AI INFRA SUMMIT

September 9-11, 2025 | Santa Clara Convention Center

2025 REPORT

DIAMOND PARTNERS

aws cadence Google Microsoft

IAN BUCK, VP OF HYPERSCALE AND HIGH-PERFORMANCE COMPUTING, NVIDIA

HIGHLIGHTS

1. NVIDIA unveiled Rubin CPX, a new GPU for long-context AI reasoning which enables up to 8 exaflops of NVFP4 compute.

2. The age of inference is here. Google revealed its internal inference has exploded: from ~9.7 trillion tokens per month in April 2024 to over 1,400 trillion tokens per month.

3. Kove, SK Hynix, Pliops and Cadence showcased significant breakthroughs in the fields of memory, storage and design innovation for large scale AI inferencing systems.

4. “AI is kicking our butts and teaching us that we know nothing about infrastructure…” said Meta’s VP of Engineering, YJ Song, as he shared their massive datacenter projects including the Prometheus Supercluster.

5. Lightmatter’s CEO demonstrated unprecedented 800 Gbps bidirectional bandwidth from photonic systems that could spur the next 1000X in AI computing.

JOHN OVERTON, CEO, KOVE

4,300 REGISTERED

250 SPEAKERS

150 PARTNERS

BARRY COOKS, VP, COMPUTE ABSTRACTIONS, AWS

ANNOUNCEMENTS AT AI INFRA SUMMIT 2025

EXCLUSIVE NVIDIA ANNOUNCEMENT

- Rubin CPX GPU: A New Class of GPU Designed for Massive-Context Inference
- Giga-Scale AI Factory Reference Designs

OTHER NEWS

- MLPerf Inference v5.1 Benchmark Results announced on-stage
- Recogni unveils new brand, and is reborn as Tensordyne
- Kove: SDM benchmark results show 5X larger AI inference workloads on Redis and Valley
- UALink, RISC-V, CXL Consortium and others showcase products on expo floor

CHARLES ALPERT, AI FELLOW, CADENCE

TOP SPEAKERS

IAN BUCK
VP of Hyperscale and High-Performance Computing
NVIDIA

BERTA RODRIGUEZ-HERVAS
Chief AI and Analytics Officer
Pfizer

MARK LOHMEYER
VP & GM, Compute and AI Infrastructure
Google

BARRY COOKS
VP, Compute Abstractions
AWS

YEE JIUN SONG
VP, Engineering
Meta

LAURA ORTMAN
CEO
Cologix

DAVID GLICK
SVP, Enterprise Business Services
Walmart

THANK YOU TO OUR 2025 PARTNERS

DIAMOND PARTNERS

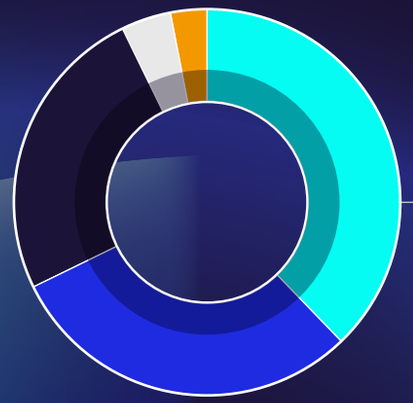
aws cadence Google Microsoft

PLATINUM PARTNERS

arm BROADCOM EDGE IMPULSE IREN KOVE Lambda

NEBIUS NVIDIA Qualcomm SIEMENS SYNOPSYS TENSORDYNE

AUDIENCE BREAKDOWN



- Tech Ecosystem.....**38%**
  Semiconductor, Systems, Systems and Device Manufacturers, CSPs/ Hyperscalers, AI Software Infrastructure, AI Application Software, Physical Infrastructure
- Enterprise AI Practitioners.....**30%**
- Hyperscaler.....**25%**
- Investors.....**4%**
- Press.....**3%**

80%

plan to **purchase/upgrade their GPUs in the next 12-18 months.** A further **54% expect to do this with memory, and 36% on interconnect**

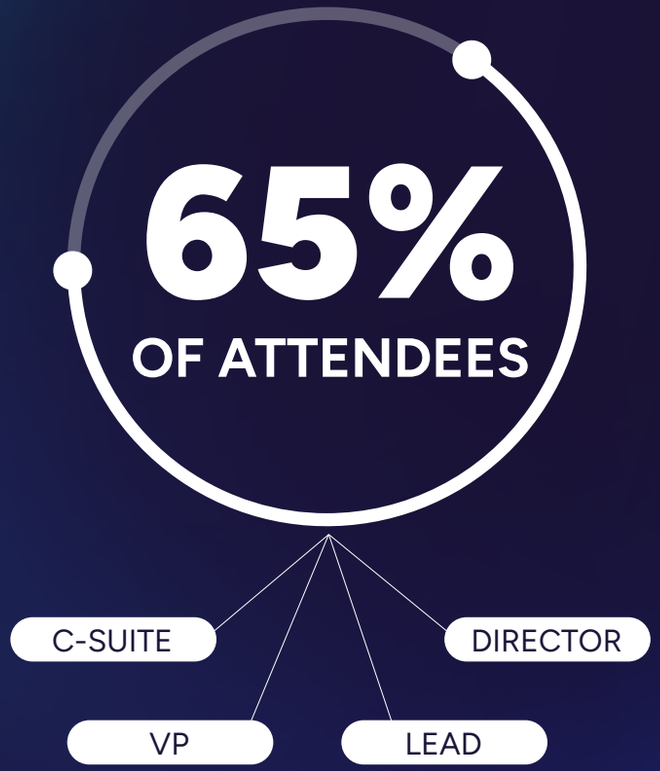
72%

annually purchase/upgrade to new hardware components (e.g memory, chips, interconnect)

YEE JIUN SONG, VP, ENGINEERING, **META**

AI TECHNOLOGISTS GATHERING FOR THE KEYNOTES AT THE AI INFRA SUMMIT

ATTENDING COMPANIES INCLUDED...



BERTA RODRIGUEZ-HERVAS, CHIEF AI AND ANALYTICS OFFICER, PFIZER

SEAN LIE, CO-FOUNDER & CTO, CEREBRAS

ALAN ROSA, CISO & SVP, TECHNOLOGY, CVS HEALTH

EXHBITION HALL - 1,470 EXPO-ONLY ATTENDEES

PRIYA KESAVAN, PRINCIPAL, AZURE AI SYSTEMS STRATEGY, MICROSOFT

AXELERA AI DEMO STAGE PRESENTATION

5,300+ VIRTUAL BUSINESS CARDS WERE EXCHANGED AT THE EXPO HALL

AI ENTHUSIAST PLAYING WITH HUMAN ROBOT

RAJOSHI GHOSH, CO-FOUNDER & CHIEF ECOSYSTEM OFFICER, PROMPTQL

DAY ONE OPENING GALA

PRESS COVERAGE

> Read the articles

GLOBAL MEDIA INCLUDED:

Bloomberg

CNBC

The Economist

DATA CENTER Frontier

Forbes

TechCrunch

semianalysis

SemiWiki
The open forum for semiconductor professionals

siliconANGLE

Silicon Valley Insider

StorageReview

TechArena™

The Information

The New York Times

THENEXTPLATFORM

MEDIA PARTNER OR PRESS PASS
FOR AI INFRA SUMMIT 2026?

> Contact Priya Khosla, Marketing Director, AI & Tech

LOOKING TO 2026

After a watershed year that witnessed a salvo of product launches, including NVIDIA's Rubin CPX, we'll return next year and dissect another year's worth of technology, defining the next 12 months in infrastructure development.

Prestigious keynotes, blockbuster announcements, and the most comprehensive industry representation anywhere in AI infrastructure – that's what makes AI Infra Summit so impactful. 8000 attendees, 400 speakers and 250+ partners will fill the halls of Santa Clara Convention Center in September 2026.

As enterprises enter the era of inference, and AI moves from servers into physical and embodied systems, we will bring you the keynotes, product launches, and technical prowess that you need to stay competitive.

Engage with us now to get your work onto AI infrastructure's ultimate stage.

NETWORKING AT THE AI INFRA SUMMIT

NVIDIA KEYNOTE ON DAY ONE


VR HEADSET AT THE AI INFRA SUMMIT


EXPO HALL AT THE AI INFRA SUMMIT

AI INFRA SUMMIT OPENING VIDEO

ANDREW NG, EXECUTIVE CHAIRMAN, LANDINGAI

DAY ONE HIGHLIGHTS VIDEO

ANKUR GUPTA, SVP & GM, SIEMENS EDA

DAY TWO HIGHLIGHTS VIDEO

MARK LOHMEYER, VP & GM, COMPUTE AND AI INFRASTRUCTURE, GOOGLE

AI INFRA SUMMIT

September 15-17, 2026
Santa Clara Convention Center

"Exactly the people I want to meet, who I want to talk to and to share our vision, our strategy, our future."

CHIEF INNOVATION OFFICER, **ALIGNED DATA CENTERS**

POWERING FAST, EFFICIENT & AFFORDABLE AI

- HARDWARE & SYSTEMS
- PHYSICAL AI
- ENTERPRISE AI
- AI DATA CENTER
- 2x DEMO STAGES

8,000 REGISTERED
400 SPEAKERS
250+ PARTNERS

2026 TRAILER VIDEO

SPONSORSHIP OPPORTUNITIES? CONTACT BEN EDWARDS
bedwards@kisacoresearch.com

> REGISTER YOUR INTEREST IN 2026