AMPERE®

Towards an Equitable AI Cloud: Leveraging Agility and Innovation

March 2-5, 2025 | Mesa, Arizona

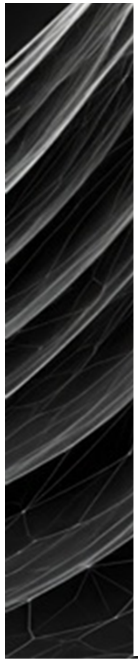
TestConX™

A MODERN SEMICONDUCTOR COMPANY BUILDING

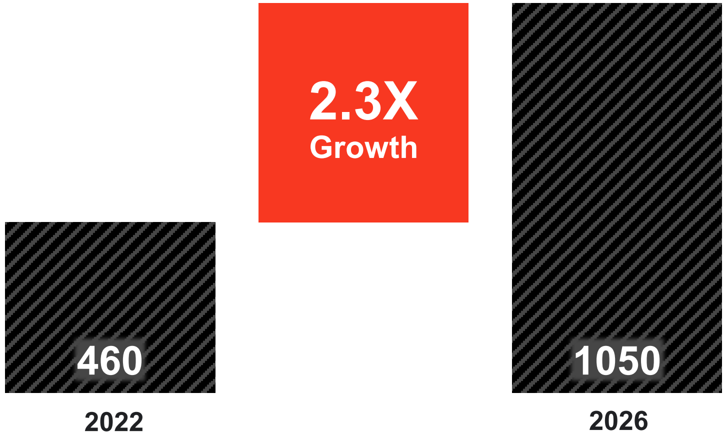
THE FIRST CLOUD NATIVE PROCESSORS

FOR THE SUSTAINABLE CLOUD

Addressing the AI Wave: Responsible Computing



Global Data Center Energy Consumption[1] (TWh)



Capacity Crisis

Strapped data centers face an **AI Growth Cycle** and are collapsing under the weight of:

Aging Legacy Infrastructure

- Low utilization
- 5+ Yr refresh cycles
- Security vulnerabilities

Constrained Power Capacity

- 15kW or less per rack in most DCs
- Average ~10kW / rack
- Regulations & moratoriums

Rising Operating Costs

- Space & power
- Software licenses
- Labor & overhead

Demands of AI Compute

- Inference adoption growing rapidly
- Rising headroom requirements
- GPUs overkill for most inference workloads

[1] Source: International Energy Agency (IEA) "Electricity 2024 – Analysis and forecast to 2024"

THE TIME IS NOW

TO CHANGE THE TRAJECTORY OF COMPUTING

The Third Phase of Compute: Cloud + AI Compute

Fundamentally Different Use Cases Require a Modern Architecture to Increase Performance and Effi

Cloud and AI Compute
(2010s-)

Client Server
(1980s-2010s)

Mainframe Computers
(1960s-1980s)

Software: **Specialized**

Software: **Monolithic**

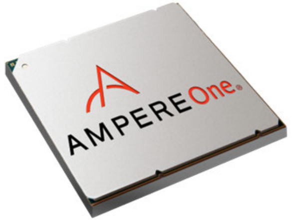
Software: **Microservices—Frameworks**

Cloud Native Processors are the Hardware Required for Cloud and AI Infrastructure and Software

Ampere's Innovative Product Portfolio

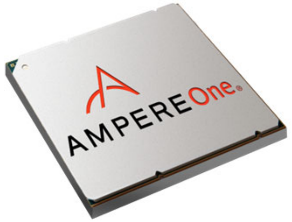
Leadership Performance Per Rack for AI Compute in All Environments

AmpereOne® Family



Up to 192 Cores 5nm
8 Ch DDR5

AmpereOne®



12 Ch DDR5

AmpereOne® M

Ampere® Altra® Family

Up to 80 Cores 7nm
8 Ch DDR4

Up to 128 Cores 7nm
8 Ch DDR4

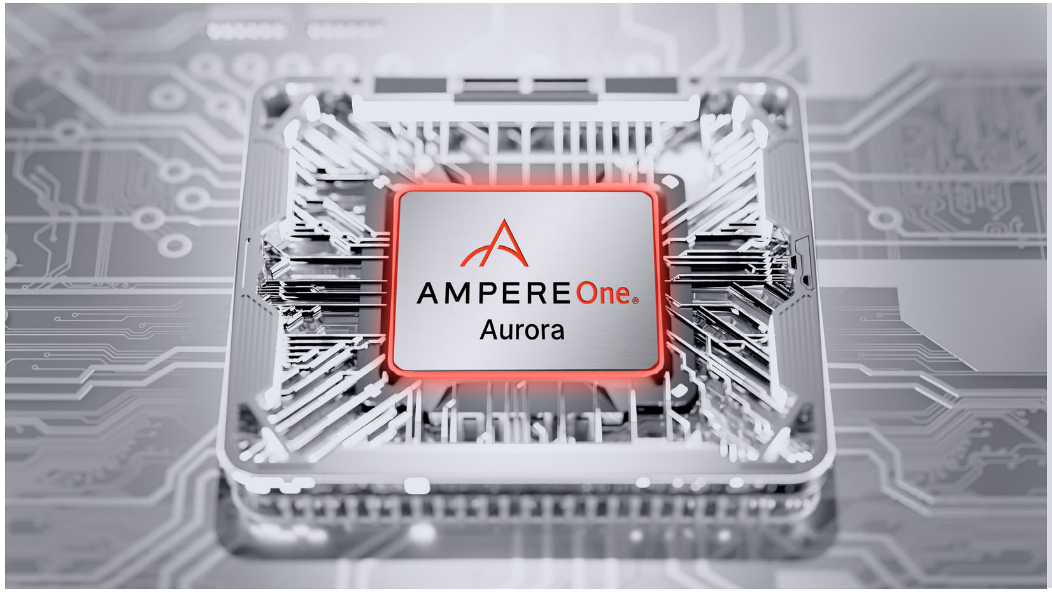
Industry Leading Compute Density and Efficiency Drove New Approaches for Current Products

The Next Advancement in AI Compute: AmpereOne® Aurora

Integrated
AI Acceleration Silicon Directly in the SOC

Air Cooled
Deployable in All Existing Data Centers

Efficient
Accelerate Meeting Our Industry's Climate Goals



- Up to 512 AmpereOne custom cores delivering over 3x the performance per rack of current AmpereOne processors
- AmpereOne Scalable Mesh that allows compute of all types to be seamlessly connected
- Ampere AI IP integrated directly in the SoC via Die-to-Die Interconnect with high bandwidth memory

The Future Demands New Innovation to Fuel Rapidly Growing AI Compute Market

AMPERE®

A Modern Semiconductor Company Building the First Cloud Native Processors for the Sustainable Cloud

Building The Cloud Native Processor

Delivering Performance and Power Efficiency

A Competitive Cloud Native Processor (CNP) & System

High core count – dense compute
Demanding per/core performance
Air cooled power efficient operation
24/7 uptime @ high utilization
5yr+ reliability in operation
Wide range of software applications

Leading edge process node
Multi-chip integration
Complex high layer substrates
Advanced packaging (2.5D, 3D)
High pin count socket
Dense system boards
Constrained thermal/mechanical

Matrix of critical ingredients | Test points | Test vectors

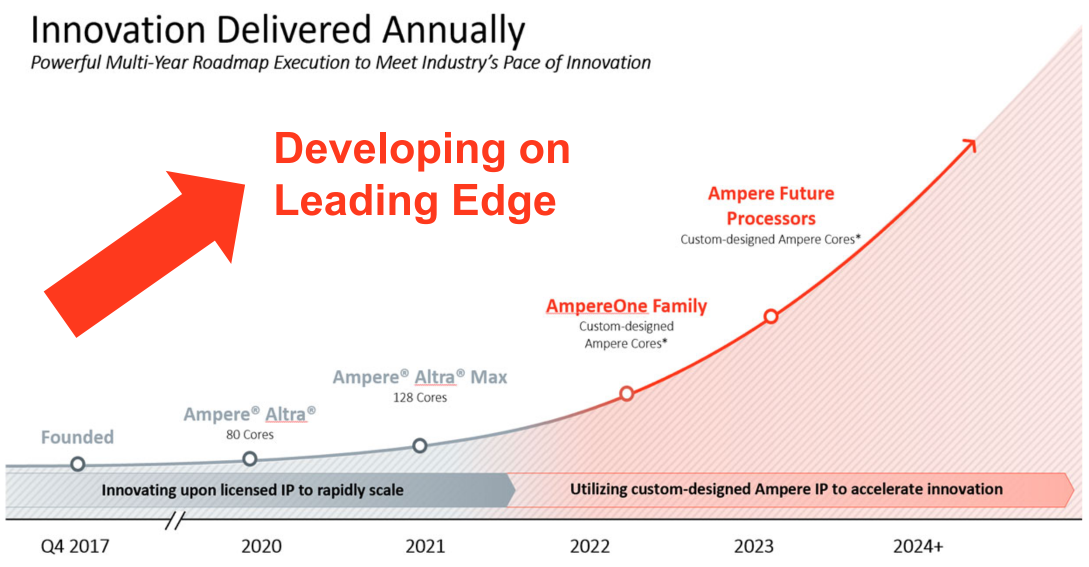
Silicon, substrate | Socket | Platform | Wafer probe | ATE | SLT | Substrate/socket test
Process | Leakage | Temp | Voltage | Pin count | Yield | Reliability | Connectivity

Die-Level Challenges

How Ampere is Addressing Them

Technology Node Related Testing Challenges

- Leading node FET/device variation → Coverage of all process variations (FF,SS,TT) and PVT corners
- High density and chiplets → Complex multi-die testing & selection pre-post assembly
- Higher system utilization in cloud with operation close to TDP → increased device stress, PVT test
- 5+ yr Reliability testing and qualification complexity → higher current densities causing thermal runaway risk under HTOL conditions



PVT: ProcessVoltageTemp
HTOL: HighTempOperatingLife
TDP: ThermalDesignPoint

Designing at the leading edge uncovers new challenges and accelerates existing ones

Design Choices for Enhanced Testing

Complex Test Hardware

- Complex Probe cards and load boards
- Complex system level test
  - Hardware and test flow and content
- Thermal and voltage controls
- Enable fine grain PVT testing
  - Ensuring test at all combinations
  - Requires multiple insertions

Monitor, Detect, Mitigate: On-die Monitors

- Spec sensors and on-die compensation
- Process monitoring, device aging compensation, voltage droop mitigation
- Limits management IP: peak current limiting, thermal sensor and limiting

Design Content for Coverage

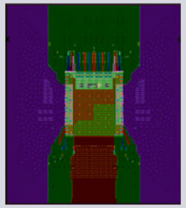
- High scan and test coverage
- High Memory-bist, IO-Bist coverage
- Ensuring high test content at Probe
  - Identifying Known-Good die at Probe

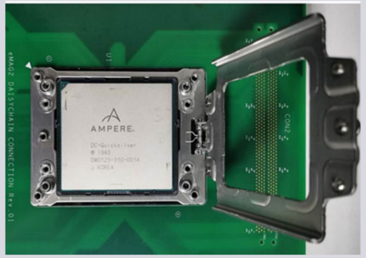
AmpereOne Production Flow

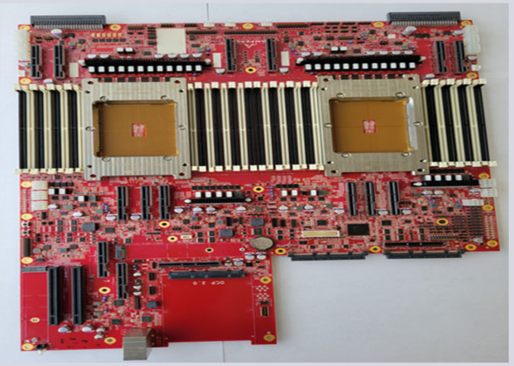
- **Wafer Probe**
- **Good Die Assembly**
- **Multiple FT Insertions**
- **System Level Test**
- **SKU Binning Mark & Fuse**
- **Final QA & Ship**

- **Test Coverage**
  - Stuck-at-fault
  - Transition-fault
  - MBIST
  - BSCAN
  - IO Loopback
  - OS Functional tests
  - Perf benchmarks
- **Binning:**
  - Core count
  - Frequency
  - Leakage

- **Process (SS, FF, TT):**
  - Multiple chiplet combinations
  - Multiple corner wafers and dies harvested using WAT and leakage
- **Voltage:**
  - Positive/negative margins as required by data-sheet and PHYs/IOs
- **Temperature:**
  - Full range of operating temperature







The Rest of the System is as Critical:

Substrate, Socket, Platform Co-design,
Co-test with Silicon-die is Critical

Ampere Server CPUs: **Package/Substrate**

Challenges:

- Chiplet trend for silicon cost reduction
- Large substrate to support increasing die count, Pin count
- Low-power & high-density Die-to-Die interconnects
- Reduced warpage for LGA socketing
- Test & screening at various packaging levels: bare die, RDL module, and package

Testing & Design Considerations :

- Via-stack designs
- Advanced die-die interconnect solutions
- Highly Engineered Materials to reduce the coefficient of thermal expansion
- Low Df/Dk build-up material choices for improved signaling
- Optimized capacitor selection and placement to reduce power supply noise

Ampere® Altra® Platforms
For Modern Compute

Packaging Strategy

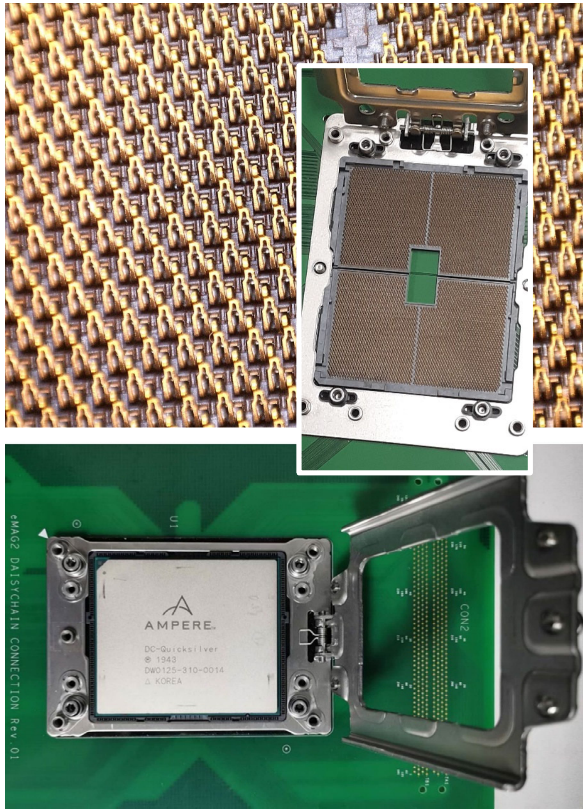
Innovations Needed:

- Design for tests with built-in features such as daisy-chain die-cracking & LGA-contact monitors as well as circuit/bump/contact redundancies
- Cost effective 2.5D and 3D packaging technologies
- Fine-pitch probing card for chip & package screening
- Strengthen Eco-system collaborations to jointly develop enabling packaging and testing technologies
- Continue pushing the limits to address ever increasing signal and power delivery challenges

Ampere Server CPUs: **Socket**

Challenges:

- Large body size and high pin count sockets (>10K contacts), various speed interfaces (DDR, CCIX, PCIe)
- SMT challenges: single vs split body sockets
- Socket warpage management & solderability
- Load application mechanism: load plate vs. heatsink-loading
- Multiple socket types:
  - ATE Socket
  - Validation Socket—enables multiple insertions >5000
  - Production Socket
  - HTOL/Burn-in Socket
  - Metal spring, Elastomer
- Leads to Inconsistent:
  - Loading Approach, Contacts, Test type, Socket Height etc.
  - Current handling, Signal integrity and Socket Cost



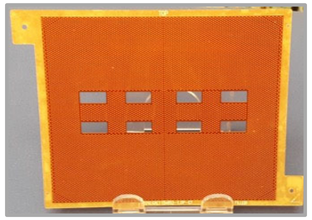
Ampere® Altra® Platforms
For Modern Compute

CCIX: Cache Coherent Interconnect for Accelerators

Socketing Technology:
Validation/Production/FT/SLT/Burn-in Socket

Innovations Needed:

- Standardization of socket technology/contact types (SLT/ATE/FT/HTOL)
- Loading Mechanism -exert right amount pressure on the exposed die and stiffener
- Socket cost reduction for high pin count (>10k contacts)
- Surface mount process for high pin count LGA/BGA and large body size production sockets



Ampere Server CPUs: **System/Platform**

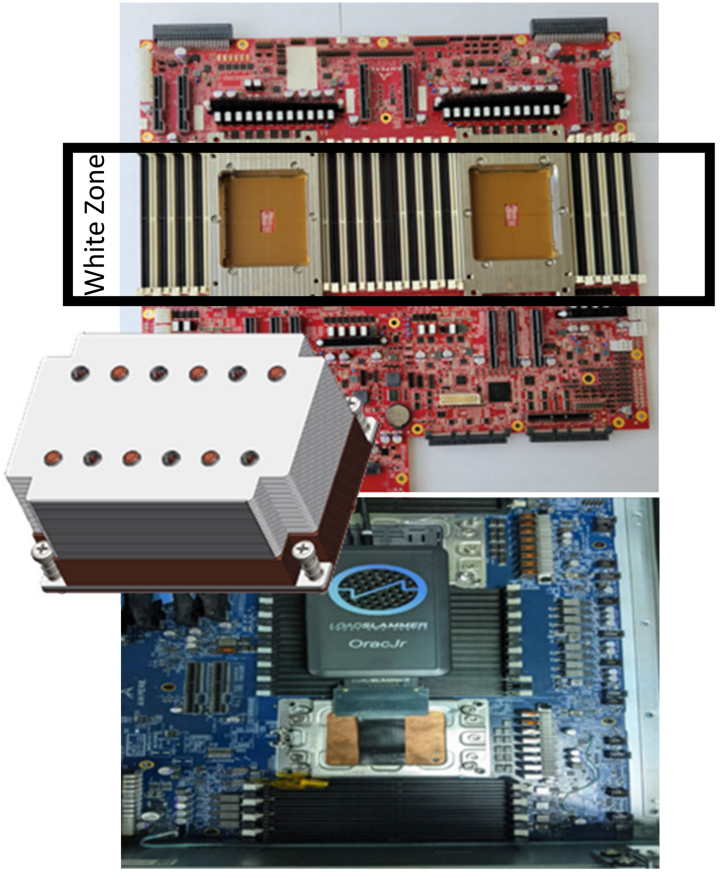
Challenges:

- Complexity in PCB, high I/O count, high-speed interfaces (DDR/PCIe), complex clocking, VR solution, routing for SI/PI
- High TDP operating conditions

Testing & Design Considerations:

- Modular board design for best SI/PI
- VRTT: special tool to validate PI/VR & Platform PDN response
- Improved thermal design and materials (TIM)
- Mechanical shock/vibe for platform stability

Ampere is working on extending air cooling to AI compute



Ampere® Altra® Platforms
For Modern Compute

VR: Voltage Regulator; VRTT: Voltage Regulator Test Tool; PI: Power Integrity;
PDN: Power Delivery Network

Platform Level Validation / Characterization

Challenges

- SI issues
- Limitation of traditional methods
- Leveraging off-the-shelf IP design & tuning features
- Platform dependency
- Limitations of SI/PI simulation methodologies

SI: Signal Integrity; PI: Power integrity

Innovation Needed

- Validation SI tool
- Beyond Industry compliance
- Optimizing feedback loop
- Cost effective and scalable tools
- Eye diagram visualization

Call to Action

- Tools for debugging
- Prioritized solutions
- FA feedback
- Hardware-based, cost-effective tools
- Advanced training and equalization techniques

In Closing

Ecosystem Partnership

Silicon/IP | **Substrate** | **Socket** | **Platform**

Ampere is a 7-year young company that is:

Pushing technology limits in our design choices across all the ingredients

Paying careful attention to ensure robust co-design and co-test across all the ingredients

Continued innovation and partnerships in industry to deliver game changing products

Presentation / Copyright Notice

The presentations in this publication comprise the pre-workshop Proceedings of the 2025 TestConX workshop. They reflect the authors' opinions and are reproduced here as they are planned to be presented at the 2025 TestConX workshop. Updates from this version of the papers may occur in the version that is actually presented at the TestConX workshop. The inclusion of the papers in this publication does not constitute an endorsement by TestConX or the sponsors.

There is NO copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies: as such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author/s or their companies.

The TestConX logo and 'TestConX' are trademarks of TestConX.