AMPERE®

Sustainable Computing for AI & Cloud Native Workloads

Matthew Erler - Architect

Aug 27, 2024

A MODERN SEMICONDUCTOR COMPANY BUILDING

THE FIRST CLOUD NATIVE PROCESSORS

FOR THE SUSTAINABLE CLOUD

Traditional Techniques No Longer Scale		Ampere® Cloud Native Processors Do
Turbo Frequency	Paradigm Shift	Power Optimized, Consistent Performance
Hyperthreading		Linear Core Scaling
Scale Up Accelerators		High Performance, General-Purpose Cores

The Ampere Roadmap: Powerful Roadmap with Rapid Innovation

Continued Commitment to Leadership Performance Per Rack for AI Compute in Air Cooled Environments

AmpereOne® Family

Up to 192 Cores 5nm
8 Ch DDR5

AmpereOne®

Shipping Now

Up to 192 Cores 5nm
12 Ch DDR5

AmpereOne® "M"

Shipping Q4 '24

Up to 256 Cores 3nm
12 Ch DDR5

AmpereOne® "MX"

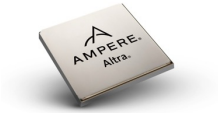
In Fabrication

Up to 512 Cores
Integrated AI Silicon
Training and Inference
Air Cooled

AmpereOne® Aurora

Next Design Product

Ampere® Altra® Family



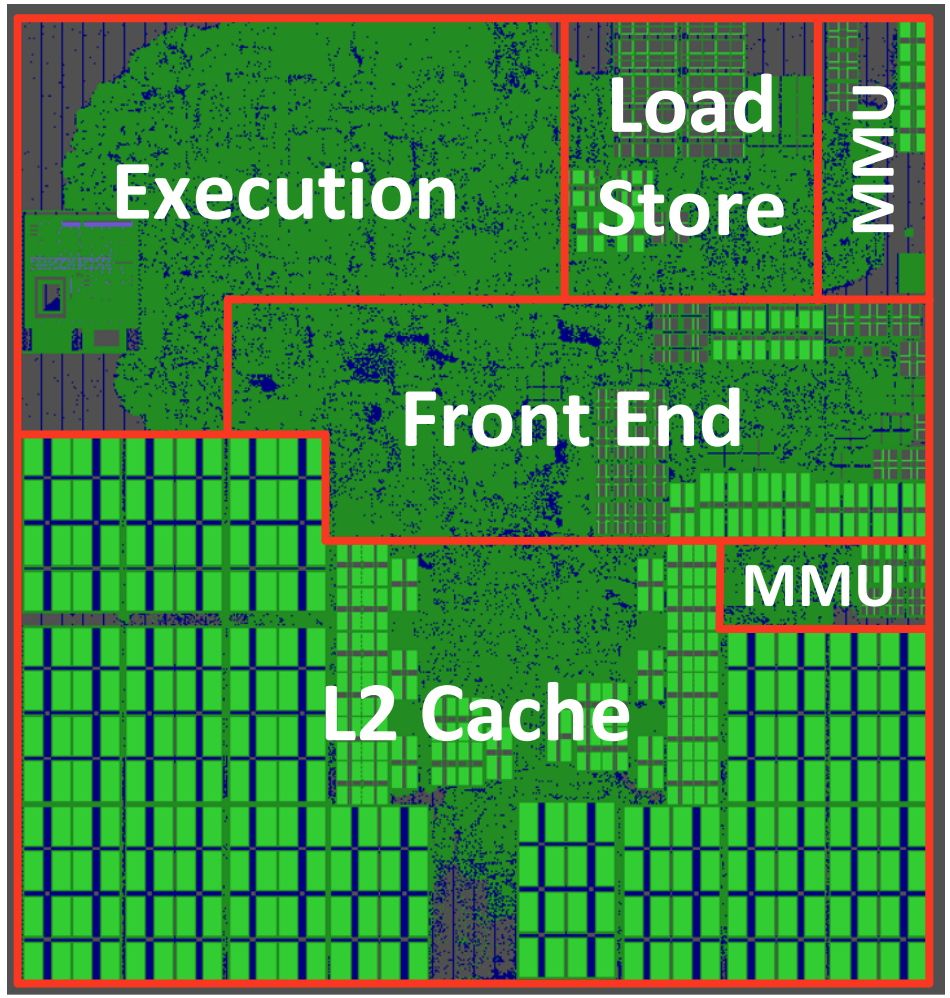
Up to 80 Cores 7nm
8 Ch DDR4
128 Lanes PCIe Gen4

Up to 128 Cores 7nm
8 Ch DDR4
128 Lanes PCIe Gen4

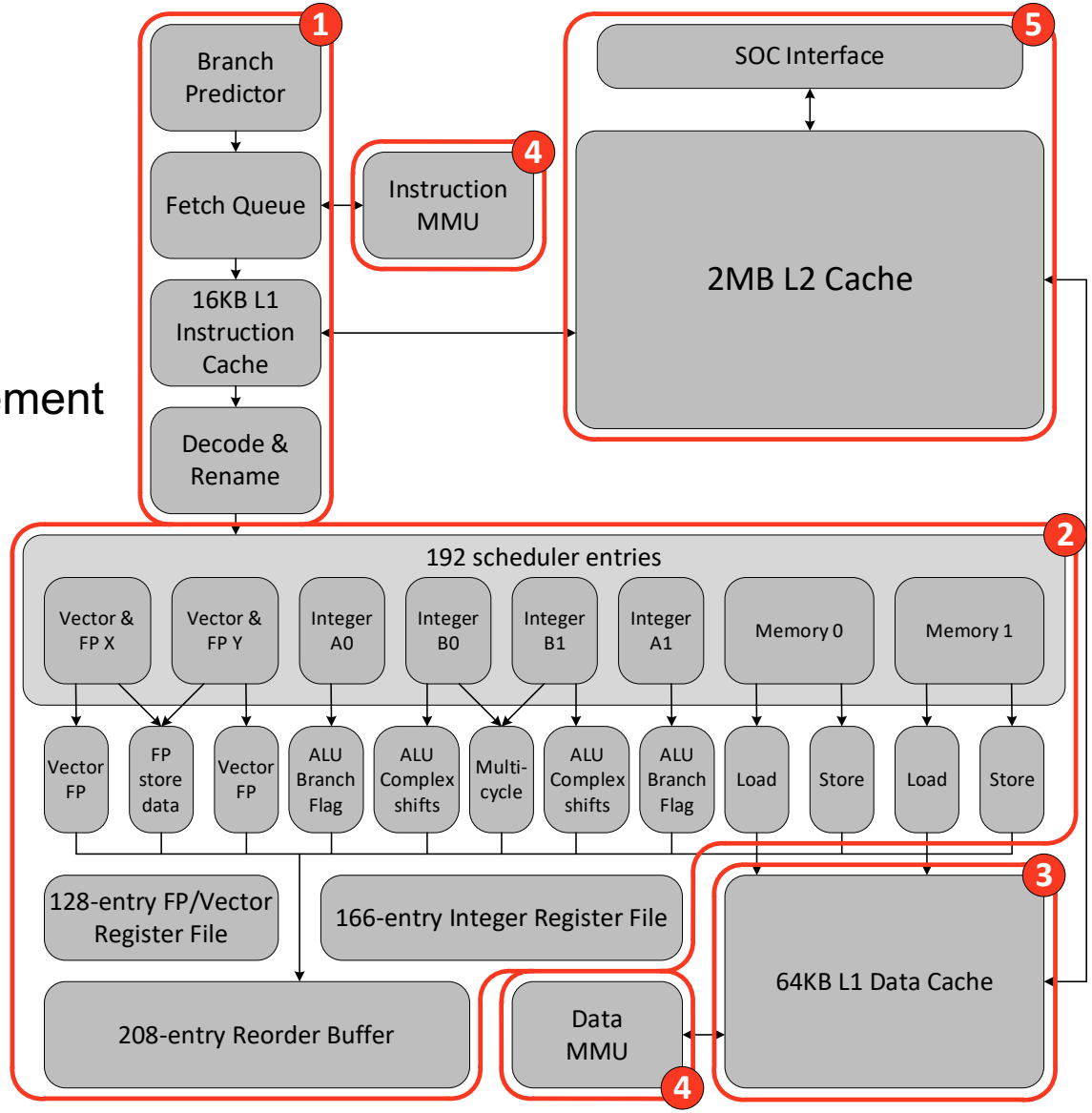
Continued Ship Support at Least Through 2030

*Remains Arm ISA Compliant.

AmpereOne® Core: Overview

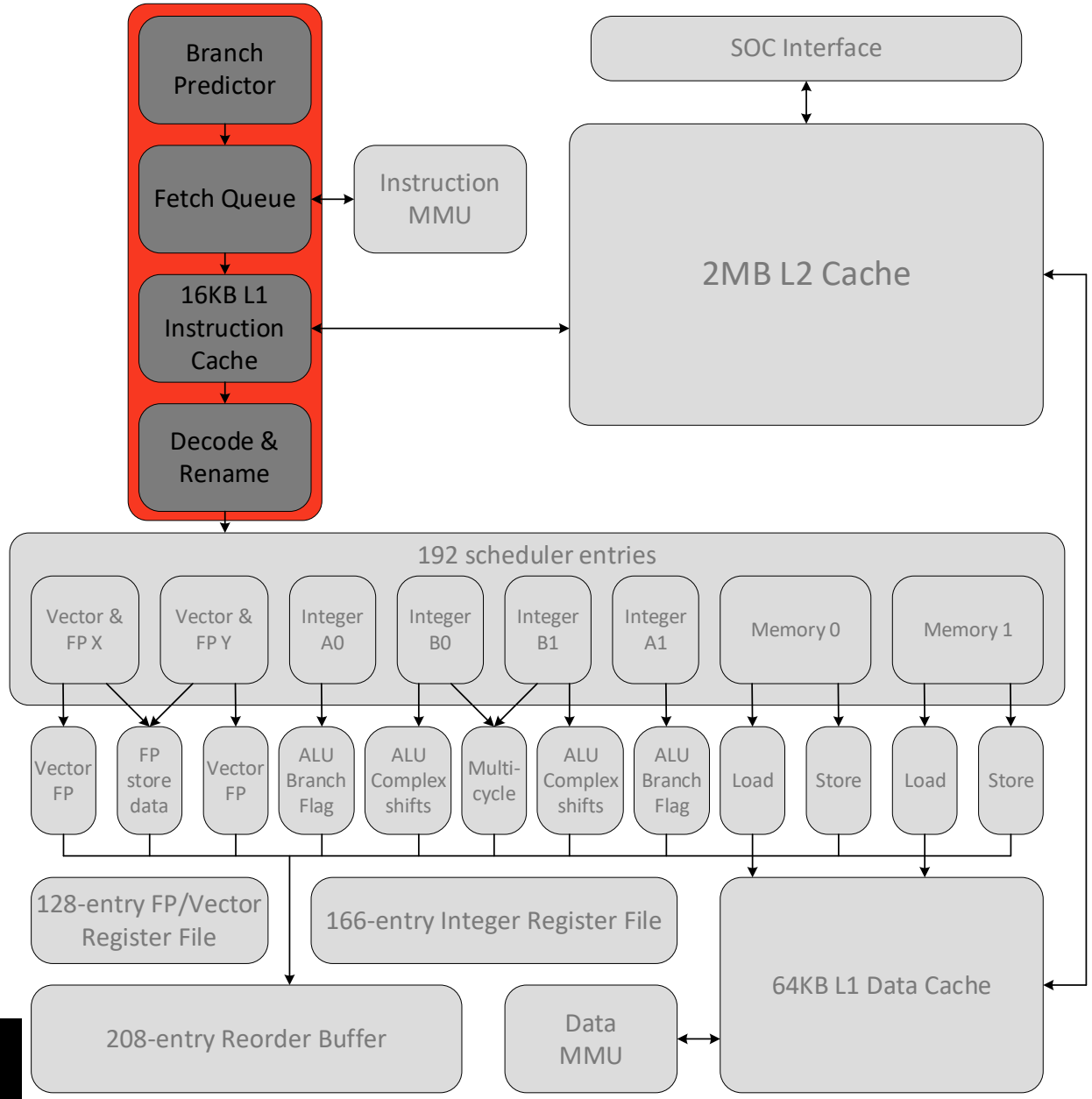


1. Front End
2. Execution
3. Load Store
4. Memory management
5. L2 Cache



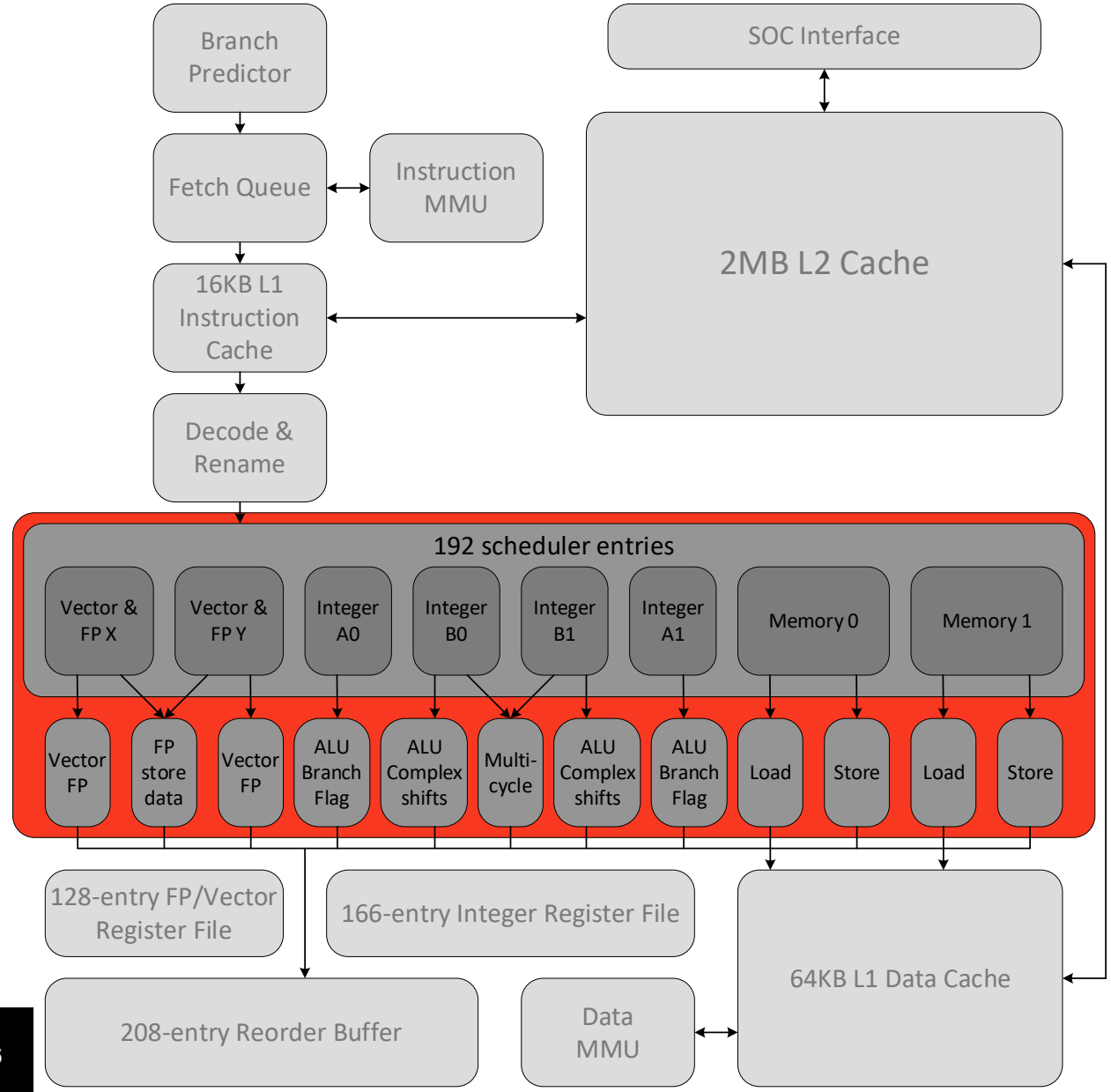
AmpereOne Core Pipeline: Fetch, Decode and Rename

- State-of-the-art branch prediction
  - 8-table TAGE direction predictor
  - L1 and L2 BTB
    - 256 entry 0-cycle
    - 8k entry 2-cycle
  - Dedicated indirect predictor
  - 10-cycle branch mispredict recovery
- Decoupled prediction & fetch pipelines
  - Low-latency, high-bandwidth interface streams code from L2 to support large code footprints
  - 32-entry instruction fetch queue
- L1 Instruction Cache
  - 16KB, 4-way set associative
  - Delivers up to 8 instructions per cycle to decode
- Decode 5 instructions → 4 micro-ops per cycle
  - Fusing of common instruction pairs into single micro-op



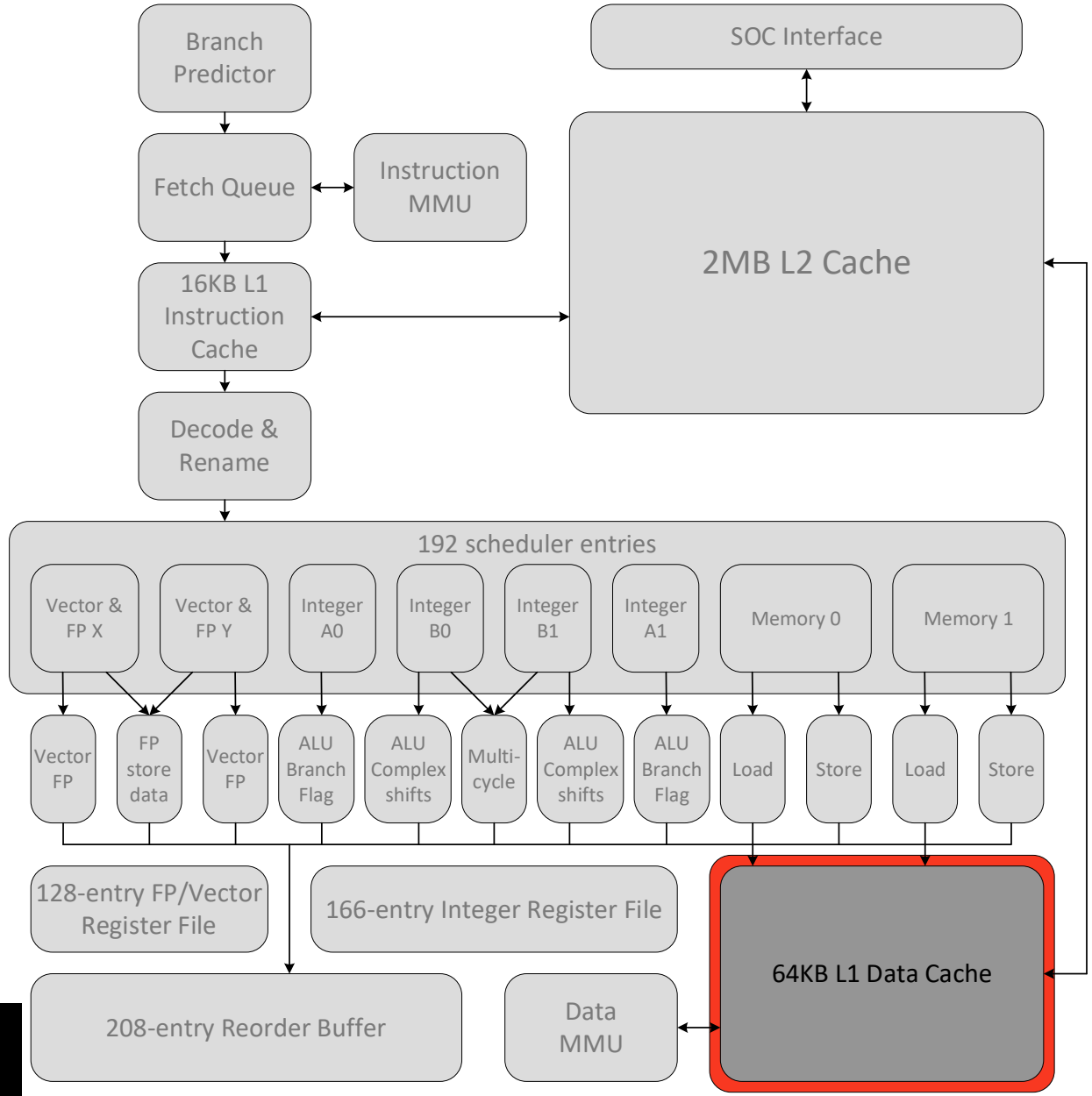
AmpereOne Core Pipeline: Execution

- 8 schedulers feed 12 execution pipes
  - Deep, power-optimized schedulers enable better out-of-order execution
- 2 load and 2 store pipes, each with full AGU
  - Memory schedulers can each issue 1 load and 1 store per cycle
  - Early store address resolution reduces memory disambiguation errors
- Largely symmetric integer and FP/vector execution pipes
  - More efficient use of execution hardware
  - More consistent performance
  - Reduced data movement conserves energy
- Execution highlights
  - 2 branches per cycle
  - Single-uop int8 MMLA for AI inference throughput
  - Bfloat16 and FP16 on both vector pipes
  - AES on both vector pipes



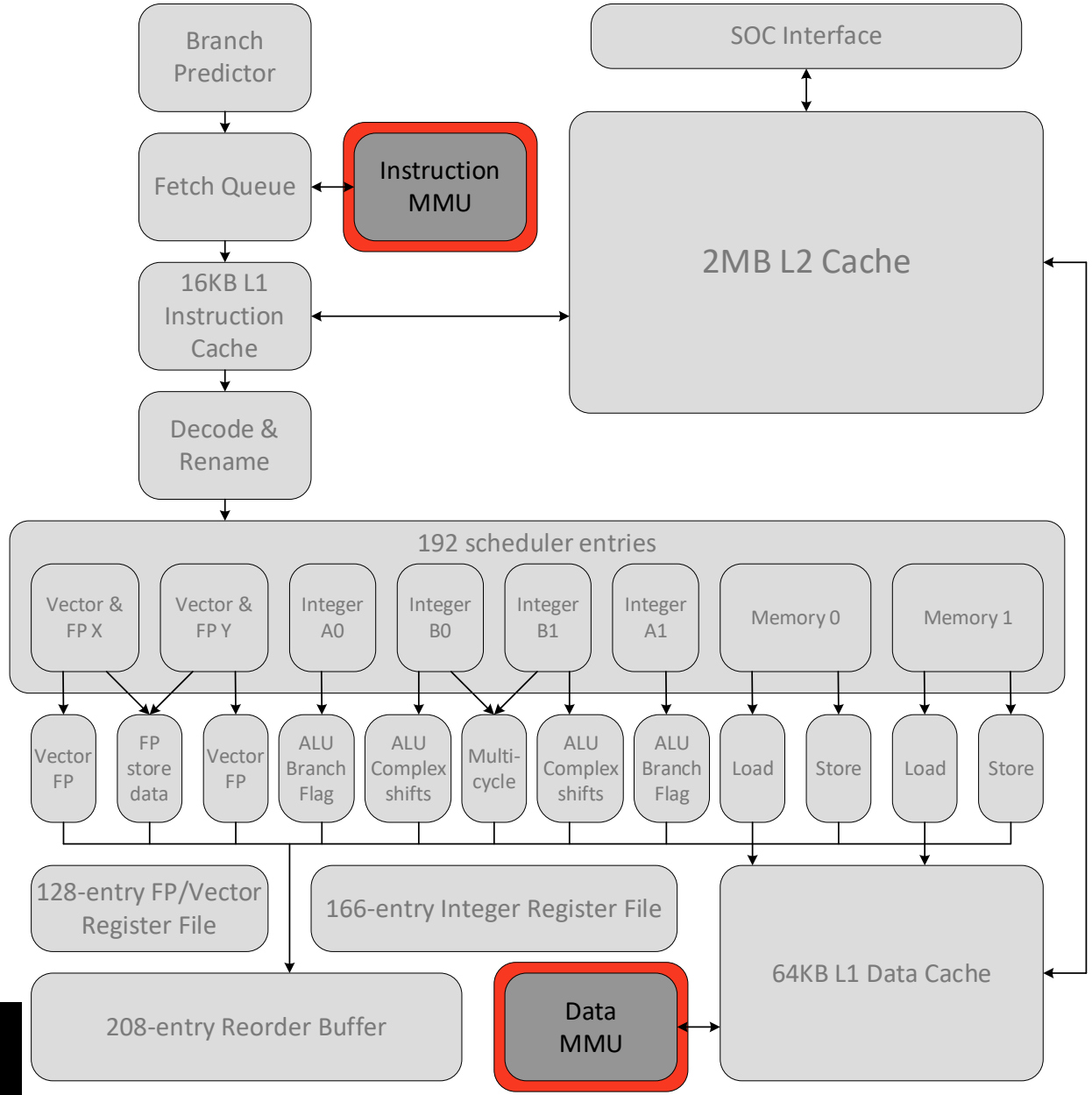
AmpereOne Core Pipeline: Load Store Unit

- 64KB, 4-way write-through DL1 cache
  - 4-cycle integer load-to-use latency for all addressing modes
  - 2x128-bit reads and 1x128-bit write per cycle
- Large store forwarding predictor
- Highly accurate L1 prefetcher
  - Minimizes excess bandwidth in a many-core SOC
- Ground-up Meltdown protections
  - Loads with permission violation will never provide data to any dependent operations
  - Multiple layers of protection for “defense in depth”
- Performant Memory Tagging implementation
  - Intended for use in production environments



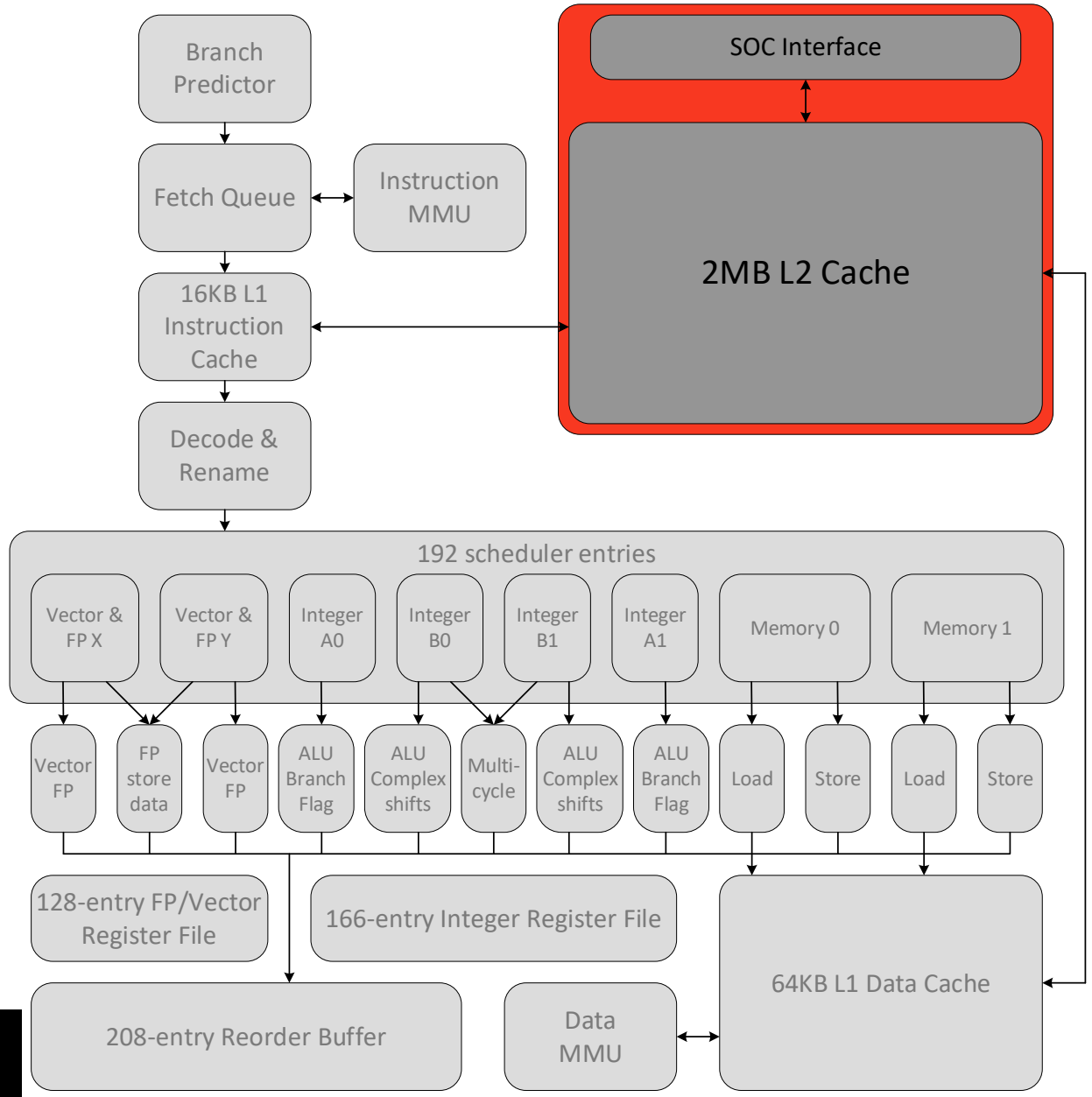
AmpereOne Core Pipeline: Memory Management

- All entries in all TLBs are ‘universal’
  - Any page size can be stored in any entry
  - Optimal support for large memory footprints and huge pages
- Level 1 TLBs
  - DL1 TLB: 64 entries, fully associative
  - IL1 TLB: 64 entries, 4-way associative
- Level 2 TLBs
  - IL2 TLB: 768 entries, 6-way associative
  - DL2 TLB: 1536 entries, 6-way associative
- Up to 8 simultaneous page walks each for instructions and data
- Dedicated L2 interface for page walks
  - Does not pollute L1 cache
- Optimized TLB maintenance response time
  - Improves performance in multi-tenant cloud systems



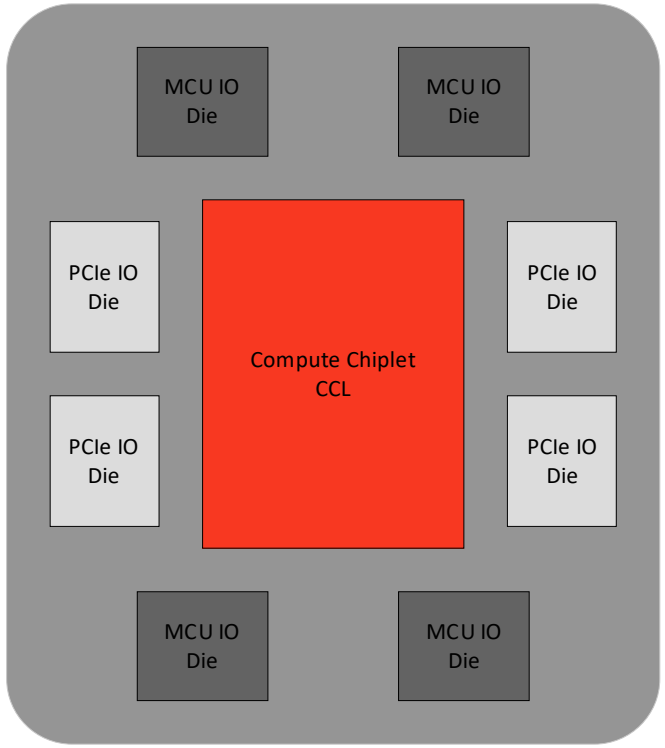
AmpereOne Core Pipeline: L2 Cache

- 2MB 8-way private L2 data/instruction cache
  - 11-cycle load-to-use latency
  - Can schedule read and write every cycle
  - Deliver full 64-byte line to L1 per cycle
  - Tracks 48 outstanding requests to mesh
- Focus on power efficiency
  - Bank size chosen to use most efficient SRAM
  - Centralized control and scheduling logic reduces power consumption in data banks
- Efficient age- and resource-based scheduler
  - Minimizes idle time due to bank conflicts
- Adaptive throttling of request rate and prefetches based on mesh traffic
  - Maximizes SOC bandwidth by preventing mesh congestion
- Multiple L2 prefetchers
  - Including modified best-offset prefetcher
  - Priority-based prefetch queue gives preference to higher-accuracy prefetches



AmpereOne® Disaggregation

AmpereOne
8 Channel DDR5



Compute, memory & PCIe subsystems on separate die

- Each on “best” manufacturing process

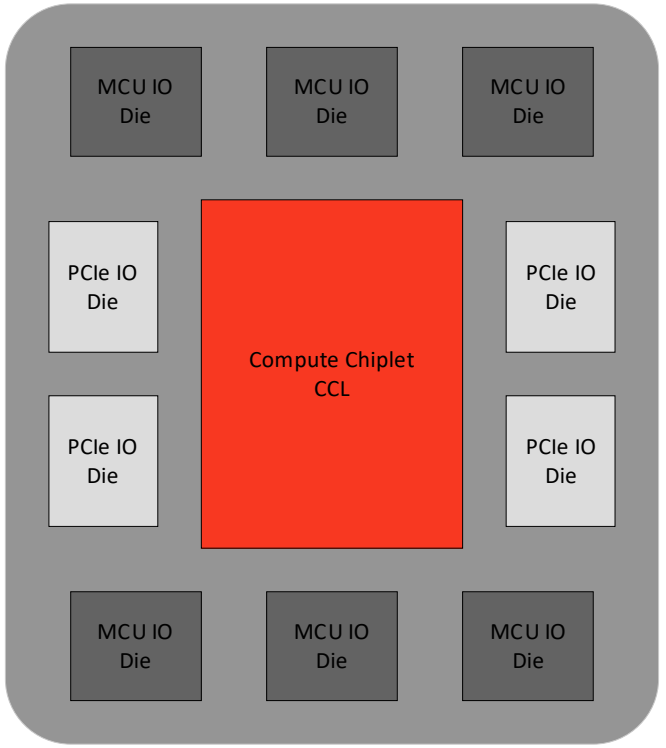
Connected via Ampere custom D2D interconnect

- Up to 2.8TB/s in each direction

Flexible and efficient architecture

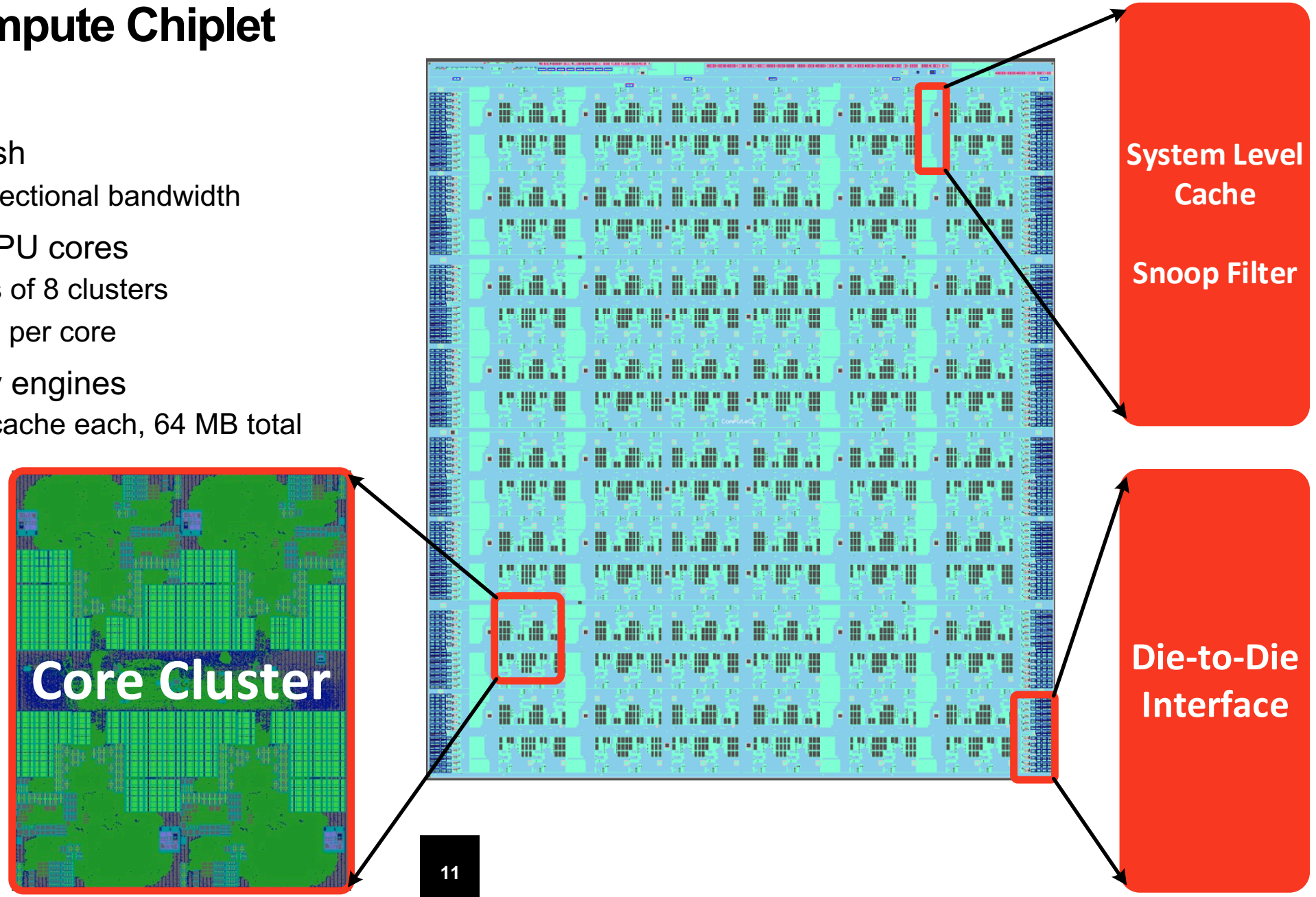
- Enables 8- and 12-channel designs with same building blocks
- Allows for rapid integration of customer IP
- Supports customization to fit unique customer IO and memory requirements

AmpereOne M
12 Channel DDR5



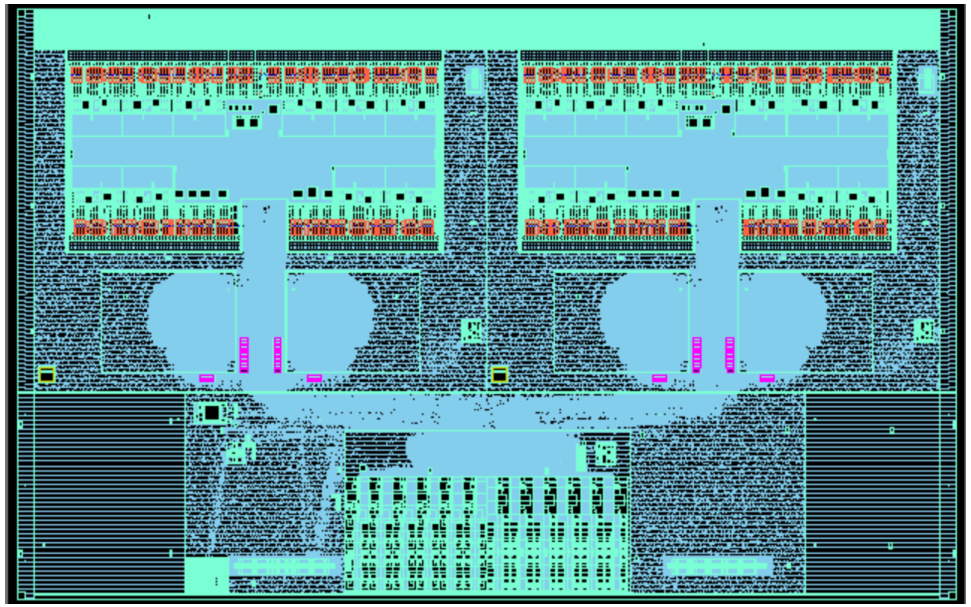
AmpereOne® Compute Chiplet

- TSMC 5nm
- Fully connected 8x9 mesh
  - Up to 5.7 TB/s cross-sectional bandwidth
- 192 Ampere® custom CPU cores
  - Arranged in 6 columns of 8 clusters
  - 2 MB private L2 cache per core
- 64 distributed coherency engines
  - 1 MB of system level cache each, 64 MB total
  - Snoop filter tracking

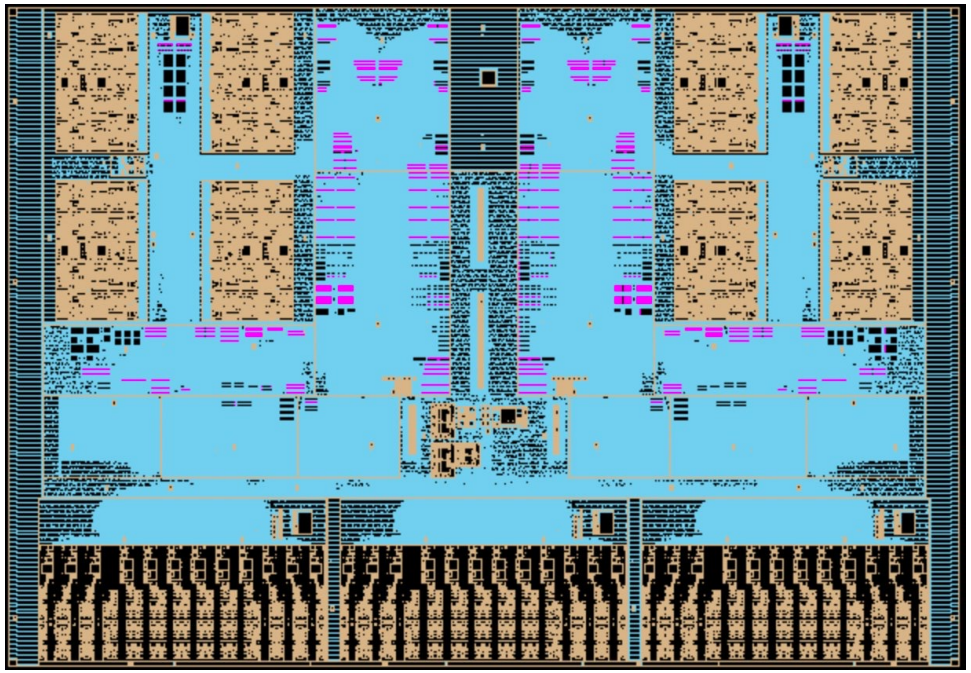


AmpereOne® I/O Chiplets

- TSMC 7nm
- MCU I/O Die
  - 4 dies per package
  - 2 channels DDR5 per die, or 8 channels total
  - Support for up to 16 DIMMs, or 4TB per socket



- PCIe I/O Die
  - 4 dies per package
  - 32 lanes of PCIe 5.0 per die, or 128 lanes total
  - 32 controllers per socket

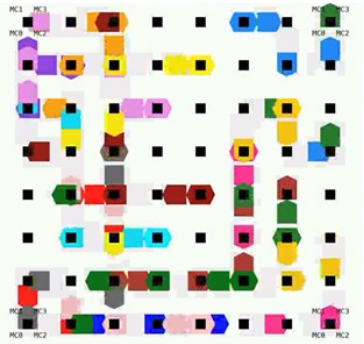


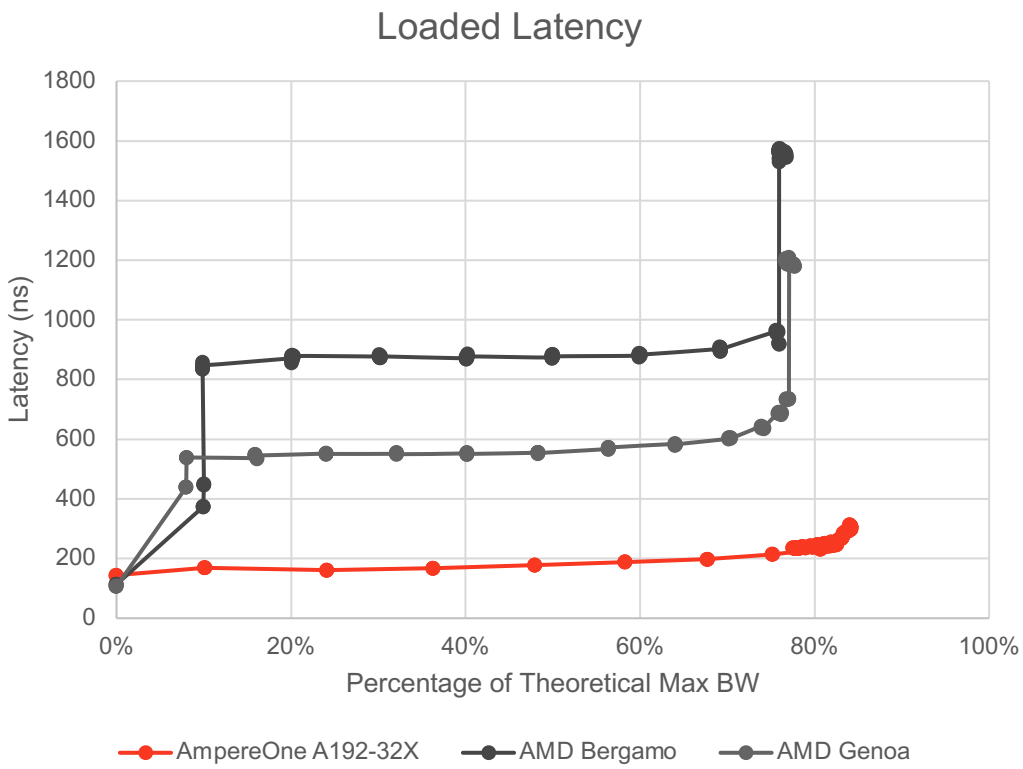
Memory Tagging

- Powerful feature for Robustness & Security
  - **Robustness**: probabilistic detection of pointer programming errors (e.g., “pointer use-after-free”)
  - **Security**: detection & mitigation of attacks that exploit memory safety vulnerabilities (e.g., buffer overflow attacks)
- Operating Principles
  - Every 16-byte memory granule gets a 4-bit ‘allocation tag’
  - Pointers carry an ‘access tag’ in upper bits of address
  - Core checks “access tag = allocation tag” for every memory access
  - Mismatch causes fault and prevents data access
- AmpereOne‘s novel implementation enables run-time use in a production cloud environment
  - Performant, precise memory tag checking for loads & stores
  - No tag-related memory capacity or bandwidth overhead

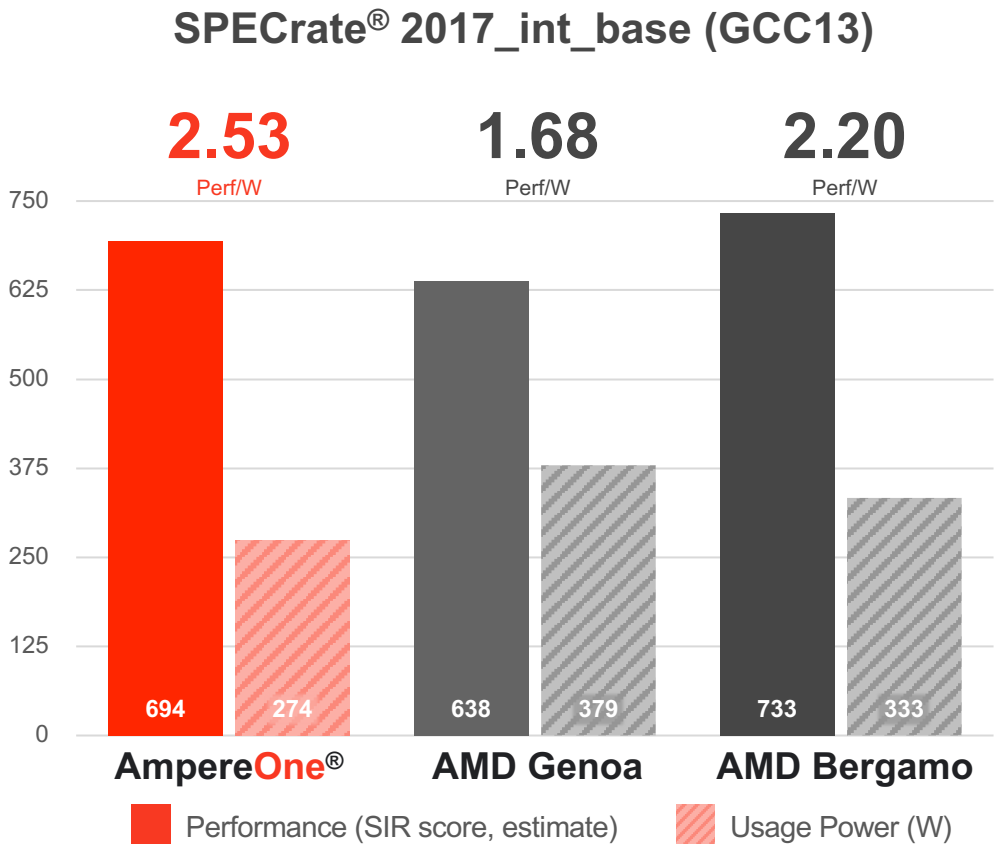
Adaptive Traffic Management

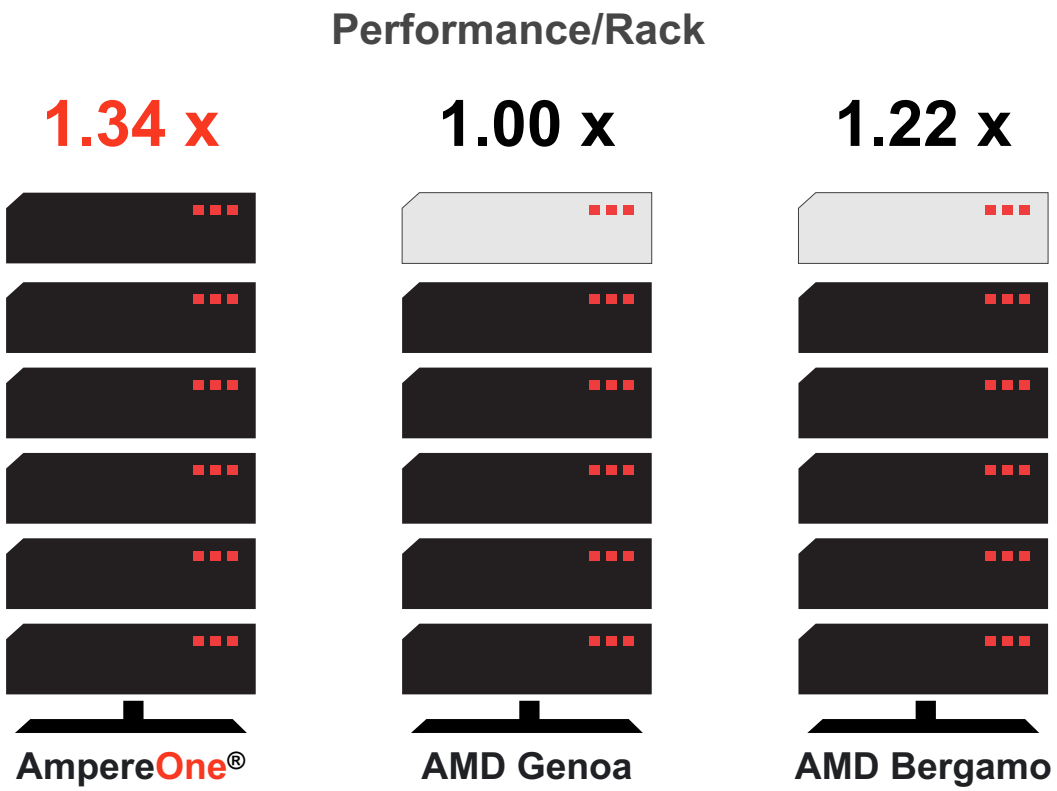
- Vital capability for delivering consistent, scalable performance
  - Consistent: minimize run-to-run variation and interference between independent concurrent workloads
  - Scalable: avoid mesh traffic congestion at high utilization
- Operating Principles
  - Memory service agents advertise their “busyness”
    - Memory controllers, caches
  - Cores modify rate and profile of requested traffic in response
- AmpereOne‘s adaptive traffic shaping allows it run a large, diverse set of workloads at optimal performance
  - Consistent performance at high CPU utilization
  - Adaptive response for different workload behaviors





AmpereOne®: Performance per Rack Leadership





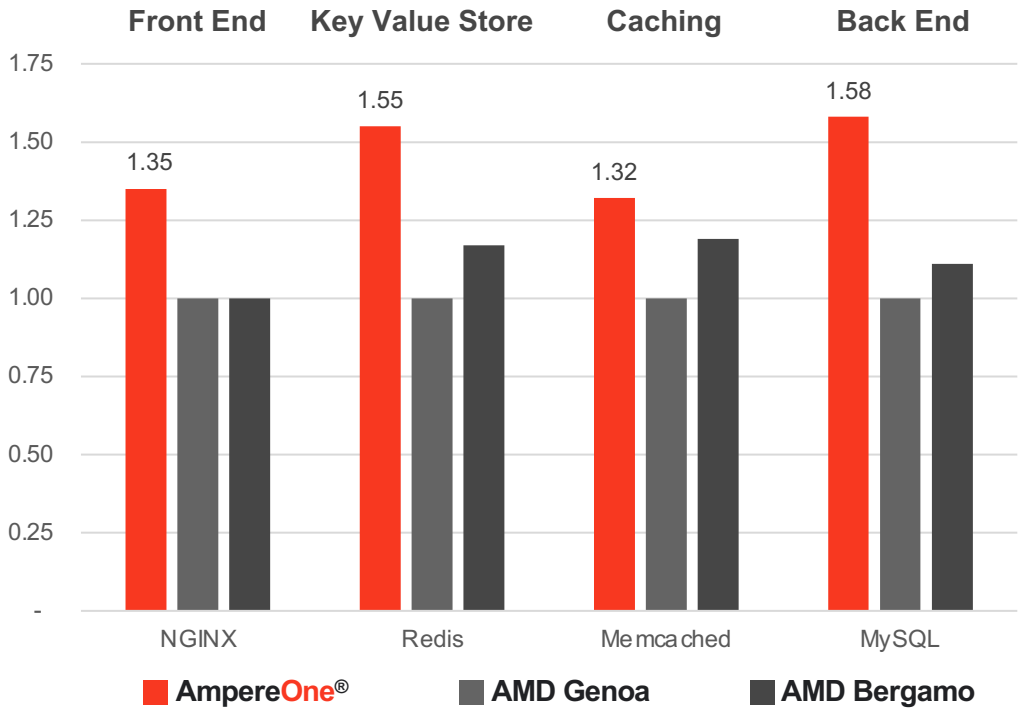
Up to 50% Better
Performance/Watt over AMD Genoa

Up to 34% Better
Performance/Rack over AMD Genoa

Rack depictions for conceptual illustration purposes only

References: slide 27

Breakthrough Performance Per Rack for Cloud Native Workloads

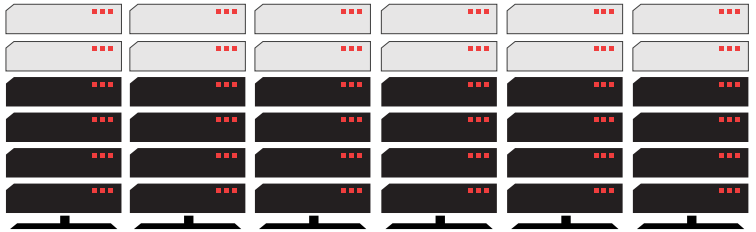


AmpereOne®

33%
Less
Rack Space

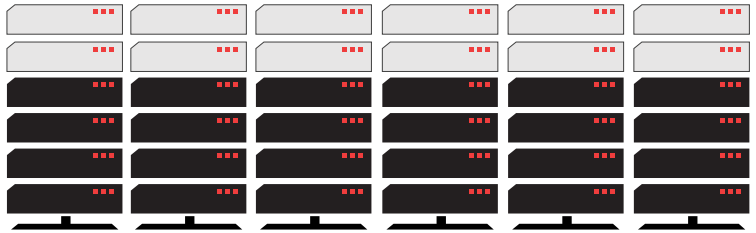
Server Count	Power Usage
81	43.6kW
Up to 15% fewer	35% less

AMD Genoa



Server Count	Power Usage
95	66.9kW

AMD Bergamo



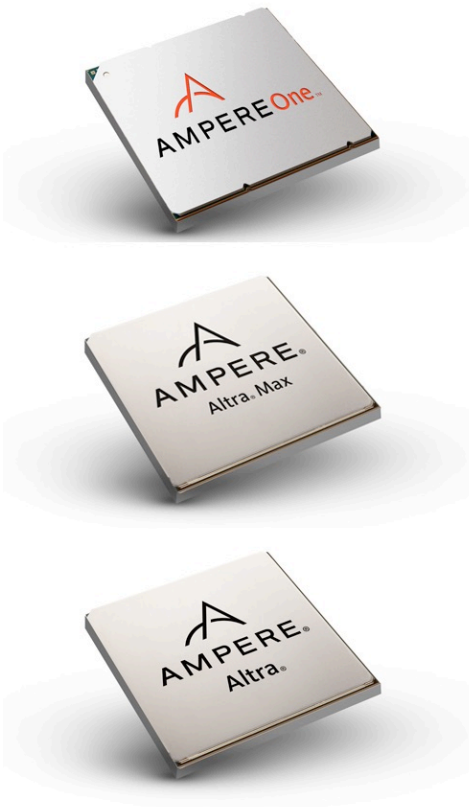
Server Count	Power Usage
90	59.1kW

For full details, See References:

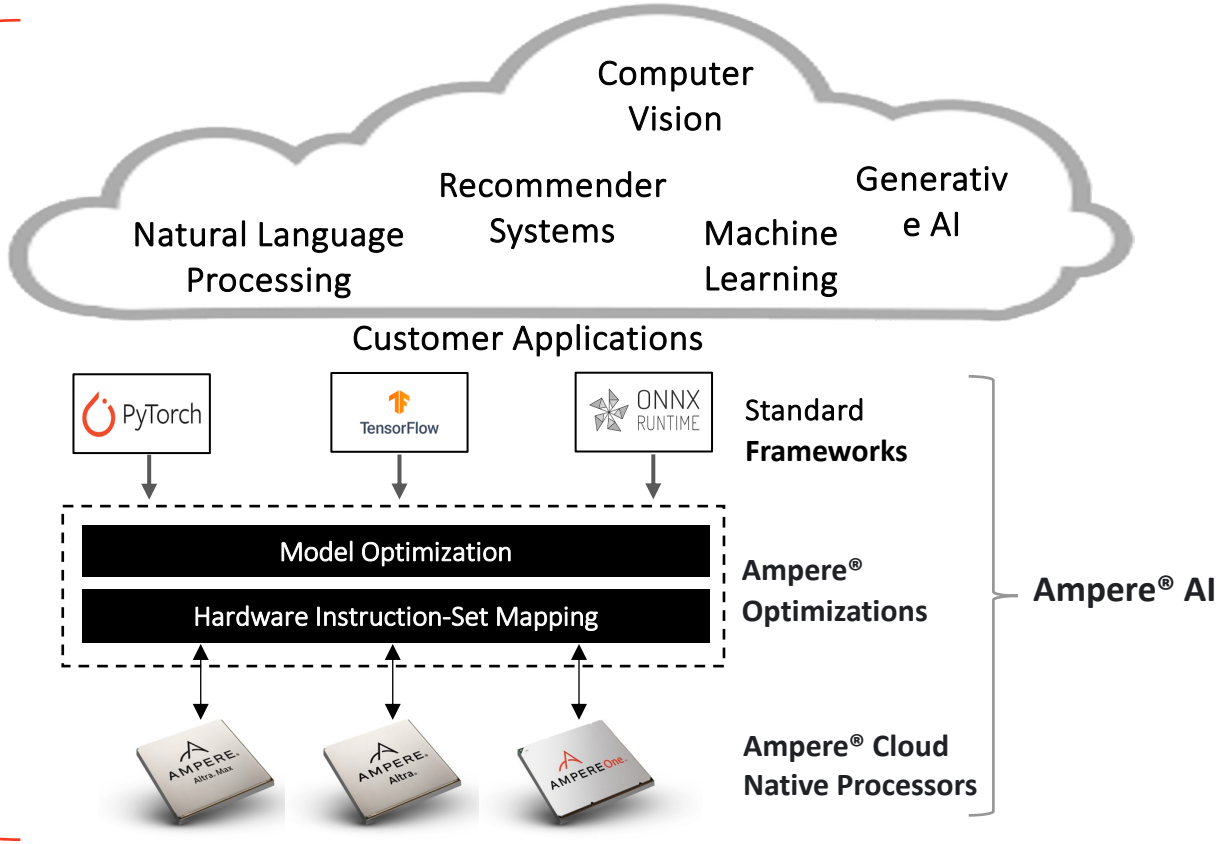
Rack depictions for conceptual illustration purposes only

References: slide 28

The Market's First Cloud Native Processors are Built for AI



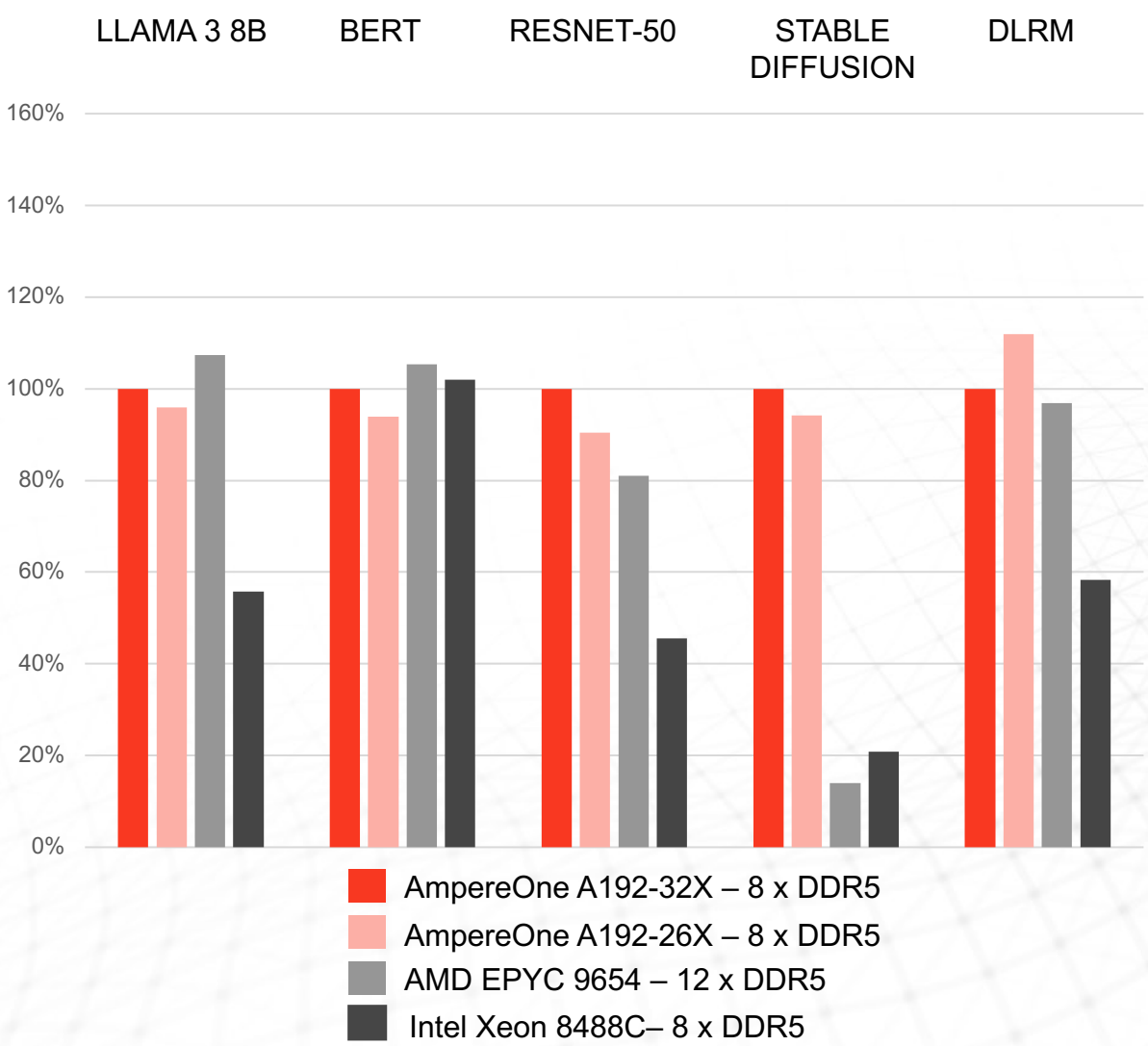
- Web Applications
- Media Encoding
- Databases
- **Artificial Intelligence**
- In-Memory Caching
- Cloud Gaming



Open Frameworks, Acceleration, Disruptive Performance for Cloud Native Workloads Including AI

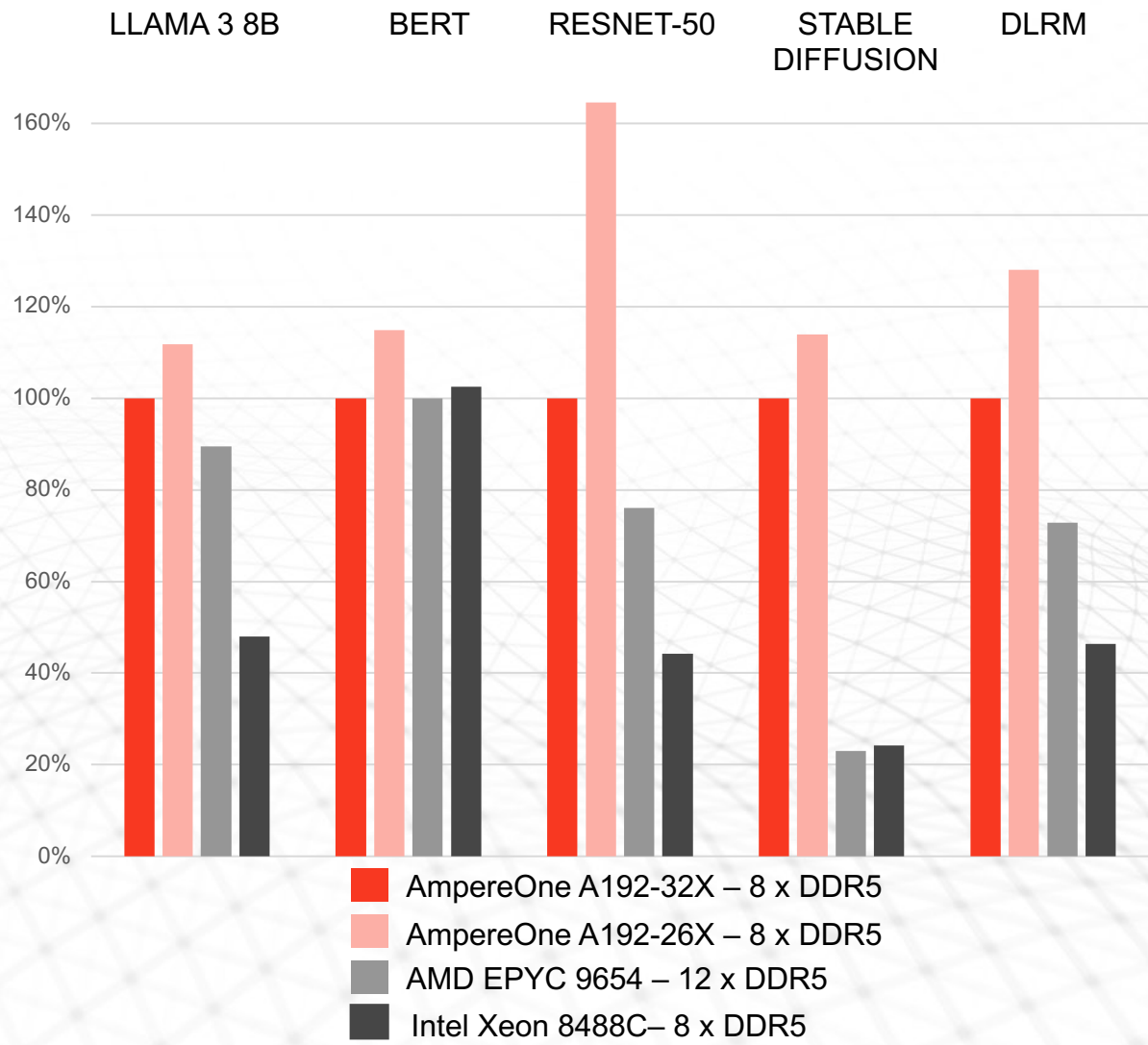
Ampere® AI: GPU-Free Inference Leadership (DDR5 platforms)

Inference Performance



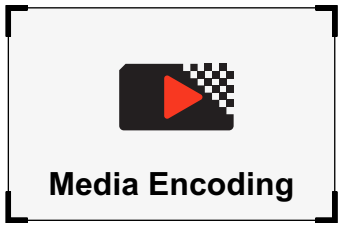
See References

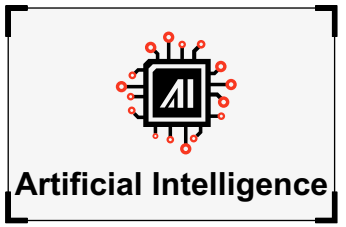
Inference/Watt

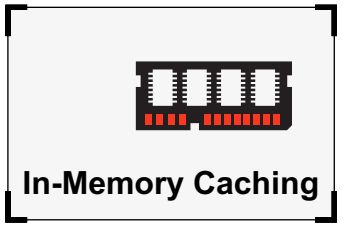


References on slide 29

The Ecosystem is Ready







Applications

Drupal
hadoop
NGINX
Solr
Spark
WordPress

Database

cassandra
MariaDB
mongoDB.
MySQL
PostgreSQL
redis

Infra Tools

ANSIBLE
CHEF
Grafana
kafka
PyTorch
TensorFlow

Networking & Storage

ceph
DPDK
MINIO
OpenSSL
OvS Open vSwitch
SPDK

Language & Runtimes

GO
GCC
node
OpenJDK
python
R

Orchestration, Virtualization & Containers

GraalVM
kubernetes
KVM
MESOS
openstack.

Operating Systems

ORACLE LINUX
SUSE
Red Hat
Alma Linux
Ubuntu
FreeBSD
debian

AMPERE®

Sustainable Computing for AI & Cloud Native Workloads

Matthew Erler - Architect

Aug 27, 2024

References

- Memory Tagging: https://amperecomputing.com/blogs/path-towards-secure-cloud-mte
- QoS Enforcement: https://amperecomputing.com/blogs/quest-for-qos

Performance/Rack AmpereOne End Notes

Performance per Rack: Rack is based on 42U rack with 12.5kW power budget. 2U and 1.0kW allocated as buffer for networking, management and PDU. Total performance per rack calculated by multiplying the performance per server with the maximum number of servers that fit in a rack (until space or power constraints are reached).

HW Configurations

AmpereOne:

- HW: 1 x AmpereOne A192-32X (192c/192t, 3.2GHz) or AmpereOne A192-26X (192c/192t, 2.6GHz), 8 x 64 GiB DDR5 5200 MHz
- OS: Fedora 38
- Kernel: 6.4.13-200.fc38.aarch64

AMD Genoa:

- HW: 1 x AMD EPYC 9654 (96c/192t, 2.4/3.55GHz), 12 x 64 GiB DDR5 4800 MHz
- OS: Fedora 38
- Kernel: 6.4.13-200.fc38.x86_64

AMD Bergamo

- HW: 1 x AMD EPYC 9754 (128c/256t, 2.25/3.1GHz), 12 x 64 GiB DDR5 4800 MHz
- OS: Fedora 38
- Kernel: 6.4.13-200.fc38.x86_64

Server Usage Power: All CPU power draw figures are based on Ampere-performed lab tests under load (for each referenced application). In order to calculate server usage power, platform power draw is added on top of CPU power draw. Platform power assumptions informed by three leading OEM server power calculator tools. Where no measured CPU power draw figures were available, manufacturer-published CPU TDP was used instead.

SPEC CPU®2017 Integer Rate Results: All SPECrate®2017_int_base performance estimates for AMD and Ampere platforms are based on GCC (10/13 compiler). See details in below table. Rack level estimates based on 1U server height and 1 socket platforms.

Processors Under Test	SPECrate®2017_int_ base score (estimated)	CPU Usage Power (W)	Performance / Watt (calculated)	Compiler
AmpereOne A192-26X	616	212	2.91	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AmpereOne A192-32X	694	274	2.53	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AMD Genoa 9654	638	379	1.68	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AMD Bergamo 9754	733	333	2.20	Community GCC 13.2 https://gcc.gnu.org/gcc-13/

Platform Power Assumptions		
Component	**Description**	**Total Power Draw**
Storage	4 x NVMe (10W ea)	40W
Networking	1 x 1GbE OCP NIC, 1 x 10/25GbE NIC, 1 x 100GbE NIC	40W
Other	Motherboard, Fans, Misc	96W
Memory	8 ch DDR4	56W
	8 ch DDR5	80W
	12 ch DDR5	120W

SPEC CPU®2017 Integer Rate Result Rack Level Calculations	AmpereOne A192-32X			AMD EPYC 9654 (Genoa)			AMD EPYC 9754 (Bergamo)		
	# Servers	Power Draw	Rack Performance	# Servers	Power Draw	Rack Performance	# Servers	Power Draw	Rack Performance
SPECrate®2017_int_base	21	11,133W	14,574	17	11,478W	10,846	18	11,325W	13,194

Performance/Rack Cloud Native Workloads: End Notes

Containerized Web Service: Rack claims based on equal weight per application: 1 full rack of AmpereOne 1U form factor servers. AmpereOne performance per rack calculated by multiplying the performance per single server with the maximum number of servers that fit in 1 full rack (until space or power constraints are reached).

Processor Under Test	NGINX Performance	Redis Performance	Memcached Performance	MySQL Performance	Compiler
AmpereOne A192-26X	140,271	156,287,956	81,771,546	337,287	Community GCC 13.2 https://gcc.gnu.org/gcc-13/
AmpereOne A192-32X	174,344	178,597,186	105,806,379	408,141	
AMD EPYC 9654	136,298	143,131,802	91,770,575	320,436	
AMD EPYC 9754	161,693	175,553,748	103,372,044	357,586	

AmpereOne application rack performance used as baseline to calculate the # of AMD Genoa and AMD Bergamo systems and power required to match AmpereOne rack performance. Total power draw (all applications combined) used to calculate the total required rack count for AMD Geno and AMD Bergamo. https://www.amd.com/en/products/specifications/server-processor.htm

CPU Usage Power (W) Web Services Applications	NGINX	Redis	Memcached	MySQL
AmpereOne A192-26X	261	245	211	206
AmpereOne A192-32X	378	300	264	280
AMD EPYC 9654	410	410	410	401
AMD EPYC 9754	410	383	363	374

Web Service Composite		AmpereOne A192-32X				AMD EPYC 9654		AMD EPYC 9754	
Application	Weight	# Racks	Max # Servers / Rack	Total Power Required	Rack Performance	# Systems Required	Total Power Required (W)	# Systems Required	Total Power Required (W)
NGINX	25%	1	18	11,415	3,138,192	24	16,948	20	14,123
Redis	25%	1	20	11,123	3,571,943,720	25	17,654	21	14,262
Memcached	25%	1	22	11,444	2,327,740,338	26	18,360	23	15,161
MySQL	25%	1	21	11,259	8,570,961	27	18,823	24	16,084
			81 total	**45,241W total**		**102 total**	**71,785W total**	**88 total**	**59,630W total**

All server-level performance and power draw claims are based on Ampere Computing LLC internal lab testing.
Server cost estimates include the same estimated costs for all Intel, AMD and Ampere platforms for the following components:

- $2,300 per server for a 1U chassis
- $400 for internal storage
- $350 for 1 DIMM of 64GB DDR5 memory

AMD EPYC server cost estimates include processor cost based on published 1KU pricing as of July 29, 2024: https://www.amd.com/en/products/specifications/server-processor.htm

- $10,625 for AMD EPYC 9654P
- $11,900 for AMD EPYC 9754

AmpereOne server cost estimates include processor cost of $5,555 SBV price (valid as of July 29, 2024).

AI Benchmarks - End Notes

System HW and SW Configurations

AmpereOne:
- HW: 1 x AmpereOne A192-32 (192c/192t, 3.2GHz), 8 x 64 GiB DDR5 5200 MHz
- OS: Fedora 38
- Kernel: 6.4.13-200.fc38.aarch64
- Framework: amperecomputingai/pytorch:1.10.0 (docker image)

AMD Genoa:
- HW: 1 x AMD EPYC 9654 (96c/192t, 2.4/3.55GHz), 12 x 64 GiB DDR5 4800 MHz
- OS: Fedora 38
- Kernel: 6.4.13-200.fc38.x86_64
- Framework: docker image + torch==2.3.1 from PyPI (python 3.10)
- hyperthreading disabled

Intel Sapphire Rapids
- HW: 1 x Intel Xeon 8488C (48c/96t, 2.4/3.2GHz), 384 GiB DDR5 4800 MHz on AWS m7i.metal-24xl
- OS: Ubuntu 22.04
- Kernel: 6.8.0-1009-aws
- Framework: intel/intel-optimized-pytorch:2.3.0-pip-base
- hyperthreading disabled

Processor Under Test	LLAMA3 8B		BERT (Large MLperf squad)		RESNET-50 (v1.5)		Stable Diffusion		DLRM (Torchbench)	
	Performance (tps)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)	Performance (ips)	CPU Usage Power (W)
AmpereOne A192-32X	219	331	21,657	389	1,929	376	6.13	381	1,339,680	306
AmpereOne A192-26X	210	284	20,345	318	1,744	207	5.77	315	1,499,630	267
AMD Genoa 9654	235	397	22,805	410	1,565	401	0.86	232	1,298,523	407
Intel Sapphire Rapids 8488C	122	384	22,097	387	879	388	1.28	329	781,161	385
	Q4, pp128, batch size = 16		Data Precision for AmpereOne SKUs: FP16 Data Precision for AMD and Intel: BF16							

Performance units of measure:
tps = tokens per second
ips = inferences per second (BERT, DLRM)
ips = images per second (Stable Diffusion, RESNET)