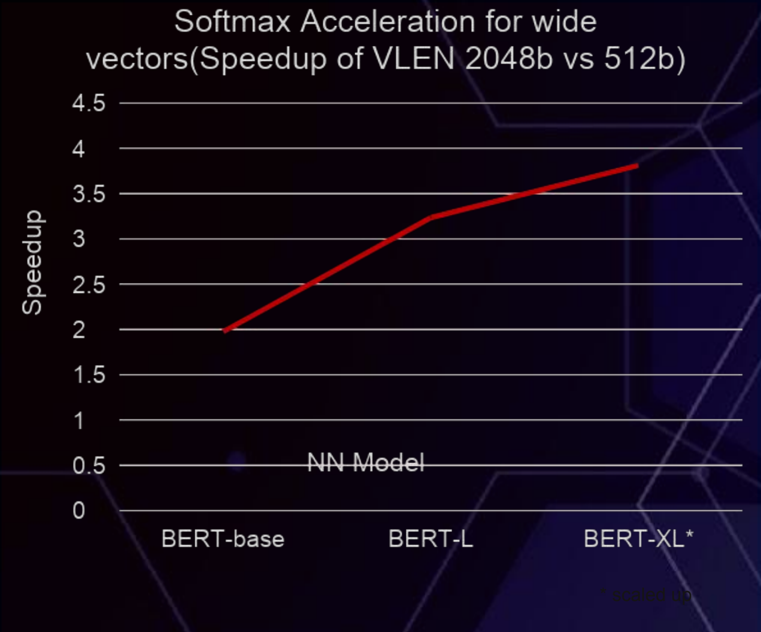


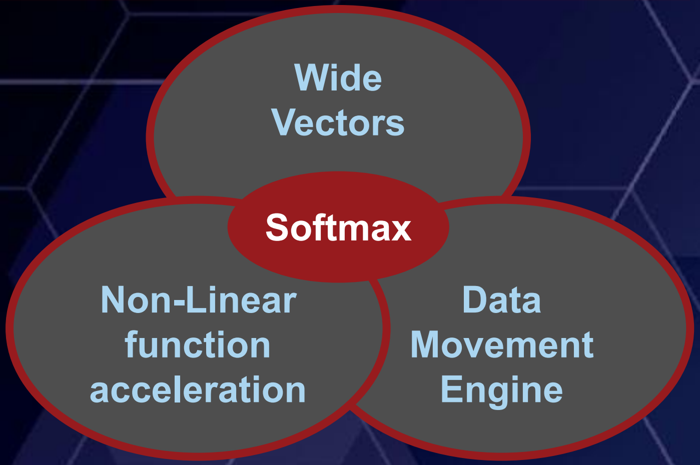
*calculated over all f16 values within function domain with absolute tolerance: 0.001, relative tolerance: 0.001

Accelerating Softmax

- Softmax needs acceleration in 3 domains; vectors, Nonlinear functions and data movement
  - Akeana data movement engine IP available to further accelerate Softmax implementations

- Example Softmax implementation running through various vector length computation
  - Auto-vectorization compiler able to efficiently map vectorizable code to Akeana vector cores
  - Ability to map over range of Transformer based models

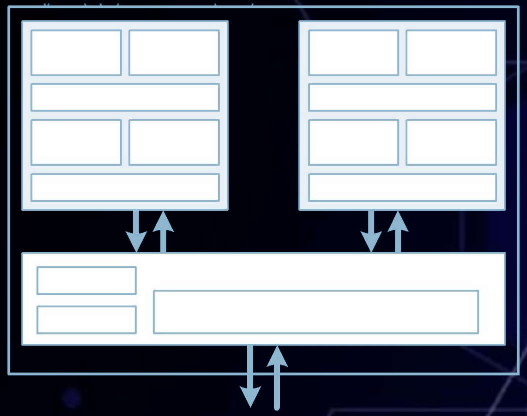




Performance through Scalability

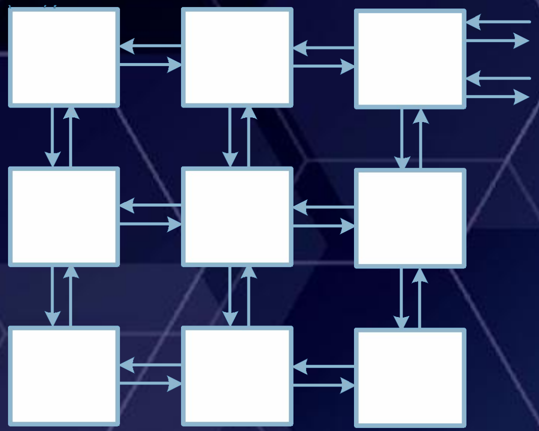
Single Coherent Cluster

- Utilizes Compute Coherent Block (CCB)
- Shared Coherent Cache, accessed in parallel by all cores
- Up to 8-cores, programmable engines, coherent operation
- Easy scalability to coherent 8-core system
- AI accelerator (GEMM), Customers hardware engines



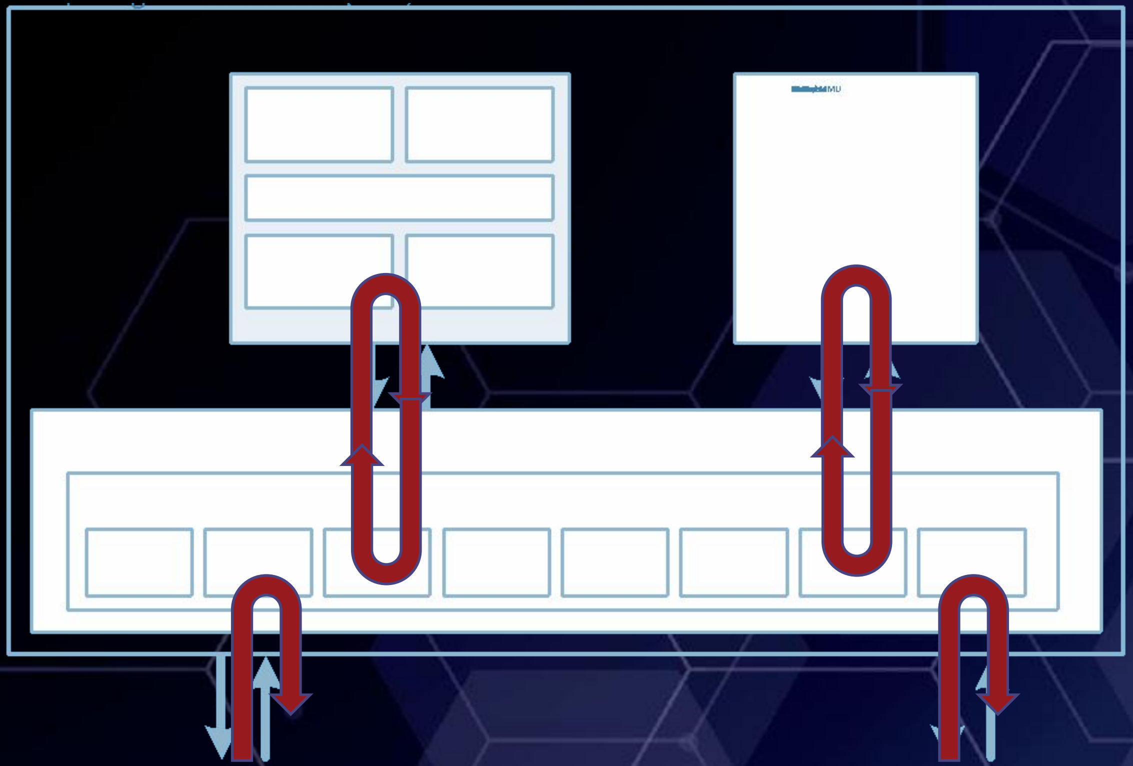
Scalable Coherent Mesh (Akeana Mesh)

- Utilizes multiple Akeana IP blocks to build up 2D Mesh array
- Single Coherent Cluster (CCB) integrated into 2D Mesh
- AMBA CHI compliant
- Can be built to connect up to 100’s of cores, fully coherent
- Akeana provides all the IP blocks, and works with customers to build these larger 2D Mesh coherent interconnect systems



Optimized Data Movement Performance

- Shared memory can be partitioned into separately accessible banks
  - Allows parallel accesses from multiple processors, Matrix Engines
  - Ping-Pong-ing of Banks when processing of large amounts of data between cores
- 2nd external port to AI CCB block
  - Allows dedicated high bandwidth Matrix Data (Activation and Weights) to be pulled into Shared L2 without blocking from Processor accesses



Performance CPU Compute

Performance Data Compute

Performance Scalable Multi-core

Performance Data Movement

Thank you

AKEANA.COM