IDC Directions

Digital Infrastructure

Fit-for-purpose tech stack for the AI and post-AI era

Ashish Nadkarni
GVP/GM, IDC

Peter Rutten

Moving to agentic AI

Is it a set of use cases?

Is it embedded within my workloads?

Is it a set of functions?

Is it about automation? Autonomous?

Is it standalone?

Is it a workload? or a set of workloads?

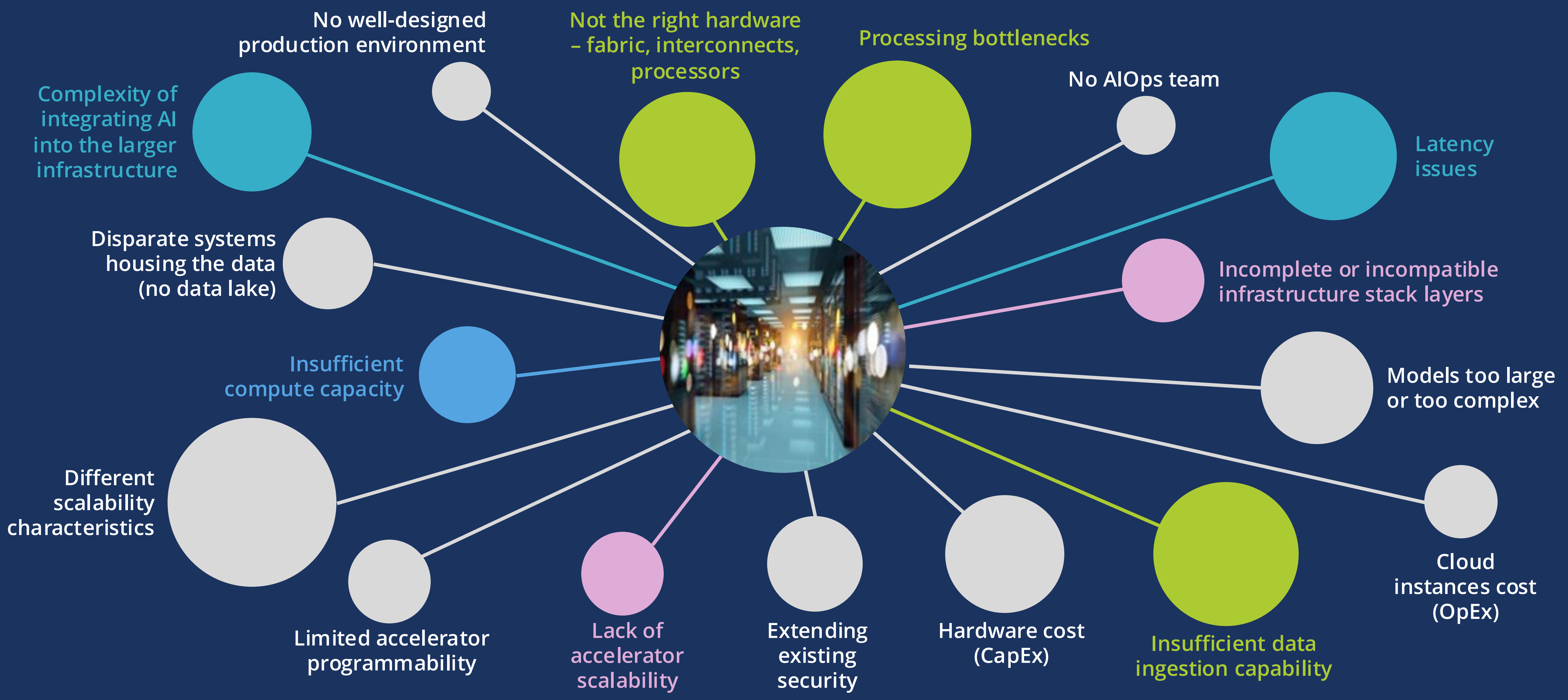
What about quantum computing?

What about training, inferencing, and RAG?

IDC

Peter Rutten

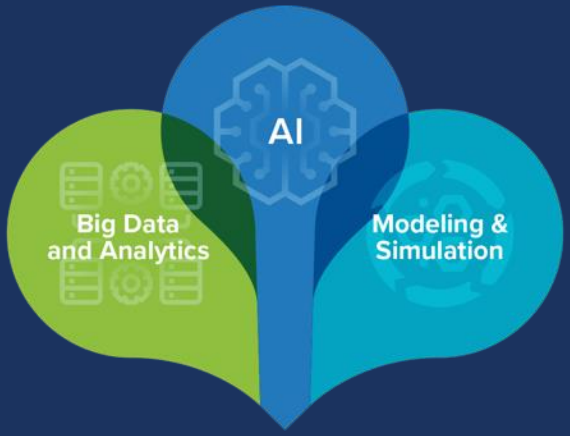
Complexities of AI workloads require a revisit of the tech stack



Peter Rutten

What is different about AI workloads?

AI has a lot in common with modeling and simulation (HPC) and analytics workloads



Performance-Intensive Computing Infrastructure

Workloads that...

- Perform large-scale mathematically intensive computations
- Process large volumes of data
- Have complex instruction sets to be executed in the shortest amount of time
- Are deployed with compressed time-to-insights objectives
- Need to scale on demand, possibly outside of the boundaries of the datacenter
- Require fit-for-purpose infrastructure, more so than general-purpose infrastructure

AI-native is evolving

Era	Enterprise	Dot Com	Cloud	AI	Post-AI
Workload	Monolithic	Tiered	Distributed	Geo-distributed	Composite
Deployment Model	On-premises	Hosted	Cloud	Hybrid	"DePIN"
Infrastructure Stack	Monolithic	Pseudo-Monolithic	Tiered	Composable	Quantum-Classical

IDC

Kuba Stolarski

New compute-intensive AI workloads require a fundamentally different tech stack

	Cloud-native	**AI-native**
Architecture	Web services	AI models
	Thousands of workloads Many nodes	One (composite) workload Thousands of nodes
Cloud Affinity	Multi-cloud	Hybrid-by-design
Ecosystem	Locked-in (Rigid)	Expansive (open to accelerate AI innovation)
Sustainability	Energy inefficient	Sustainable by design

IDC

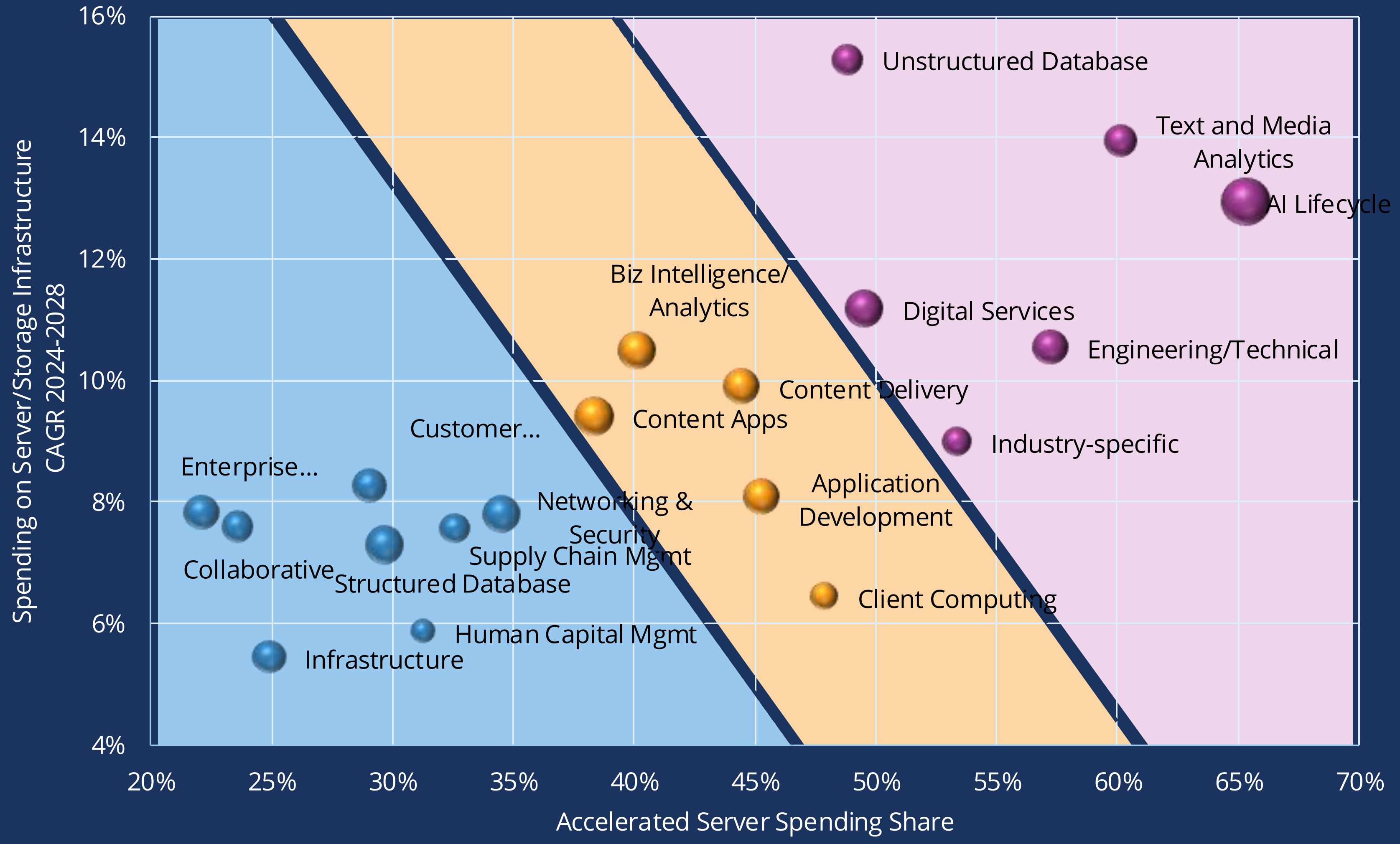

Kuba Stolarski

Infrastructure evolution for AI-native

	Cloud-native	AI-native
Operating strategy	Expert-led	Policy-driven
Consumption model	I&O driven	Model-driven
Observability	Human created, Machine readable	Machine created, Human readable
Control	System-specific	Workload-driven
Access	Restful	Model-specific
Deployment	Storage and computing platforms	Control and management
System architecture	Cloud-centric	Model-centric

Natalya Yezhkova

Enterprise workloads become AI-native



- Top enterprise workloads requiring more accelerated infrastructure are also among the fastest growing
- Most workloads with moderate level of demand for accelerated computing support AI development and analytics
- Workloads with colder demand for accelerated servers tend to be more mature, some are typically considered mission-critical

Source: IDC's Worldwide Semiannual Enterprise Infrastructure Tracker: Workloads, November 2024


Heather West

The true meaning of xPUs

(The intersection of CPUs, NPUs, GPUs, and QPUs)

Pre-processing

Local

Scientist
Engineer
Developer

Classical Compute
CPU | NPU
Simulations | Linear algebra
Optimization

On-prem/ Public Cloud

AI-assisted software platforms
CPU | GPU
Quantum algorithm development
SDK libraries | Hybrid solvers
Hybrid resource orchestration

Distributed hybrid compute

On-prem/ Public Cloud

Classical compute
(HPC/AI Supercomputers)
CPU | GPU

Quantum Compute
QPUs
Gate-based
Quantum annealers
Analog quantum systems

\+

Classical compute
CPU | GPU
Controls and
Measurement | AI-assisted
error correction

On-prem/ Public Cloud

Post-processing

On-prem/ Public Cloud

Classical compute
CPU | GPU
Circuit knitting
Circuit embedding

Scientist
Engineer
Developer

Classical Compute
CPU | NPU
Data analysis

Local

IDC

Quantum-centric supercomputing in the world of agentic AI

Infrastructure evolves in response to increased computational, data logistics and connectivity requirements

2024
Large Model Centric

Classical

- Mix of cloud and dedicated locations
- Supports rapid innovation
- Human centered IT operations augmented incrementally with AI

2025-26
Fit-for-Purpose Optimized

Classical | Quantum

- Agents enable access to wider range of model types and links across different AI types
- Infrastructure selection determined by security, privacy, cost and performance tradeoffs
- Inferencing strategies for edge/campus/branch become a top priority

2027+
Agentic Connectedness

Quantum <> Classical

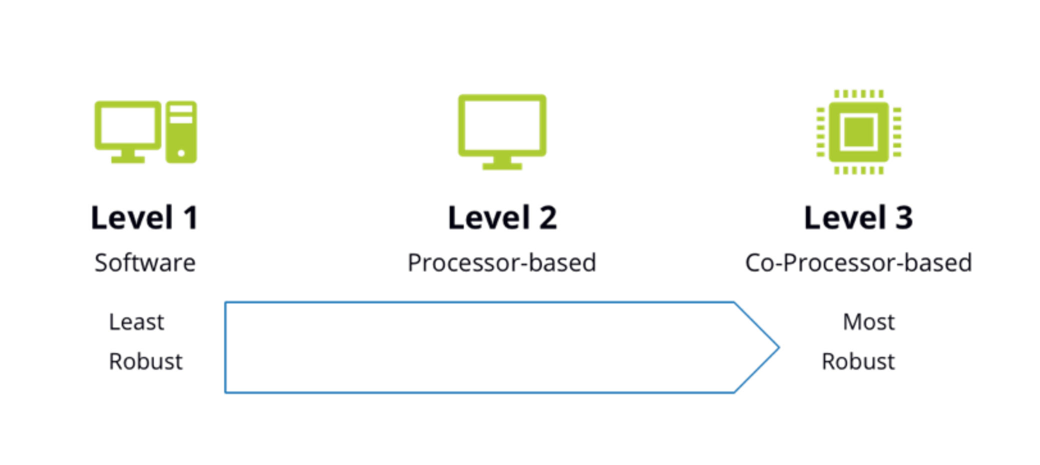
- Robust Agent to Agent Connectedness Drives Highly Distributed Infrastructure that cuts across quantum and classical computing
- Data Logistics and Resilience Demand Agent to Agent Optimization
- Autonomous operations enables continuous scaling and lifecycle operations efficiency

IDC

Dave Pearson

Composable infrastructure approaches offer better scaling

- Virtual networking
- Distributed security
- Full embedded offload
- AI native workload optimization
- Full management offload




Carol Sliwa

Demand for scale-out, software-driven file and object storage

- AI lifecycle is the top workload driving spending on enterprise compute and storage infrastructure
- Data-intensive AI and analytics workloads are driving the use of software-driven, server-based storage to ease the scaling of capacity and performance
- Scale-out distributed file storage and parallel file systems are the leading storage technologies in use with AI workloads
- Security, performance, and scalability are the most critical priorities for organizations selecting on-premises storage infrastructure for AI workloads

IDC

The evolving and expanded role of networking in AI

Enterprise investments in AI are a secular growth driver for the datacenter networking market.

Historically, growth trajectory has been levered to cyclical drivers such as technology transitions and to traffic growth.

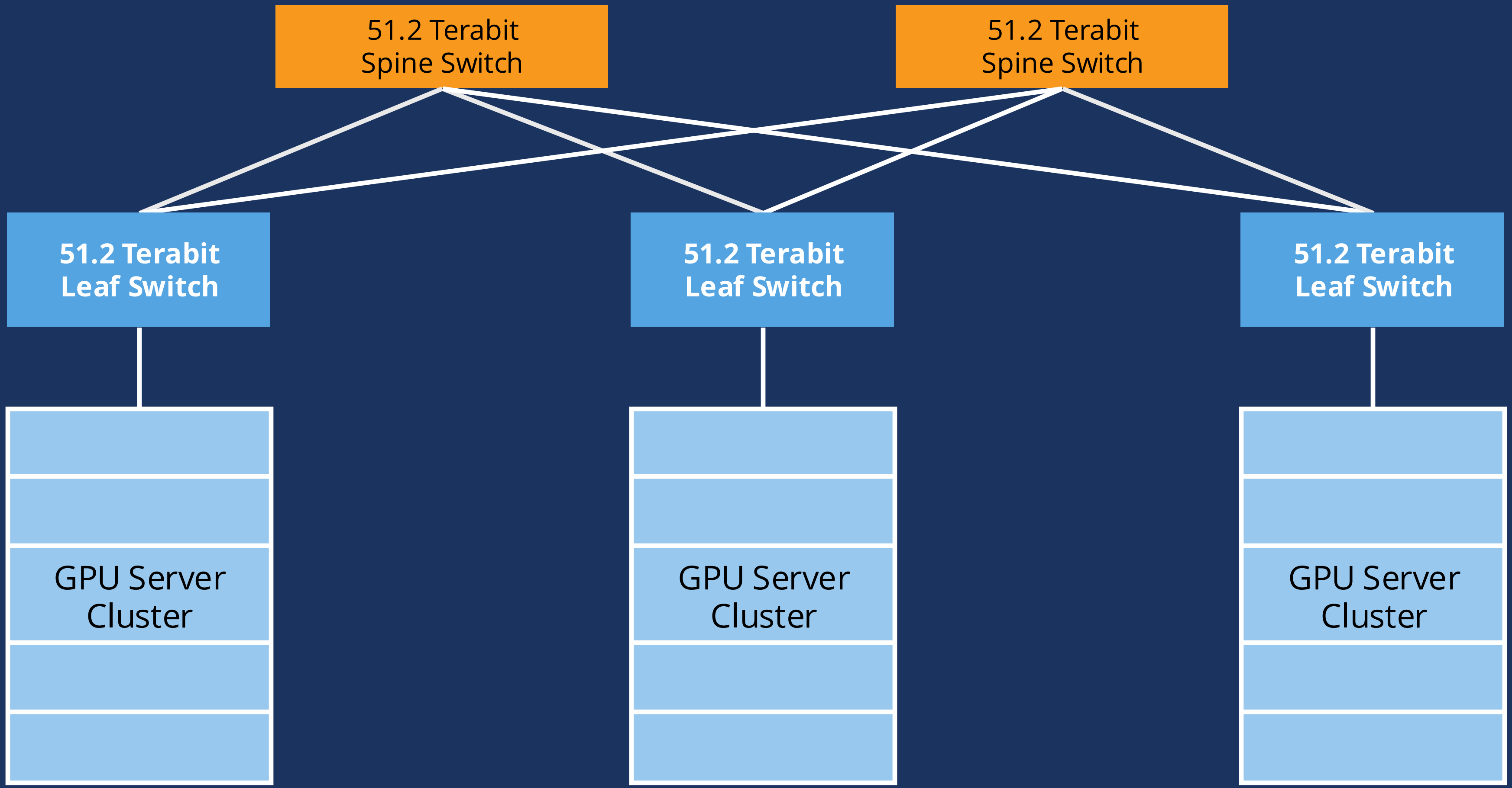
Interoperability	**Multicloud, Hybrid cloud**
Scaling	**Remove restrictions on silicon**
Simplification	**OpEx and management**
Advanced access	**AI network fabrics and clustering**
Sustainability	**Reduced power draw**

IDC

Ethernet “AI fabric” datacenter switch network enable low-latency interconnection of AI GPU clusters

AI fabric characteristics

- Higher proportion of “elephant flows”
- Job completion times is a critical success metric
- Immense volume of data traversing the fabric. GPT3 trained on 300B words!



IDC

Madhumitha Sathish

Hybrid infrastructure approaches bond together multiple deployment strategies

Self-built and/or managed infrastructure

Interoperability and workload mobility

Deployed in traditional data center, co-lo or dedicated cloud infrastructure

- Data confidentiality
- Private AI
- Customized analytic and dev stacks

Computing platforms and systems	**Storage systems**	Storage and computing **Infrastructure software**
NVIDIA, intel, WEKA.io, cisco, DELL, ddn	XILINX, AMD, FUJITSU, IBM, ORACLE, Red Hat	Hewlett Packard Enterprise, vmware, Lenovo, SUPERMICRO, and others

Infrastructure as a Service

Deployed as dedicated or public cloud infrastructure as a service

- Consumption-based pricing and support
- Standard APIs
- Built-in advanced analytics and dev services

Public and dedicated cloud IaaS (compute and storage)	**PaaS**	**SaaS**
IBM, ORACLE, Hewlett Packard Enterprise	aws, Microsoft Azure, DELL, Lenovo	Google Cloud, and others

IDC


Phil Goodwin

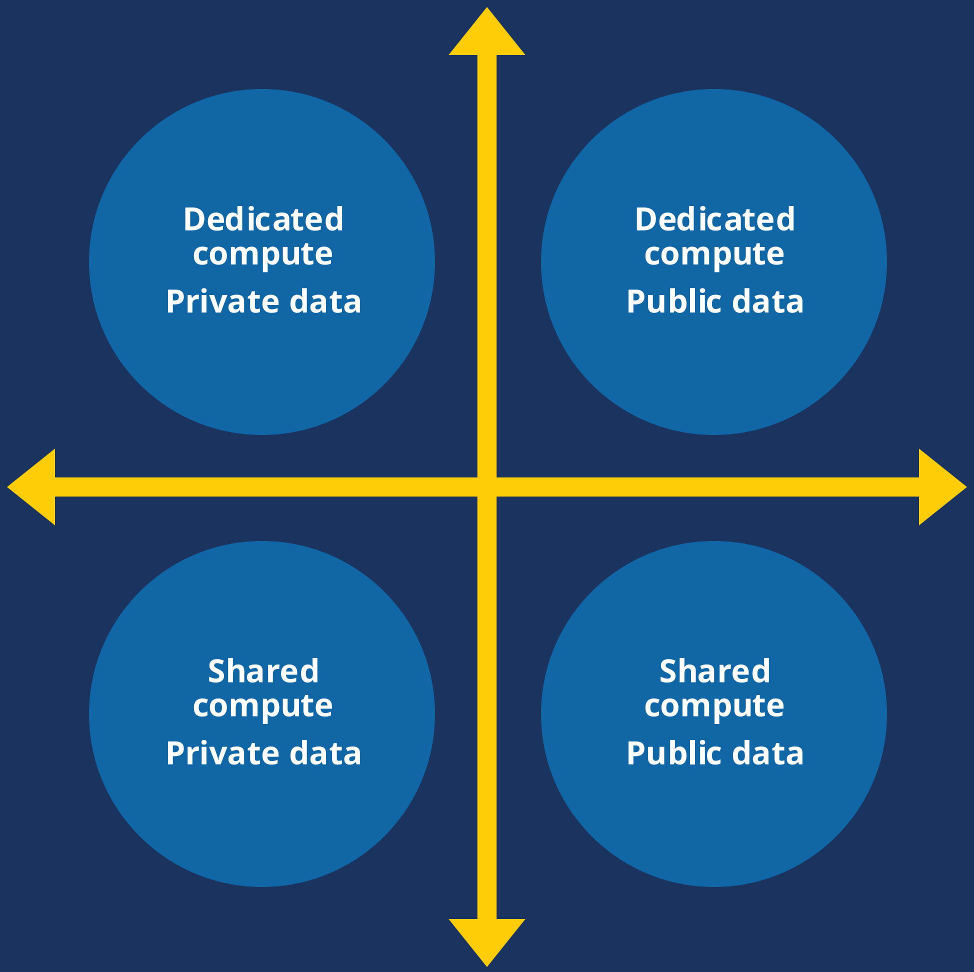
Placement of the AI infrastructure is driven by data-centric decision criteria

Asset inventory

Tagging

Stewardship

Security and compliance



Phil Goodwin

Elements of an AI-ready data storage infrastructure

Requirements

Requirements	Attributes
Performance	• Performance intensive computing • Storage latency • Network bandwidth and quality of service (QoS)
Service levels	• Data availability • 9's of infrastructure availability • High availability/failover
Data logistics	• Data protection, replication and recovery • Data placement • Data sovereignty
Data trust	• Data inspection, provenance, classification and governance • Data security, encryption and masking • Zero trust architecture

AI workloads

Data ingest from external sources

Data creation from internal sources

IDC

Elements of an AI-ready data storage infrastructure


Phil Goodwin

AI-ready data storage infrastructure

Storage software

- Data protection
- Workload migration
- Orchestration
- Data replication
- Reporting

Data exploration

- Data discovery
- Data classification
- Indexing
- Tagging
- Masking
- Copy data mgmt.

AI models / Data staging

- LLMs
- Vector DB
- Agents

Governance

- Provenance
- Sovereignty
- IP ownership

Data trust

- Security
- Zero trust architecture
- Access control

Storage hardware

- All Flash
- Hybrid Flash
- Capacity
- Performance tiers
- Latency
- IOPs
- GPU-direct

Cloud

- Dedicated Cloud
- Public Cloud
- Hybrid Cloud
- Open AI
- Co-pilot

IDC

Olga Yashkova

Peek into the future: decentralized physical infrastructure (DePIN)

DePIN (Decentralized Physical Infrastructure Network) combines physical infrastructure (computing, storage, networking) with blockchain technology to create decentralized networks for various applications. Depending on the project, many DePin networks offer a pay-as-you-go and/or hybrid consumption models

Physical infrastructure
Users can **share real-world physical resources** like servers, storage systems, charging stations, communication networks, etc.

Blockchain technology
Provides a **transparent, secure, and decentralized ledger** for recording transactions and ownership.

Token-based incentives
Incentivizes users to contribute resources, participate in governance, and provide services on the network.

Smart contracts and Consensus Mechanism
Self-executing contracts with the terms of the agreement directly written into code **automate processes and enforce rules** without intermediaries.

Benefits of DePINs

- Removes intermediaries and simplifies processes.
- Enables anyone to participate and benefit from the network.
- Removes dependance on centralized systems, making the network more resilient to failures.
- Promotes innovation by empowering communities to build and manage infrastructure.

IDC

Heather West

Post-quantum cryptography is an important consideration when examining "tech stack security"

Technology description

PQC systems and applications use cryptographic algorithms that are secure against attacks from classical and/or quantum computers

Adoption

NIST PQC standards will be released in 2024. Per NSM-8 and NSM-10, federal agencies are migrating to PQC to be quantum resilient by 2031.

Benefits

PQC algorithms are designed so that the security is based on new hard math problems that are unsolvable by classical and quantum computers.

Risks

Shor's/Grover's algorithms provide the speedup needed to factor large number and find discrete logarithms that are used to protect today's data.

Critical success factor

Successful PQC implementation allows for crypto-agility and full comprehension of the IT encryption landscape and vulnerabilities.

Investment

Preparing for PQC should include developing knowledge, identifying data with long-shelf value, and developing a PQC roadmap.

IDC

Mary Johnston Turner

How providers can guide IT buyers towards an "AI-ready infrastructure"

Consider AI workload diversity
- ✓ Performance? Standalone or embedded?
- ✓ Centralized or distributed model activities

Implement a decision framework
- ✓ Validate reference architectures for production scale during POCs
- ✓ Establish goals around service level agreements and objectives

Hybrid works best for most scenarios
- ✓ Placement is important to reduce latency and time to value
- ✓ Move compute to data, not the other way around

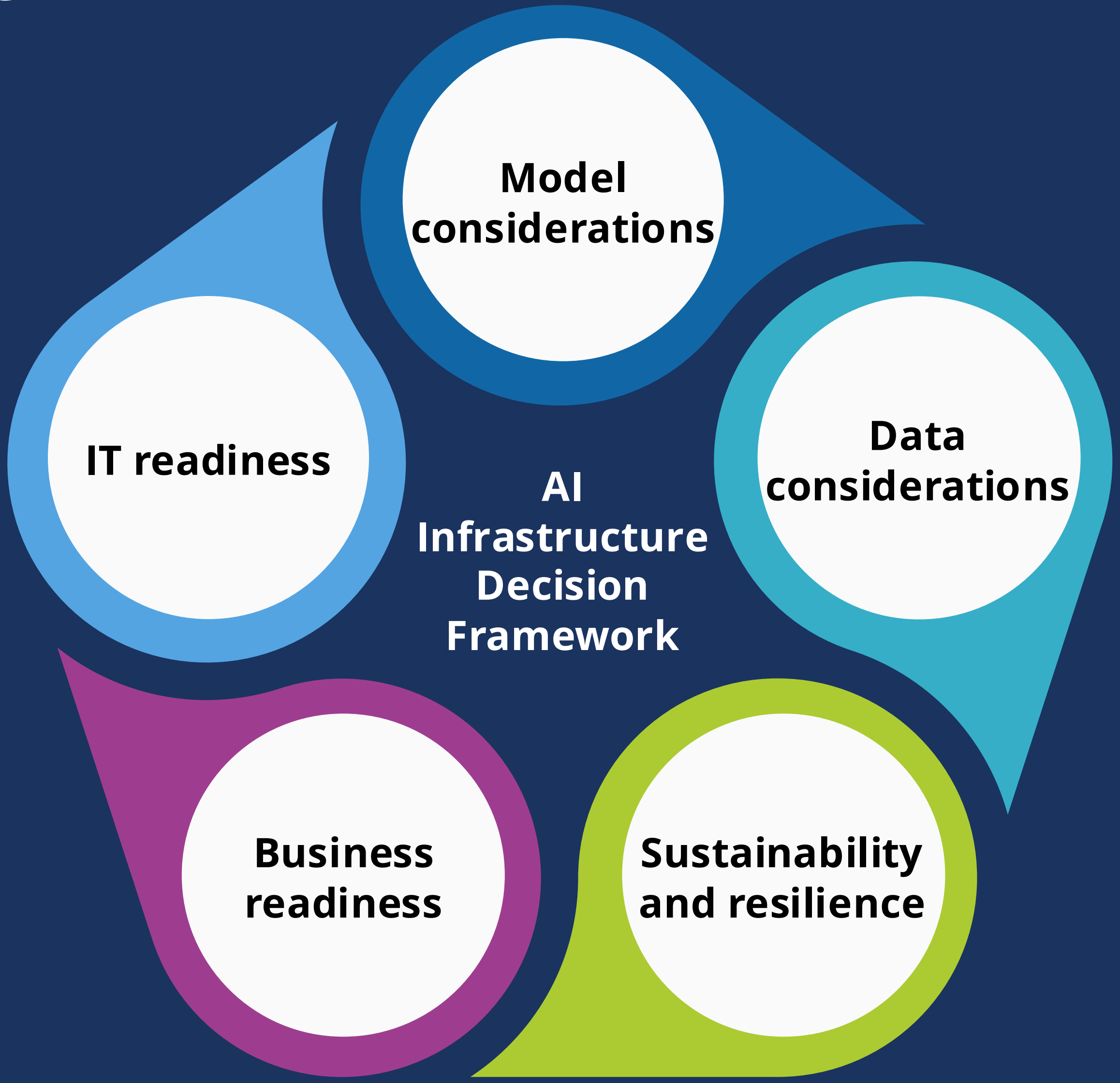
Look beyond GPUs
- ✓ The urgency question
- ✓ Make appropriate fit-for-purpose choices

Other considerations
- ✓ Staff and skills
- ✓ Technical debt

Mary Johnston Turner

How business leaders can guide their organizations towards an "AI-ready infrastructure"

A decision framework ensures POC infrastructure choices can scale as needed in production

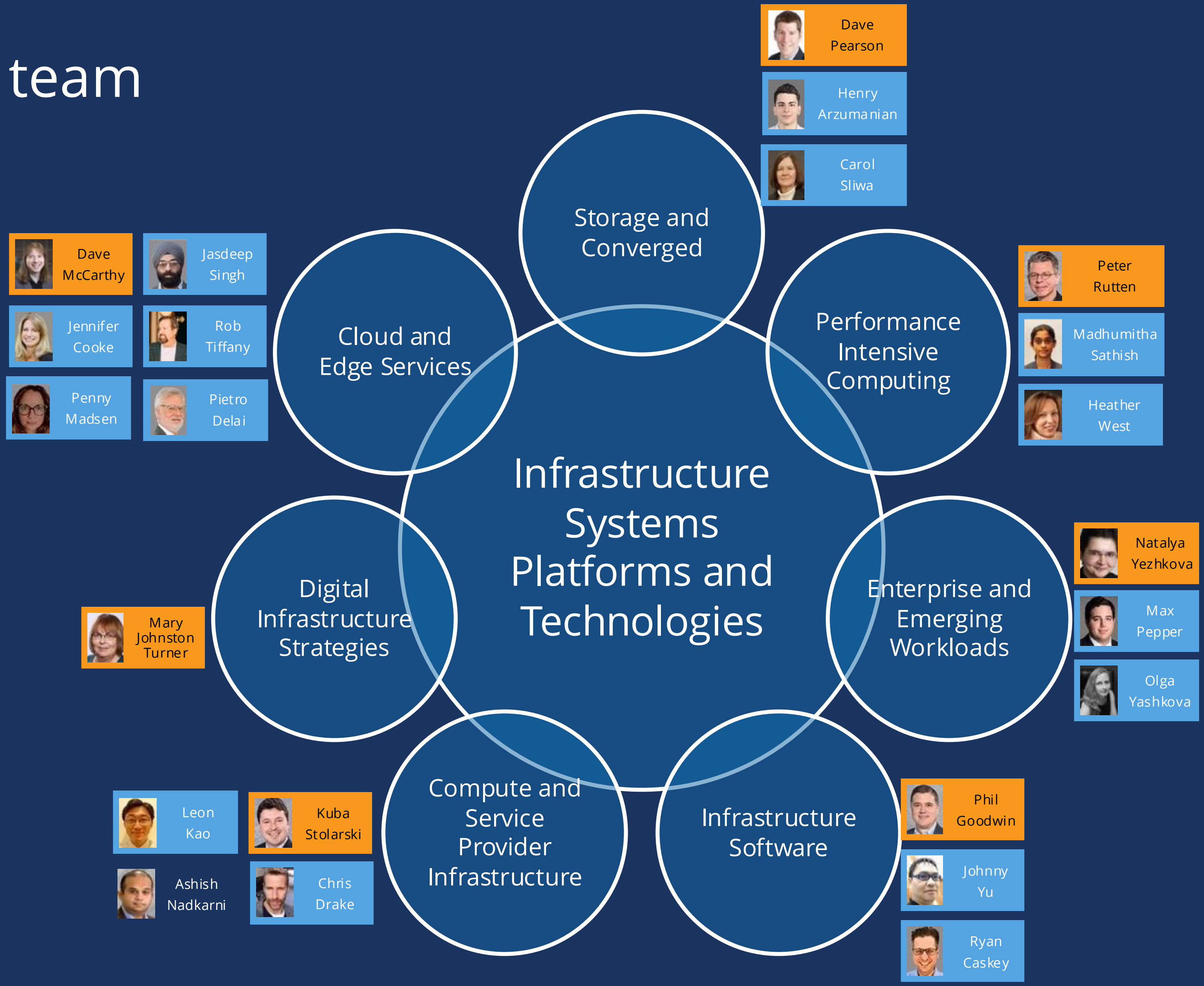


IDC

Elements of the tech stack – How IDC approaches infrastructure research

Partners	OEM and ODM Direct Vendors	Cloud Service Providers	Digital and Managed Service Providers
Operating strategy	Traditional IT	Hybrid IT	Cloud-first
Workloads	Enterprise Workloads	Emerging Workloads	Performance Intensive
Deployments	(Shared) Public Cloud	(Dedicated) Private Cloud	Traditional (non-Cloud)
Infrastructure systems	Computing Systems	Storage Systems	Converged, Hyperconverged and Composable
Infrastructure software	Storage and Computing Platforms	Control and Management	Observability and Automation
Infrastructure hardware	Enabling Technologies	Computing Platforms	Data Persistence Platforms

Meet the team



IDC

In closing…

IDC forecasts Artificial Intelligence to contribute **$22.3 Trillion** to the global economy through 2030 and drive **3.5% of global GDP** in 2030

What this means for you:

- ✓ Increased spending on AI solutions and services driven by accelerated AI adoption
- ✓ Economic stimulus among AI adopters, seeing benefits in terms of increased production and new revenue streams
- ✓ Impact along the whole AI providers supply chain, increasing revenue for the providers of essential supplies to AI solutions and services providers

For additional information

Ashish Nadkarni
anadkarni@idc.com
LinkedIn/in/ashishnadkarni

IDC.COM

LINKEDIN.COM/COMPANY/IDC

TWITTER.COM/IDC