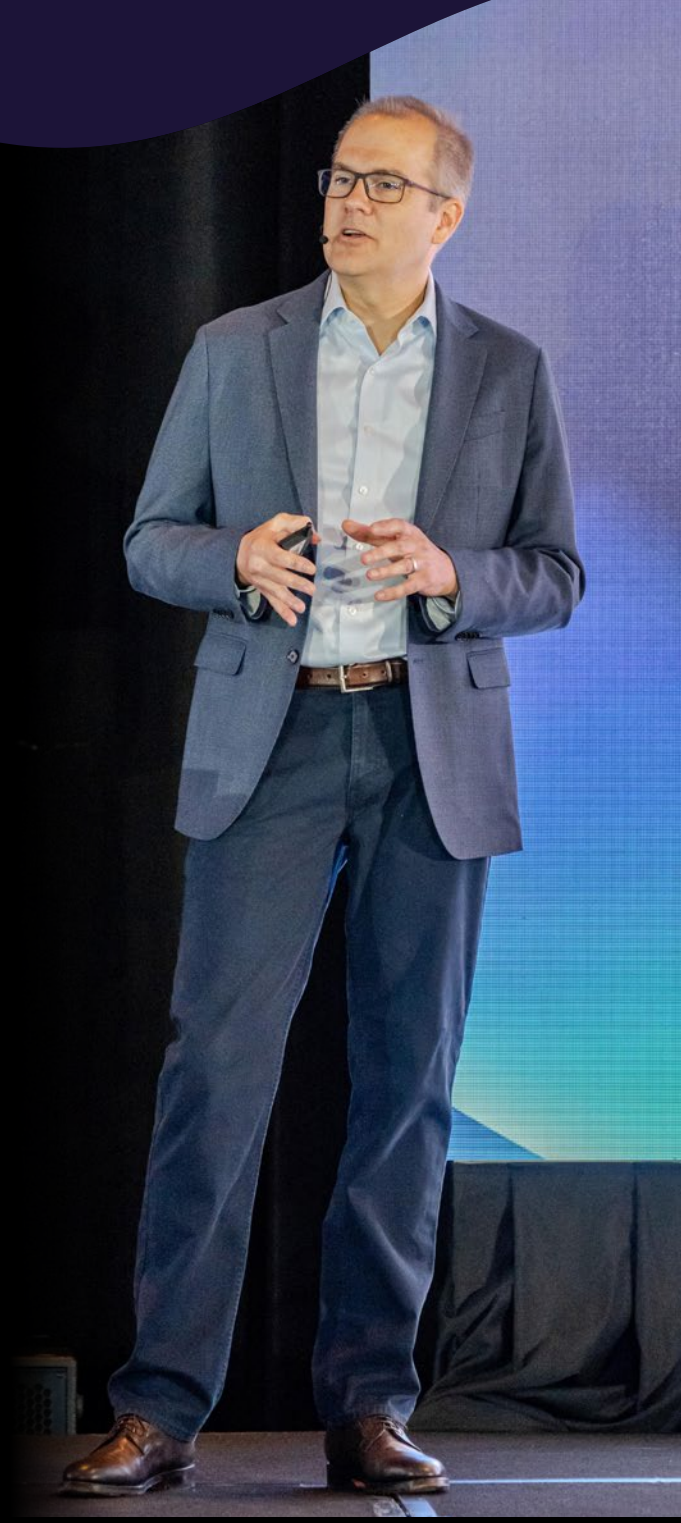
AI INFRA SUMMIT

HYPERSCALER INSIGHTS

PARTNERS

aws Google Meta NVIDIA



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

KEYNOTE
ADVANCING INNOVATION IN AI INFRASTRUCTURE

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

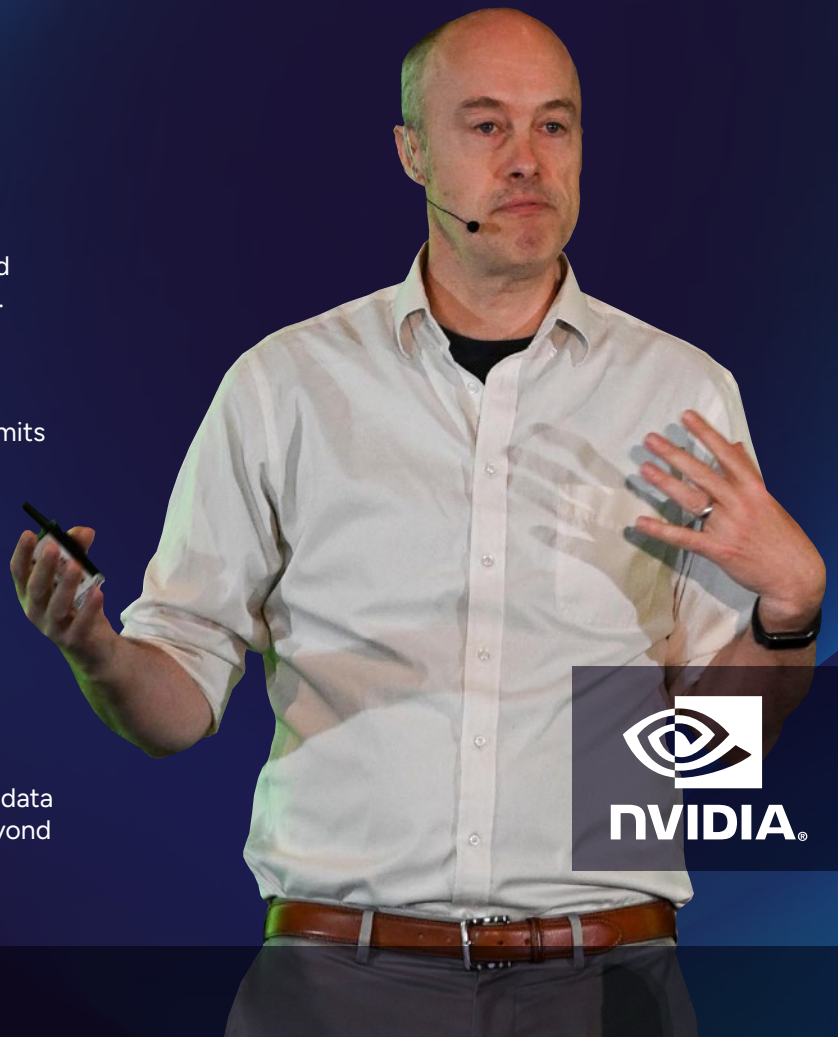
01 **AI Infrastructure Challenges**
Complex trade-offs across intelligence, responsiveness, cost, throughput, and energy efficiency, requires long-term planning for hardware and silicon development.

02 **Technological Innovations**
NVLink X for rack-scale disaggregation, Spectrum-X Ethernet for massive GPU interconnects, and advanced numerical methods like NVFP4 for efficient computation in both inference and training.

03 **Long Context Processing**
Advanced coding and video generation require processing millions of input tokens, pushing the limits of current models and creating a need for specialized hardware.

04 **Rubin CPX Processor**
The Rubin CPX GPU, a new processor specifically optimized for massive context length processing (1,000,000+ tokens) with triple attention acceleration cores and enhanced video encoders/decoders, launching in late 2026.

05 **AI Factory Gigascale Reference Design**
Collaborating with the community and industry partners to develop a reference design for future data centers, addressing power, cooling, electrical, and mechanical challenges to scale AI factories beyond individual racks.



KEYNOTE
ADVANCING META'S AI INFRA

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

01 **Infrastructure Evolution**
From simple web servers in leased data centers to a global network of custom-built data centers, terrestrial backbones, and subsea cables, learning to manage distributed systems, failures, and complexity.

02 **GPU Cluster Challenges**
Running GPU clusters differed significantly from CPU fleets, resembling HPC supercomputers with high-bandwidth, low-latency networks and specialized memory systems, requiring dedicated "AI zones" within data centers.

03 **Mega-Cluster Development**
Meta built 24K H100 GPU clusters within single data center buildings and, unprecedentedly, emptied five production data centers to create a 129K H100 cluster, involving massive logistical and engineering challenges.

04 **Future Scale**
Planning even larger clusters, including the "Prometheus" (1 gigawatt, by 2026) and "Hyperion" (5 gigawatts), demonstrating a mind-boggling scale of infrastructure build-out.

05 **Hardware Diversity**
The AI workload's rapid evolution has led to a proliferation of hardware solutions, including NVIDIA's GB200/300, AMD's MI300X, and Meta's own custom silicon optimized for specific workloads like ranking and recommendation.

KEYNOTE
WHAT'S NEXT FOR THE FOUNDATIONS OF AI

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

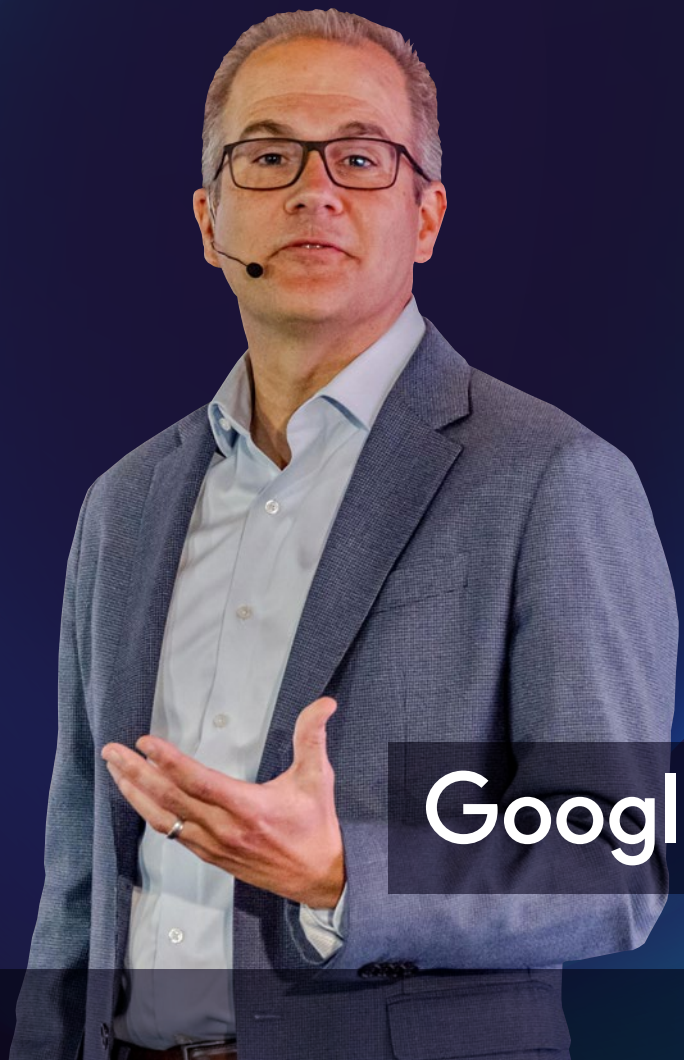
01 **AI Growth & Inference Focus**
Google observed a 50x increase in tokens processed in 12 months, reaching 1,400 trillion tokens per month.

02 **Power Efficiency Metrics**
A comprehensive methodology for measuring power consumption, finding that a single Gemini prompt uses approximately 0.25 watt-hour, 30mg of CO2, and five drops of water, comparable to watching nine seconds of TV.

03 **Full-Stack Optimization**
Through model architectures, software, custom hardware, data center operations, Google reduced the energy consumption per Gemini prompt by 33 times in one year.

04 **Accelerator Choices**
NVIDIA GPU-based services (G4, A4, A4X on Blackwell platforms) and its own industry-leading TPU platform, now in its Ironwood generation, specifically designed for inference.

05 **TPU SuperPod Innovations**
Scaling to 9,216 chips in a super pod with 1.77 petabytes of HBM, featuring optical circuit switching (OCS) for resilience and fifth-generation liquid cooling.



KEYNOTE
PROVIDING THE INFRASTRUCTURE OF TOMORROW, TODAY

BARRY COOKS, VP, COMPUTE SERVICES, **AWS**

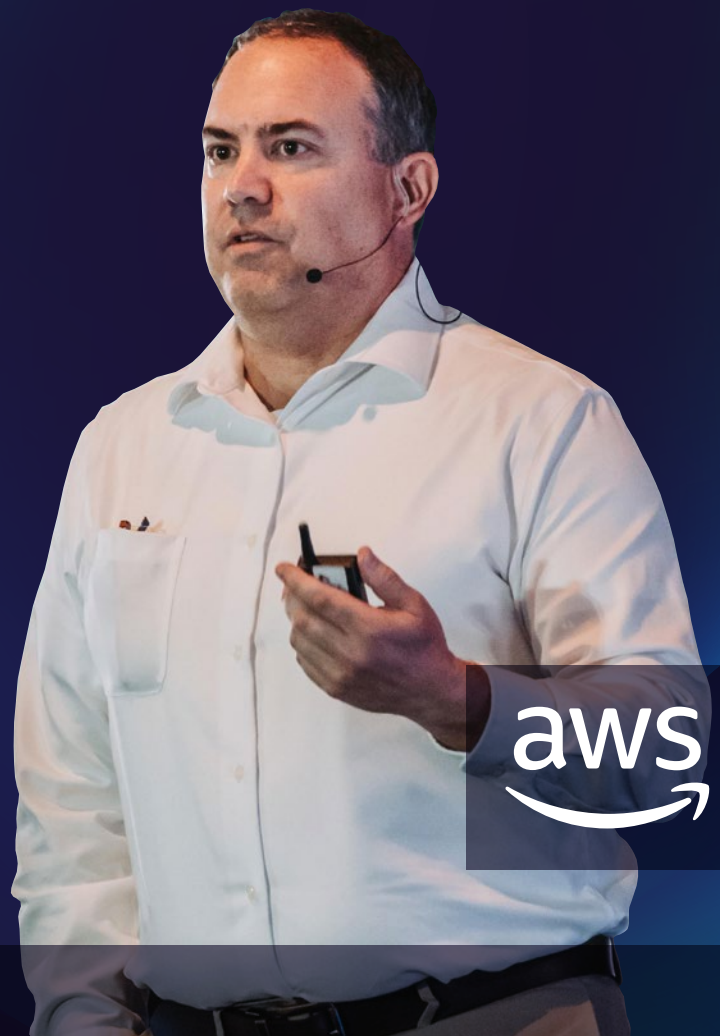
01 **Network & Connectivity**
The largest private network (37 regions, 117 AZs, 6M km fiber) optimized with Elastic Fabric Adapter (EFA) and Scalable Reliable Datagram (SRD) for low-latency, high-throughput connections.

02 **Advanced AI Network**
The 10P10U network for AI, targeting tens of petabits/second and sub-10 microsecond latency, utilizing Scalable Intent and Routing (SIR) for rapid issue resolution.

03 **GPU Compute Offerings**
A wide range of accelerated compute, including NVIDIA Blackwell-based P6 instances (8 chips, 1.4TB HBM) and Ultra Servers (72 GPUs, 360 FP8 petaflops), with new water cooling for high-power devices.

04 **Custom AI Chips**
Tranium chips for price-performance leadership in AI training, using systolic arrays for efficiency, with Ultra Server variants available.

05 **Resilience & Checkpointing**
Addressing failure at scale with SageMaker HyperPod, offering automatic recovery, checkpointing, and new tiered checkpointing for faster recovery (seconds vs. minutes) and advanced observability tools.



PRESENTATION
SCALING MISSON-CRITICAL INFRASTRUCTURE: THE ART OF BALANCING INNOVATION & EFFICIENCY

ANAHITA MOURO, GLOBAL LEAD, DATA CENTER FIELD RELIABILITY & QUALITY, **GOOGLE**

01 **Cost of Failure**
A single failure during large language model training can cost between $4 million and $200 million, plus months of production delay, highlighting the financial impact of unreliability.

02 **Industry Conflict**
Hyperscalers face a core conflict between design teams pushing for the latest technologies and the mission-critical mandate for efficiency and reliability, creating a high-stakes balancing act for all teams involved.

03 **Financial Impact of Changes**
Changes can result in 6-9 months of delay and a 15-20% increase in CAPEX for large projects, alongside opportunity loss from delayed capacity delivery.

04 **Data-Driven Design**
The industry needs better practices for designing for performance by integrating and mining data from design, commissioning, delivery, and operational costs across various databases to train AI models, improve predictability, and close the design feedback loop.

05 **Cost of Variation Metric**
Google developed a metric to measure the cost of variation from efficient delivery, tracking hard and soft costs of deviations across project stages to emphasize the importance of managing changes as early as possible.

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

8,000 REGISTERED
400 SPEAKERS
250+ PARTNERS

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

2026 TRAILER VIDEO

SPONSORSHIP OPPORTUNITIES?
CONTACT THOMAS NOSZCZAK
Thomas.Noszczak@kisacoresearch.com

> REGISTER YOUR INTEREST IN 2026