AI INFRA SUMMIT

HARDWARE & SYSTEMS INSIGHTS

INCLUDING

arm cadence Google

KOVE LIGHTMATTER Meta

JOHN OVERTON, CEO, KOVE

KEYNOTE
DRIVING THE NEXT 1000X LEAP IN COMPUTING WITH PHOTONIC AI INFRASTRUCTURE

NICK HARRIS, FOUNDER & CEO, **LIGHTMATTER**

01 **Interconnect Bottleneck**

The next 1000X in AI performance will be driven by interconnect, as current scaling of compute units outpaces I/O capabilities, leading to a “shoreline” problem where chips lack sufficient communication bandwidth.

02 **Performance Divergence**

Compute throughput has increased by 10^7X since 2000, while interconnect has only improved by 10^2-10^3X, causing chips to wait for data and limiting AI data center efficiency.

03 **Lightmatter’s Product Offerings**

Lightmatter provides two main product lines: the L-series (3D-stacked CPO I/O tile for high bandwidth density) and the M-series (3D-stacking entire GPU/switch complexes on a photonic substrate).

04 **Breakthrough in Link Architecture**

Lightmatter has achieved a breakthrough by packing 16 colors of light (8 transmit, 8 receive) into a single optical fiber, delivering 800 gigabits per second, matching pluggable transceivers in a fraction of the size and cost.

05 **Energy Efficiency**

This new photonic link architecture reduces energy consumption to 2.6 picajoules per bit (4.6 pJ/bit including SERDES), a 4X improvement over pluggable optics (17-20 pJ/bit), significantly cutting data center power usage.

06 **Faster Model Training**

Using Lightmatter’s photonic scale-up technology, AI model training times can be reduced by up to 3X for models with a high number of experts, offering a significant economic advantage for foundation model development.

KEYNOTE
WHAT AI COULD DO WITH LIMITLESS MEMORY

JOHN OVERTON, CEO, **KOVE**

01 Memory Bottleneck

The industry's traditional view of DRAM as "stuck in a box" creates self-limiting memory architectures, leading to high costs (60% of server cost) and significant memory stranding (50%).

02 The Core Problem

When memory runs out, systems resort to disk, which is 125 to 1,000 times slower than DRAM, severely impacting performance and scalability.

03 Kove's "Unlimited DRAM"

Kove's software allows servers to dynamically request and receive memory from a pooled resource, effectively providing "unlimited DRAM" regardless of hardware generation or location.

04 Dramatic Performance Improvements

Kove enables 60x faster model training, 3-5x faster vLLM inference, 100x container density, and 12-54% power reduction.

05 Benchmark Validation

Benchmarks against Redis and Valkey show Kove's remote memory (150m away, 5x larger) performing 11-42% faster than local DMA for Redis and 1-10% faster for Valkey.

KEYNOTE
DESIGN FOR AI AND AI FOR DESIGN

CHARLES ALPERT, AI FELLOW, **CADENCE**

01 **AI Compute Growth**
The rapid growth of AI compute power (2X per year) is causing a power crunch, leading to massive investments in data centers, estimated at $5-8 trillion.

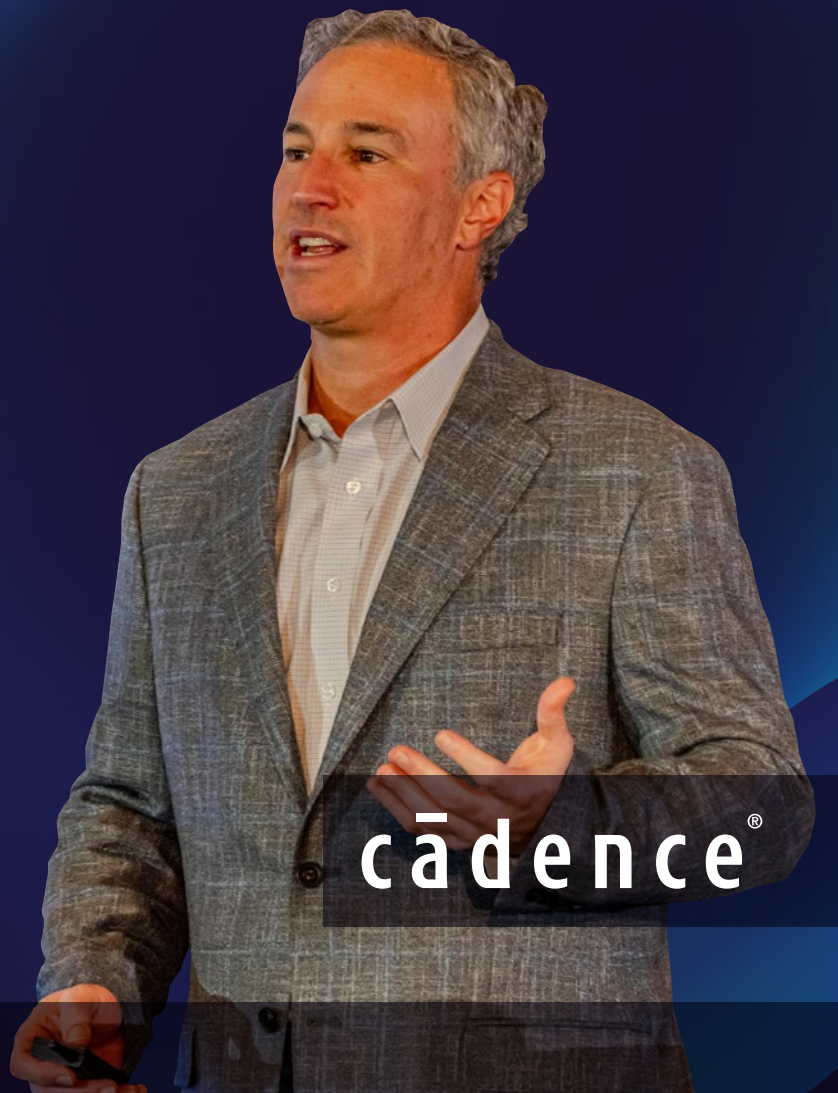
02 **AI in Chip Design**
Cadence software is integral to designing AI chips, and AI is increasingly embedded within Cadence's own systems, with 50% of chips today using AI, projected to reach 90% in three years.

03 **Agentic AI Levels**
Cadence is developing AI from optimization AI (e.g., Cerebrus) and conversational LLMs to complex reasoning, agentic workflows, and ultimately full autonomy, akin to self-driving car levels.

04 **Verification Workflow Automation**
Agentic AI is used to automate complex verification workflows, from specification analysis and verification plan generation to RTL generation and iterative debug/optimization.

05 **Hardware Acceleration for Simulation**
Cadence leverages GPUs and custom hardware (Palladium, Millennium) to significantly accelerate complex simulations for 3D-ICs (e.g., 10-18X speedup) and other silicon design problems.

06 **Data Center Digital Twin**
Cadence's Reality software provides a unique digital twin for data centers, enabling live simulation, capacity increase, and energy reduction, supported by an ecosystem of 14,000 models



KEYNOTE
ARCHITECTING INTELLIGENCE FOR THE NEXT DECADE ON ARM

MOHAMED AWAD, SVP & GM, INFRASTRUCTURE BUSINESS, **ARM**

01 **Pervasive Platform**
Arm is the world's most pervasive compute platform, with over 300 billion chips shipped, supported by a massive software ecosystem of 22 million developers, creating a virtuous cycle of hardware and software growth.

02 **Neoverse Launch**
In 2019, Arm launched Neoverse, shifting to building specific CPUs, interconnects, and system IP for the infrastructure market, significantly lowering investment and accelerating time to market for partners.

03 **Neoverse Success**
Neoverse has shipped over 1 billion CPU cores, bringing significant efficiency to data centers, exemplified by AWS Graviton accounting for 50% of their deployed compute in the last two years.

04 **Ecosystem Adoption**
Neoverse has achieved massive adoption, with major cloud providers like AWS, Google, and Microsoft optimizing software for Arm, predicting 50% of top cloud service provider compute will be Arm-based by 2025.

05 **Arm's Role in Custom AI**
Arm Neoverse CPUs are central to custom AI systems, managing frameworks, schedulers, and operating systems, with partners like Nvidia building custom silicon and racks around Arm technology.

PANEL
CUSTOM VS GENERAL COMPUTE: THE NEXT WAVE OF COMPUTE PRODUCTION

01 Custom Silicon Strategy

Mahesh (Meta) explains their vertical integration to tune custom silicon for specific AI workloads like recommender systems, while Leo (Google) highlights their 10+ years of TPU development, offering it as a cloud service and informing infrastructure.

02 Adiabatic Computing for Efficiency

Andrew (Vaire Computing) introduces adiabatic reversible computing as a solution to fundamental energy and heat problems in data centers, aiming for hundreds to thousands of percent efficiency gains at the circuit level.

03 Power as Key Optimization

Mahesh (Meta) and Leo (Google) agree that power efficiency is the critical system parameter, with Meta focusing on optimizing standard cell libraries and Google leveraging chip design, software (e.g., Mixture of Experts), and liquid cooling.

04 Startup Challenges & Hyperscaler Engagement

Andrew (Vaire Computing) and Ingolf (Euclyd) discuss the challenges for startups, noting hyperscalers only seek extreme technologies when facing fundamental limits, and the broader data center market offers opportunities.

05 Advice for Startups

Leo (Google) advises startups to offer "eye-popping" innovation while minimizing software change friction, while Mahesh (Meta) stresses the need for operational efficiency, reliability, and software stack compatibility for vendor success.

Andrew Sloss
VP Technology
Vaire Computing

Mahesh Maddury
AI Architect
Meta

Leo Leung
Director, Product Management,
AI & Computing Infrastructure
Google Cloud

Ingolf Held
VP, Product
Euclyd

MODERATOR:
Brett Simpson
Co-Founder & Senior Analyst
Arete Research

EXCLUSIVE INTERVIEWS

FROM OUR HARDWARE & SYSTEMS LEADERS

REGISTER YOUR INTEREST IN ATTENDING IN 2026