PRESENTATION

Tying it all together: Considerations for AI-Ready Datacenters

4:05 PM - 4:25 PM

PRESENTER

Andrew Buss

Senior Research Director, Datacenter Infrastructure and Services, **IDC**

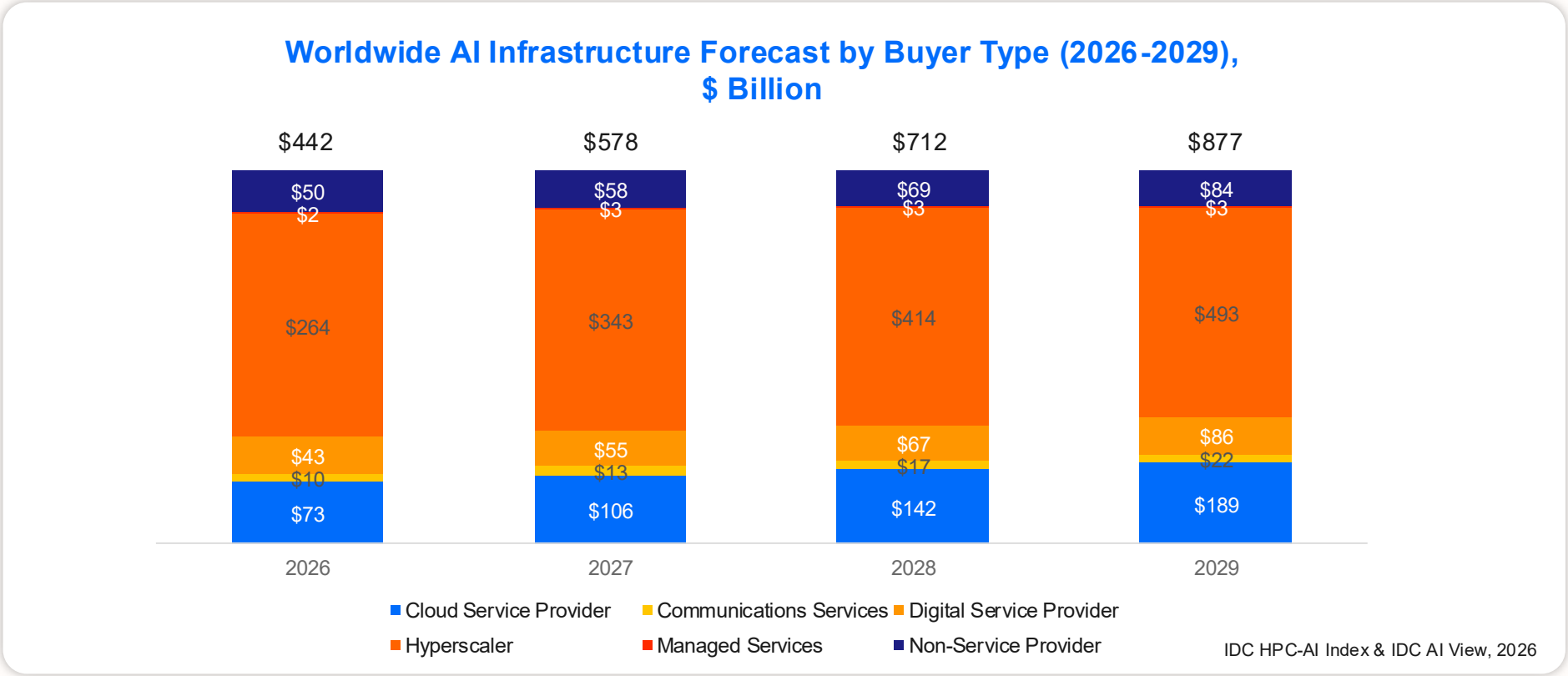
PRESENTER

Madhumitha Sathish

Research Manager, Performance Intensive Computing, **IDC**

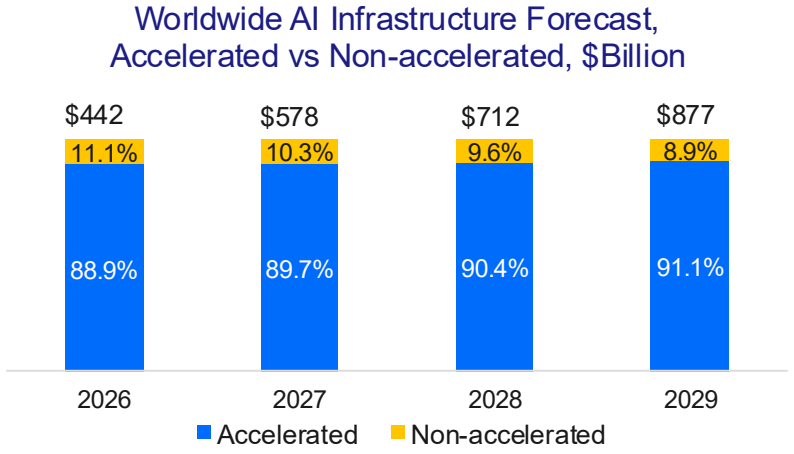
IT budgets are being disrupted by AI infrastructure

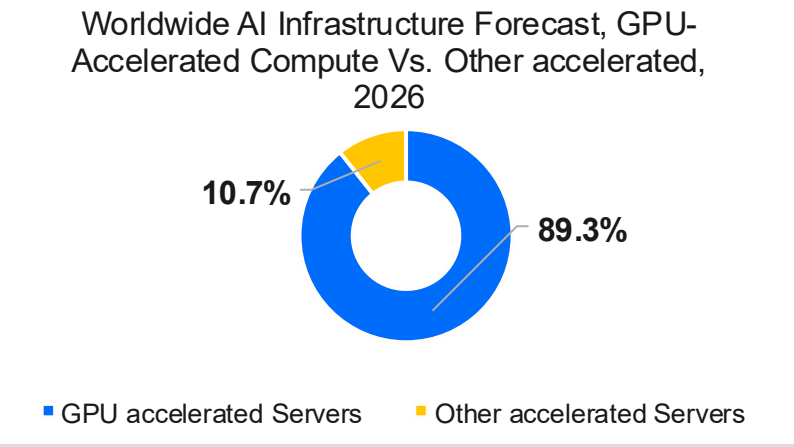
- Organizations expect AI to become their largest workload in 2 ½ years – upsetting IT budgets that had been predictable for many years
- Spending on AI servers and Storage will grow from $442 billion in 2026 to $877 billion in 2029, with 6 distinct buyer categories:



AI readiness demands major infrastructure overhauls

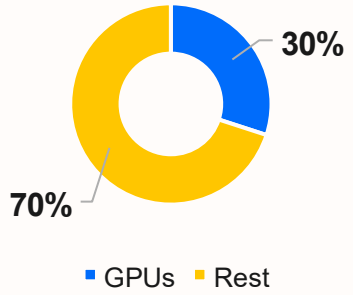
> 50% of general-purpose infrastructure needs an AI-driven overhaul



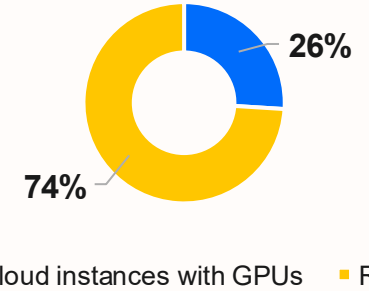


For enterprises, hardware (servers, storage, networking) represents **47%** of the entire AI initiative budget, inclusive of cloud, software, 3rd party services, and staff.

Within hardware, servers with GPUs are the largest cost factor – 30% of the entire hardware budget



Outside of hardware, cloud instances with GPUs are the largest cost factor – 26% of the non-hardware budget

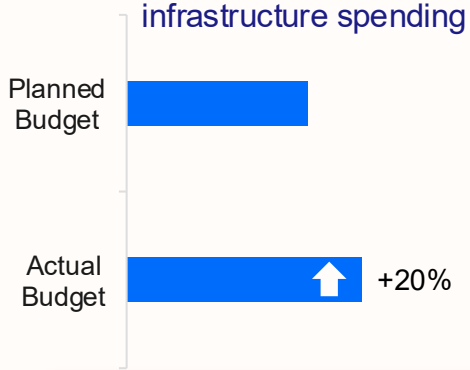


Underestimated infrastructure, overestimated ROI

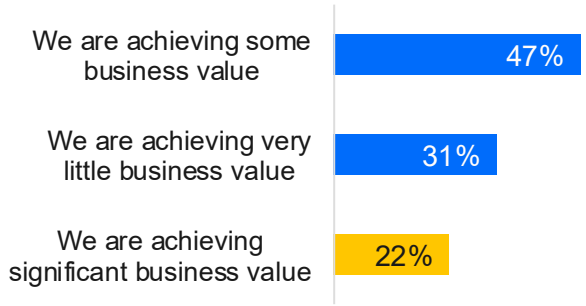
33% of AI initiatives require mid-project budget adjustment due to ONE Specific reason:

Unanticipated IT Infrastructure Spending

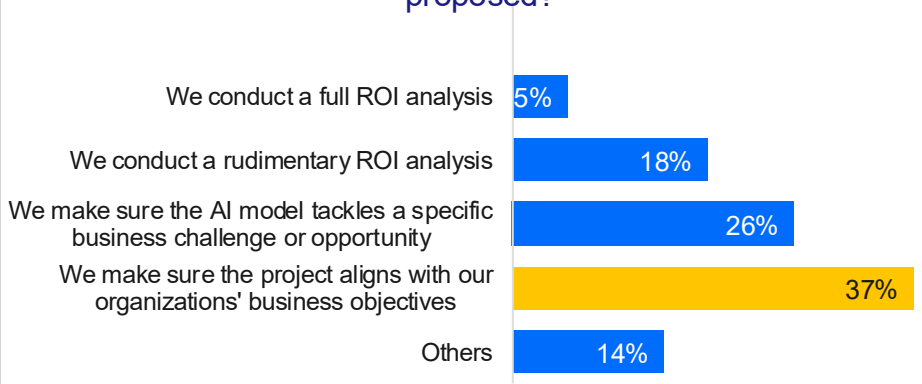
Average budget increase of 20% as a result of unanticipated IT infrastructure spending



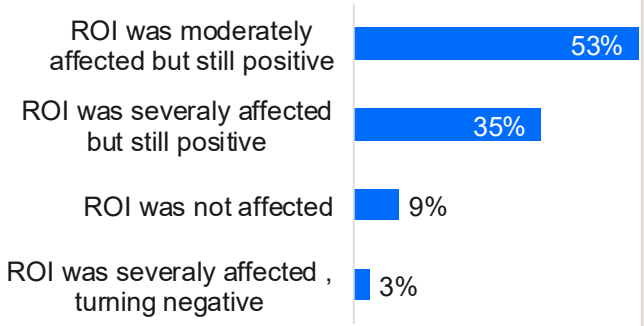
How does your organization view the business value achieved from all its current (in production) AI projects?



How does your organization typically assess the business value of AI use cases that have been proposed?



How did adjusting the IT infrastructure budget for these AI initiatives typically impact your overall ROI?



AI ROI breaks when infrastructure is ignored

AI Use Case Initiated

Only 8% have AI initiatives originate in a cross-functional team with all stakeholders, including IT

ROI Modeling

Only 1 in 5 include IT in ROI formulation

INFRASTRUCTURE REALITY

7 cost drivers rarely assessed upfront:

Model type | Parameter count | Training data volume | Required accuracy | Time to value | Query size | Response time

3 Capital Impacts:

Compute density | Power & cooling | Datacenter allocation

AI infrastructure decisions determine enterprise capital exposure.

AI is moving fast — Infrastructure will decide its future

Where is AI Moving?

1. Updated datacenter
2. New datacenter
3. Legacy datacenter
4. Public cloud
5. Modular datacenter

Most enterprises are expanding or rebuilding owned infrastructure — not defaulting to public cloud.

What's Straining the System?

1. Increasing energy costs
2. Processor complexity
3. Infrastructure optimization skill gaps

AI infrastructure risk is operational, not experimental.

What Determines Long-Term Viability?

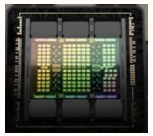
1. Thermal management & cooling
2. Dynamic power management (throttling)
3. Smaller footprint

AI sustainability is determined by power efficiency and cooling capacity.

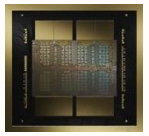
Power density per rack is accelerating enormously in AI datacenters as reasoning and Agentic AI take hold

TDP per SXM GPU (W)

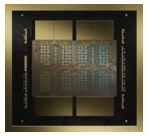
→ 2020 – A100 — 400



→ 2022 – H100 — 700



→ 2024 – B200 — 1,000



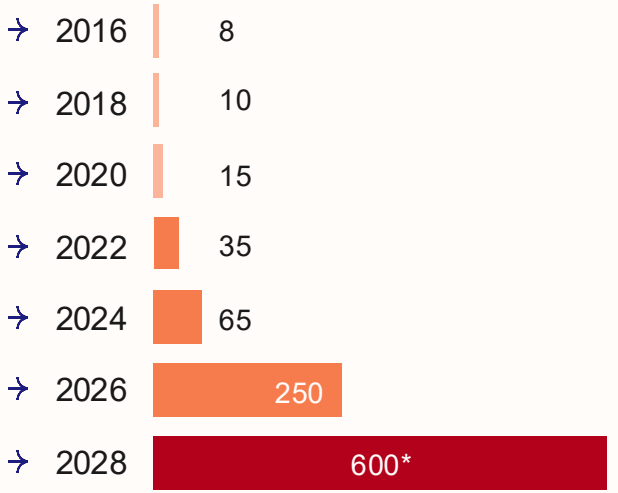
→ 2025 – B300 — 1,400

Datacenter GPUs per Rack

2022 – 32 air-cooled GPUs

2025 – 96 Liquid Cooled GPUs

Evolution of kW per Rack for GPU-Accelerated GenAI Compute



* - NVIDIA Feynmann GPU generation planned power density per rack disclosed at NVIDIA GTC 2025

A little bit here, a little bit there, and pretty soon you're talking multi-GW datacenters

A little bit here, a little bit there, and pretty soon you're talking multi-GW datacenters

Mid-size AI Datacenter

1,000 Racks
Average 3-rack power density 300KW

300MW

A little bit here, a little bit there, and pretty soon you're talking multi-GW datacenters

Mid-size AI Datacenter

1,000 Racks
Average 3-rack power density 300KW

300MW

Large AI Datacenter

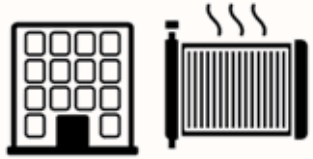
10,000 Racks

Average 3-rack power density 300KW

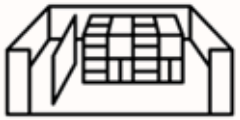
3 GW

Extreme power densities of AI need effective liquid cooling options

Elements of a direct-to-chip liquid cooled datacenter that turn it into an AI factory:

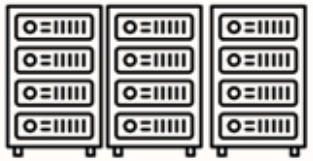


Datacenter facilities-side plumbing, fluid, and heat exchanger



Facilities-side liquid distribution manifolds and heat exchangers in datahalls

Coolant distribution units (CDUs) in racks, rows, or datahalls



Hybrid or 100% direct-to-chip cooling loop with advanced technical cooling fluid

The suitability of different air and liquid cooling approaches based on rack-level power density

Air cooling

Rear-door heat exchanger

Hybrid direct-to-chip liquid cooling and rear-door heat exchanger

100% direct-to-chip liquid cooling

Adopters and Enablers of AI-ready datacenters

Hyperscalers, Frontier providers, and AI Service Providers

aws
ANTHROPIC
Alibaba
CoreWeave
Google
Meta
Microsoft Azure
NSCALE
NEBIUS
ORACLE
OpenAI
xAI

Datacenter Facilities Services Providers and hosting providers

Aligned Adaptive Data Centers
CyrusOne
colt Data Centre Services
DATAONE SUSTAINABLE DATACENTER
DIGITAL REALTY
EQUINIX
NTT DATA
STACK INFRASTRUCTURE

Enterprises and public institutions

BriCS Bristol Centre for Supercomputing
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
NRIS
OLIVIA
Lawrence Livermore National Laboratory

Datacenter Critical Infrastructure Providers

ABB
EATON
HITACHI
Schneider Electric
VERTIV
CoolIT systems
GE VERNOVA
Johnson Controls
SIEMENS

Key takeaways for AI-ready datacenters

1. The sheer scale of AI related compute and associated infrastructure investment and deployment is at a level we have not seen before in enterprise IT

2. The drive to maximise efficiency and scale for AI to drive down the cost of token generation is leading to a high level of vertical integration at the rack and cluster/datahall level

3. The unprecendented high levels of power density of the new and upcoming engineered AI systems platforms has fundamentally changed demands on datacenter design, power delivery, and cooling requirements

4. For successful AI factory rollouts, AI-ready datacenters need to be co-designed and optimized alongside the rack-scale AI system engineering

5. AI-ready datacenters will be optimized around convenient modular building blocks, with datacenter critical infrastructure vendors already offering solutions able to scale in 10MW units for power delivery and cooling

IDC

DIRECTIONS