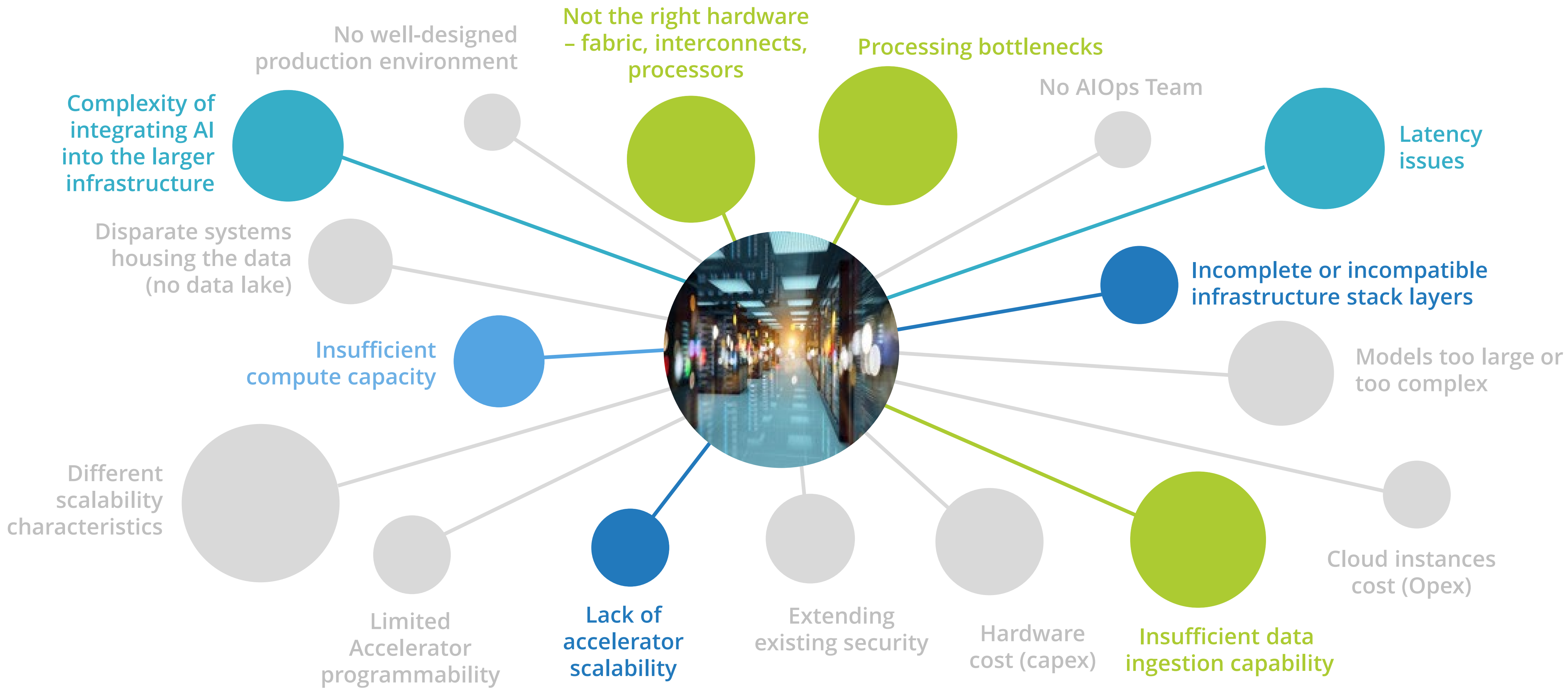
Performance Intensive Computing: Your AI-ready Infrastructure

Extracting value from data in a timely manner

Ashish Nadkarni
Worldwide Infrastructure Research

The current state of Infrastructure for AI

Many AI projects fail because of improper attention to infrastructure



Complexity of integrating AI into the larger infrastructure

No well-designed production environment

Not the right hardware – fabric, interconnects, processors

Processing bottlenecks

No AIOps Team

Latency issues

Disparate systems housing the data (no data lake)

Incomplete or incompatible infrastructure stack layers

Insufficient compute capacity

Models too large or too complex

Different scalability characteristics

Cloud instances cost (Opex)

Limited Accelerator programmability

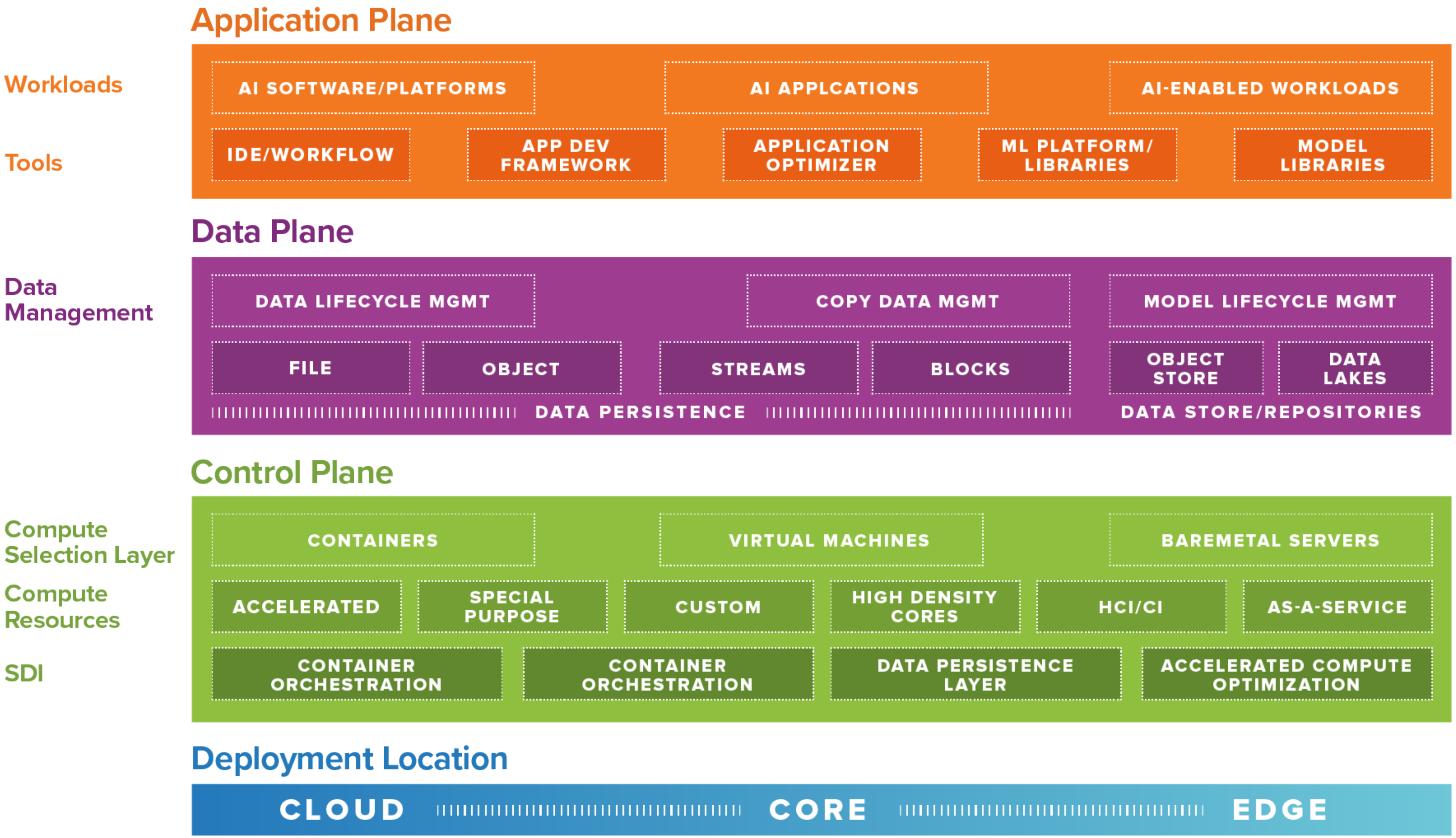
Lack of accelerator scalability

Extending existing security

Hardware cost (capex)

Insufficient data ingestion capability

A homegrown AI Infrastructure Stack is not often designed for scale



Application Plane

Workloads: AI SOFTWARE/PLATFORMS · AI APPLCATIONS · AI-ENABLED WORKLOADS

Tools: IDE/WORKFLOW · APP DEV FRAMEWORK · APPLICATION OPTIMIZER · ML PLATFORM/ LIBRARIES · MODEL LIBRARIES

Data Plane

Data Management: DATA LIFECYCLE MGMT · COPY DATA MGMT · MODEL LIFECYCLE MGMT

FILE · OBJECT · STREAMS · BLOCKS — DATA PERSISTENCE

OBJECT STORE · DATA LAKES — DATA STORE/REPOSITORIES

Control Plane

Compute Selection Layer: CONTAINERS · VIRTUAL MACHINES · BAREMETAL SERVERS

Compute Resources: ACCELERATED · SPECIAL PURPOSE · CUSTOM · HIGH DENSITY CORES · HCI/CI · AS-A-SERVICE

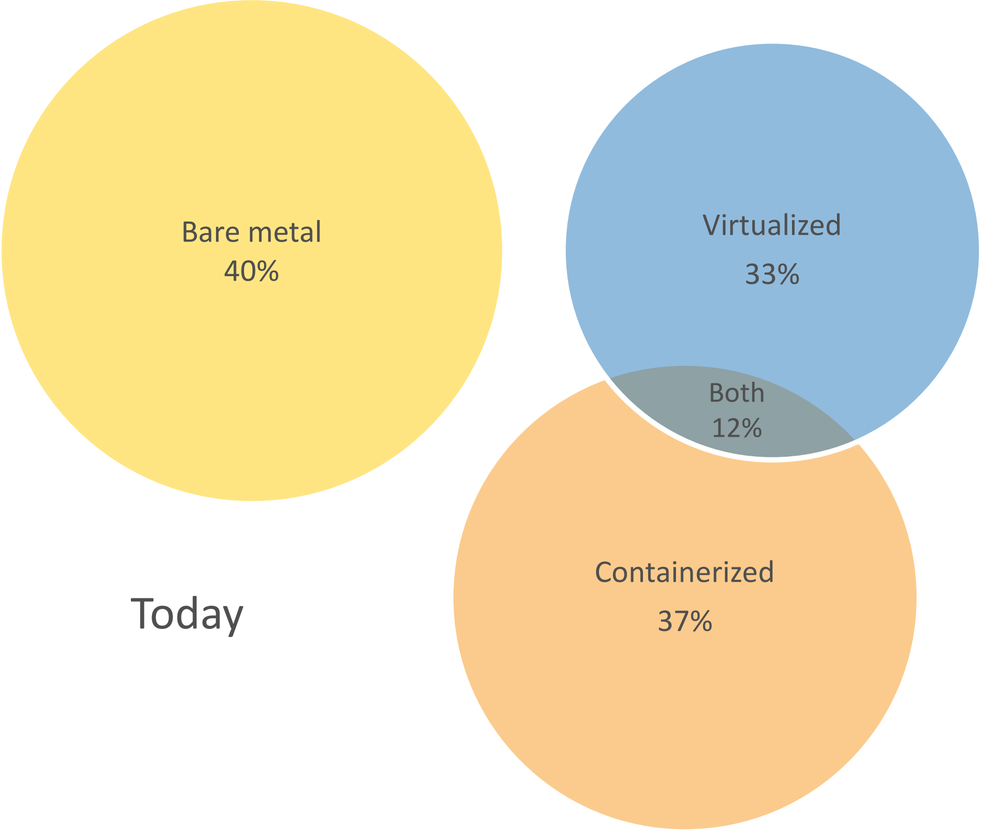
SDI: CONTAINER ORCHESTRATION · CONTAINER ORCHESTRATION · DATA PERSISTENCE LAYER · ACCELERATED COMPUTE OPTIMIZATION

Deployment Location

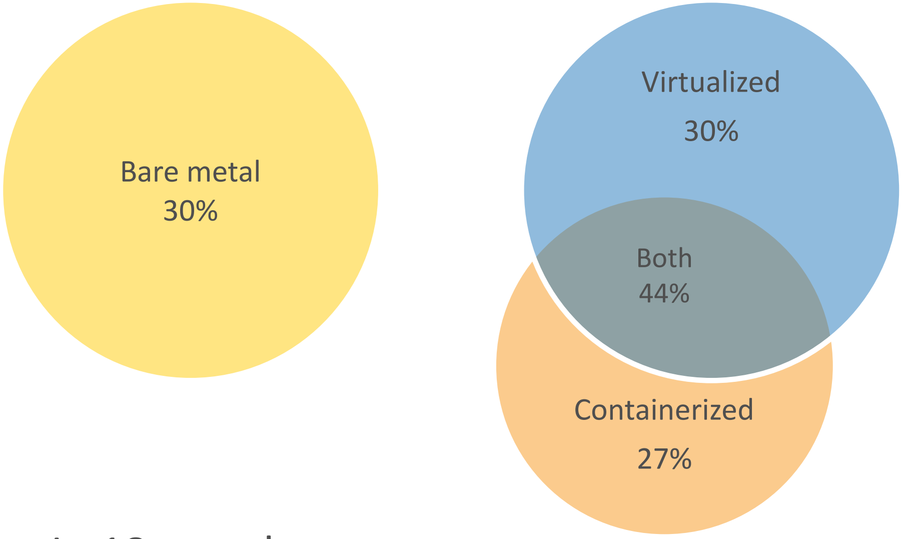
CLOUD · CORE · EDGE

Lack of proper computing considerations can cause problems with outcomes

Less than half (45%) of the compute infrastructure currently used for running AI workloads is virtualized or a combination of virtualized and containerized. This is expected to increase to 74% in the next 18 months



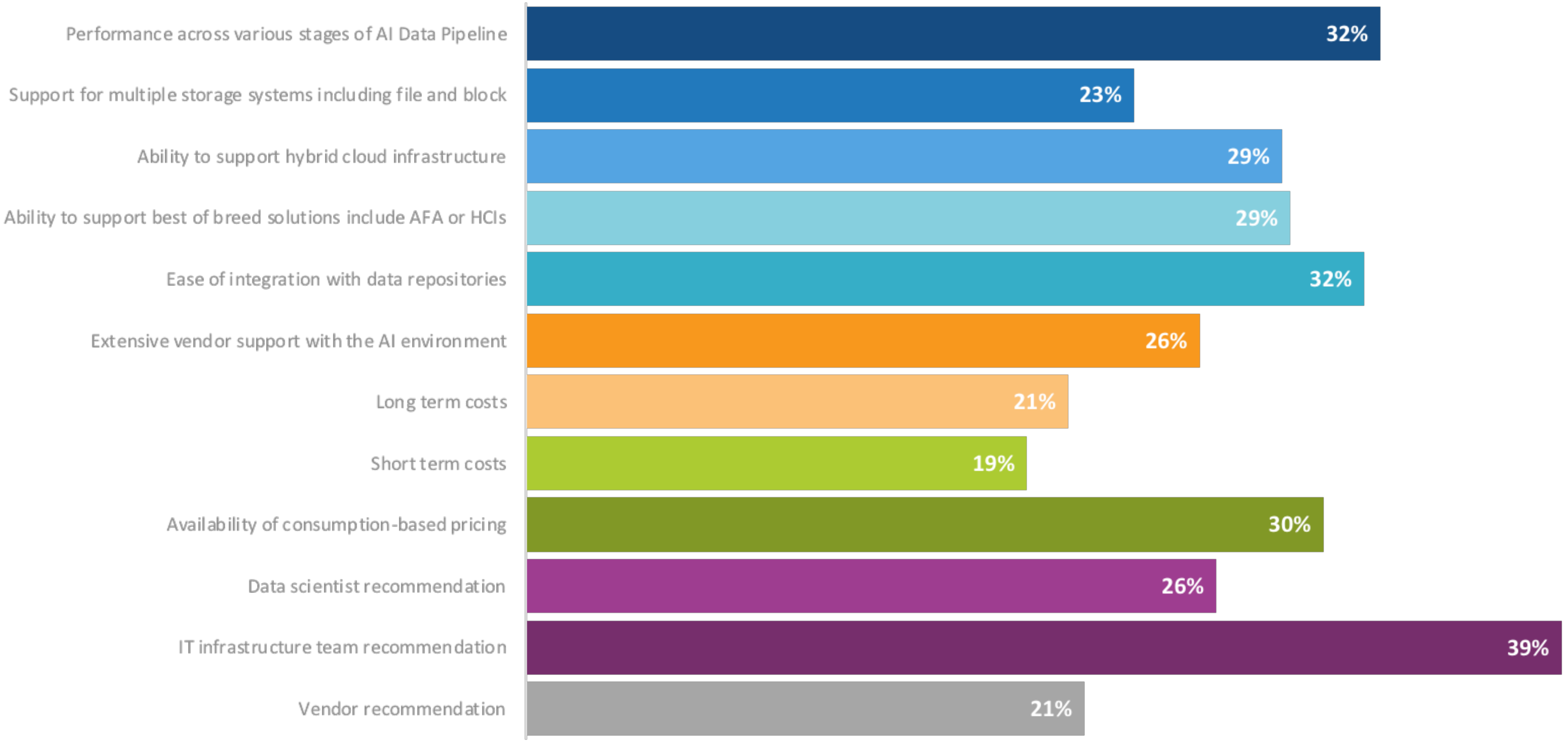
Today



In 18 months

Source: 2023 AI View, September 2023, IDC; N=400

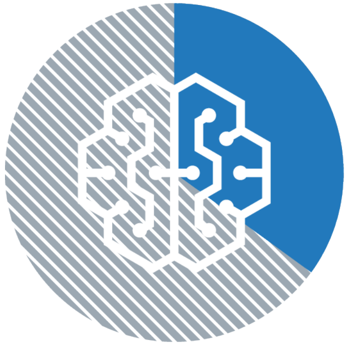
Storage Infrastructure requirements are often not well understood



Q. What were the key requirements when selecting storage infrastructure for AI in you datacenter/colocation provider/edge location?

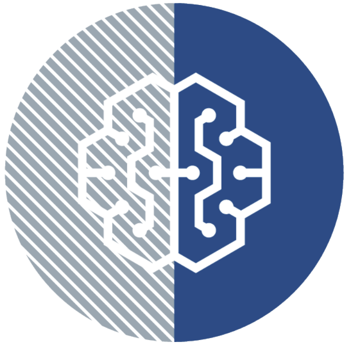
Source: 2023 AI View, September 2023, IDC; N=400

Artificial intelligence is not always standalone; It is being infused into Enterprise Applications



35%

The average number of applications using some form of AI/ML or DL TODAY

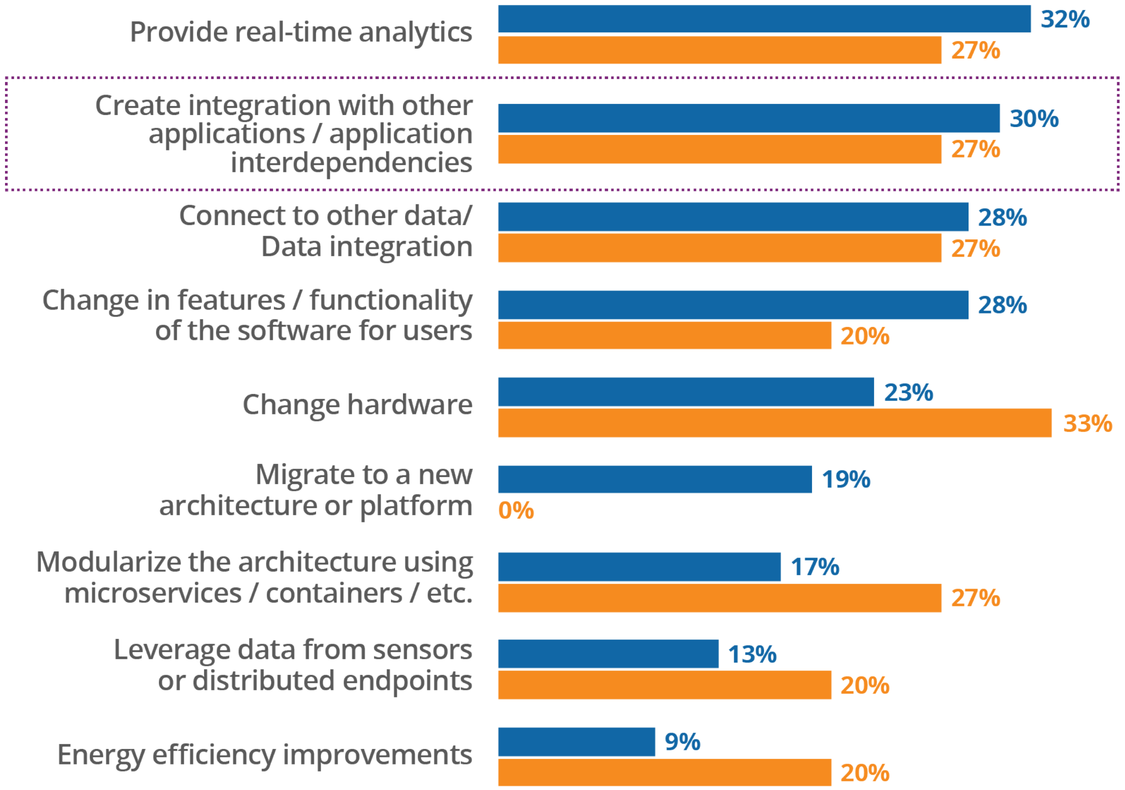


50%

IN TWO YEARS

AI applications will be integrated with other applications across the cloud portfolio

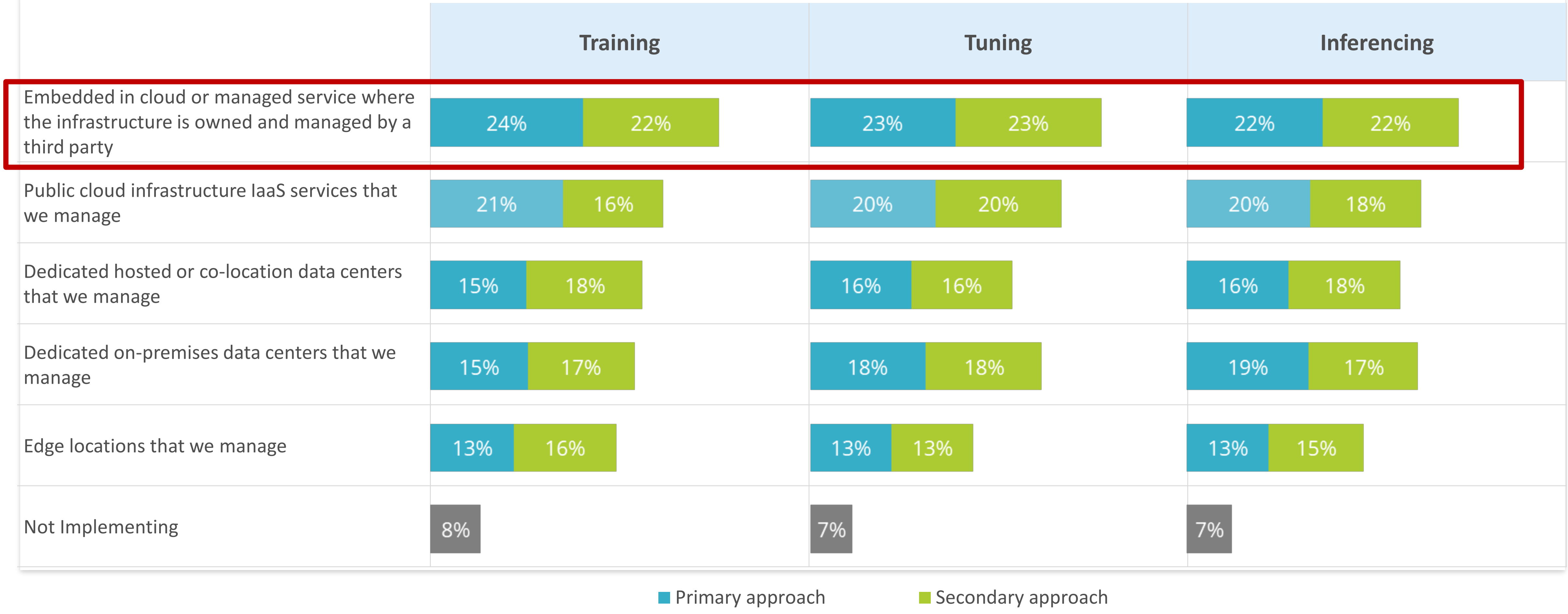
What type of transformation do you expect for your ■ AI software services and ■ AI lifecycle applications?



N = 1,357 Source: QA17, QD11 Cloud Pulse 2Q23, August 2023 IDC

Public Cloud is the de facto approach for a range of Generative AI activities (but it may not always be the best one)

Over the next 18 months, what will be the primary approach which your organization deploys and manages infrastructure resources for GenAI? What will be your secondary approach?



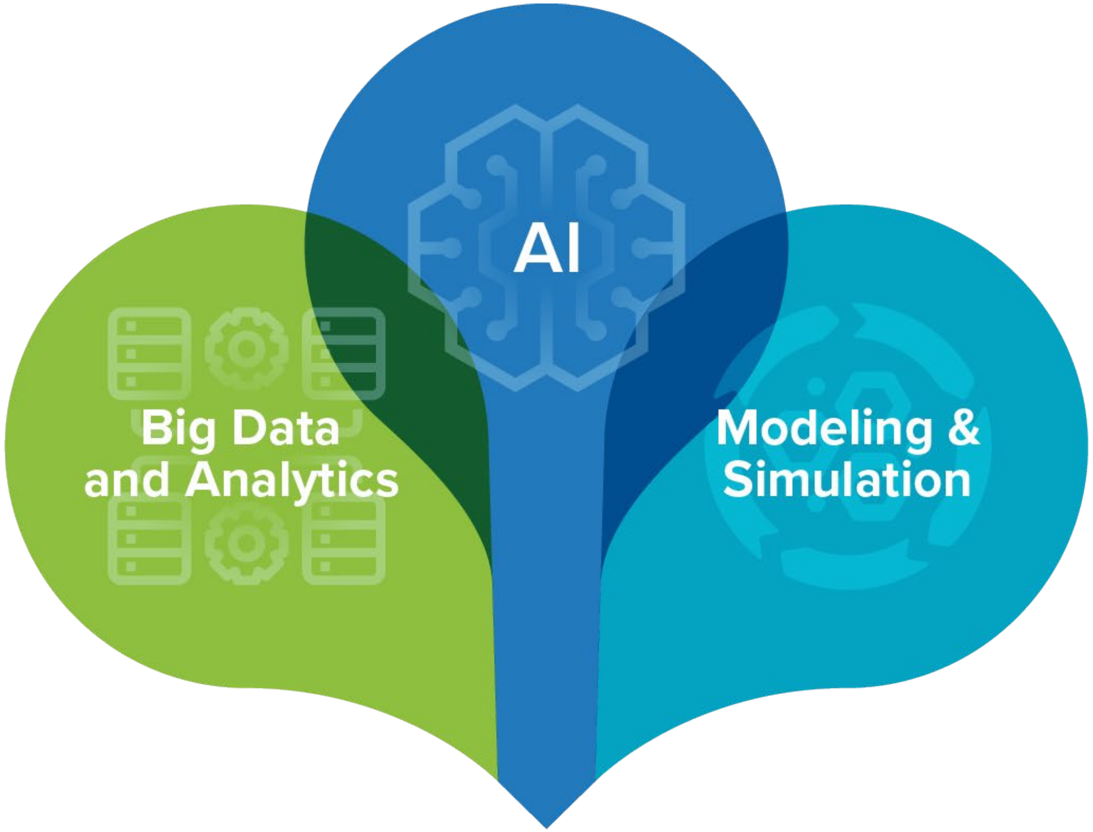
	Training	Tuning	Inferencing
Embedded in cloud or managed service where the infrastructure is owned and managed by a third party	24% 22%	23% 23%	22% 22%
Public cloud infrastructure IaaS services that we manage	21% 16%	20% 20%	20% 18%
Dedicated hosted or co-location data centers that we manage	15% 18%	16% 16%	16% 18%
Dedicated on-premises data centers that we manage	15% 17%	18% 18%	19% 17%
Edge locations that we manage	13% 16%	13% 13%	13% 15%
Not Implementing	8%	7%	7%

🟦 Primary approach 🟩 Secondary approach

Source: Future Enterprise Resiliency & Spending Survey Wave 7, IDC, August 2023, N=883 Data weighted by IT spend (500+ emp size)

Considering an AI Ready Infrastructure

IDC is seeing a Convergence of Three “Workload Groups” onto One Infrastructure Approach



Performance-Intensive Computing Infrastructure

Borrowing scaling approaches from HPC (modeling and simulation) infrastructure for AI workloads

What are Performance Intensive Computing Workloads?

IDC defines workloads are applications and their associated datasets

Workloads that **perform large-scale mathematically intensive computations**

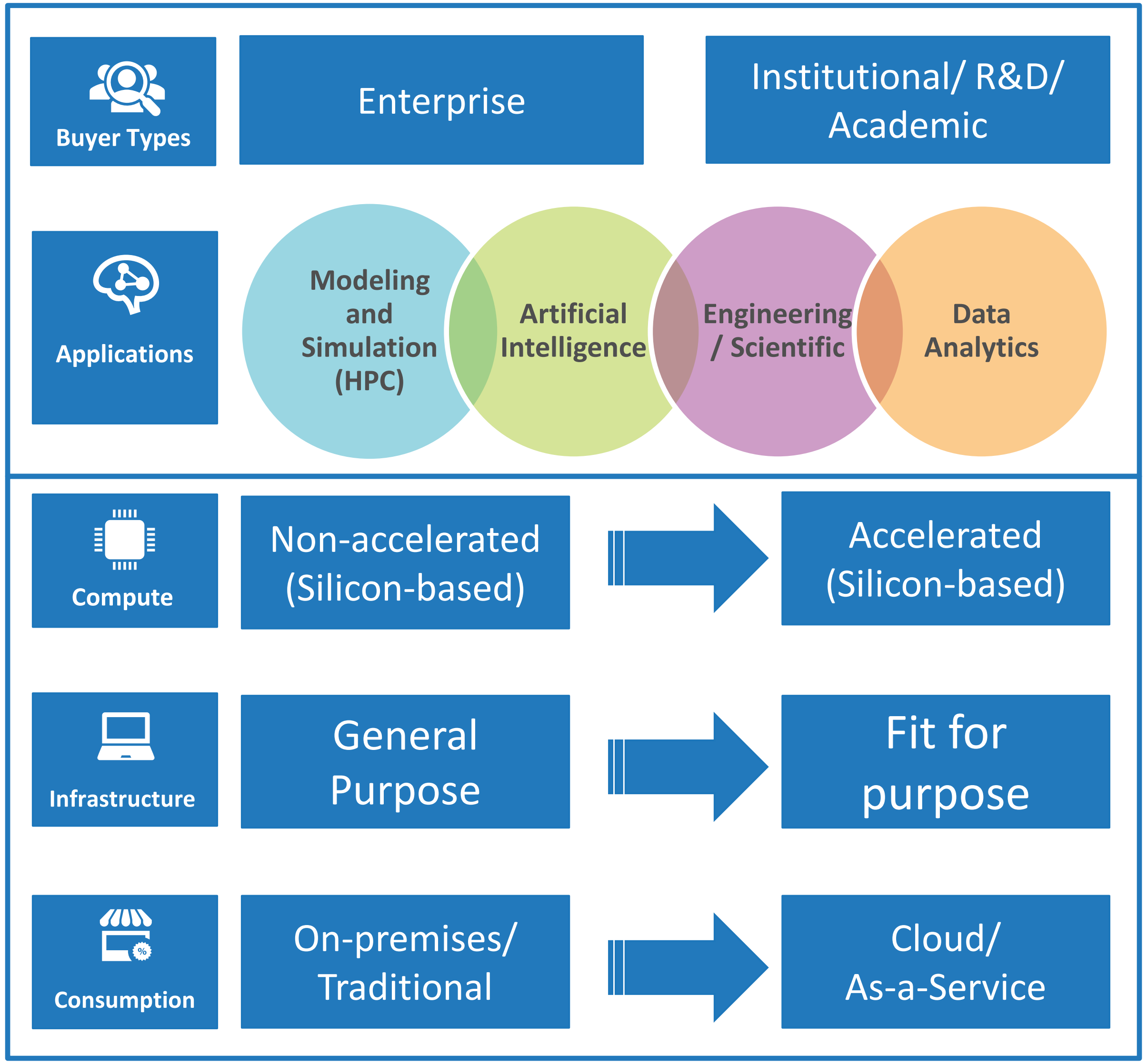
Workloads that **process large volumes of data**

Workloads that **complex instruction sets to be executed in the shortest amount of time**

Workloads that **are deployed with compressed time-to-insights objectives**

Use cases

- **Artificial Intelligence and Machine Learning (AI/ML)**
- **Modeling and simulation (M&S)**
- **Big Data and Analytics (BDA)**
- **Engineering, technical and industry specific**



Buyer Types

Enterprise

Institutional/ R&D/ Academic

Applications

Modeling and Simulation (HPC)

Artificial Intelligence

Engineering / Scientific

Data Analytics

Compute

Non-accelerated (Silicon-based)

Accelerated (Silicon-based)

Infrastructure

General Purpose

Fit for purpose

Consumption

On-premises/ Traditional

Cloud/ As-a-Service

Two Principal Deployments, or as a Hybrid Approach

Self-Built and/or Managed PIC Infrastructure

PIC as a Service (PIC as a Service)

- Deployed as **traditional IT or dedicated cloud infrastructure**

- Deployed as **dedicated or public cloud infrastructure as a service**

Computing Platforms	Storage Systems	Storage and Computing Infrastructure Software	Public and Dedicated Cloud IaaS (Compute and Storage)	PaaS	SaaS
NVIDIA intel WEKA.io cisco DELL ddn	XILINX AMD FUJITSU IBM ORACLE Red Hat	Hewlett Packard Enterprise vmware Lenovo SUPERMICRO and others	IBM ORACLE and others	aws Microsoft Azure	Google Cloud

A Framework for selecting the right infrastructure stack
Part 1 – Cultural/ Deployment (What and How)

For Lower Costs >>>	Prefer on-premises or Collocation (Private Cloud) if	Prefer off-premises (Public Cloud) if
AI Initiatives	On-going AI initiatives from a significantly busy team	On- and off AI initiatives
System utilization	The ability to keep utilization rates very high (keep expensive processors busy)	No ability to keep utilization rates high
IT Skills	Possess in-house skills for for complex AI deployments	Limited IT skills for AI deployments, leave alone complex
Facilities	No limitations on datacenter floorspace, power, and cooling capabilities	Limited floor space, power, and cooling
Opex friendly options	System vendor can provide consumption-based pricing	System vendor can provide capital only pricing

A Framework for selecting the right infrastructure stack
Part 2 – Use case specific (Model considerations)

For Lower Costs >>>	Prefer on-premises or Collocation (Private Cloud) if	Prefer off-premises (Public Cloud) if
Model Iteration	Many model training iterations	Fewer model iterations
Model Scaling	High scaling needs	Lower scaling needs
Model Accuracy	Highly customized	Little to no customization
Model customization	Heavily customized or Inference only	No API changes or customization
Model Performance	High performance requirements	Lower performance requirements

A Framework for selecting the right infrastructure stack
Part 3 – Use case specific (Data considerations)

For Lower Costs >>>	Prefer on-premises or Collocation (Private Cloud) if	Prefer off-premises (Public Cloud) if
Data sensitivity	Highly sensitive data, strict data compliance requirements, proprietary data	Data is not proprietary, no compliance requirements or has been completely sanitized
Data isolation	Model data cannot mix with public data, requires isolation	Model data Data can safely mix with public data, does not require isolation
Time to Value	Not time critical	Highly time critical

When choosing a Cloud Provider for AI

Building differentiation in AI Infrastructure will be the next battle ground for cloud providers

New architectures
Internally developed and 3rd party CPUs and GPUs

Sovereign clouds
Addressing regulatory compliance for data and operations

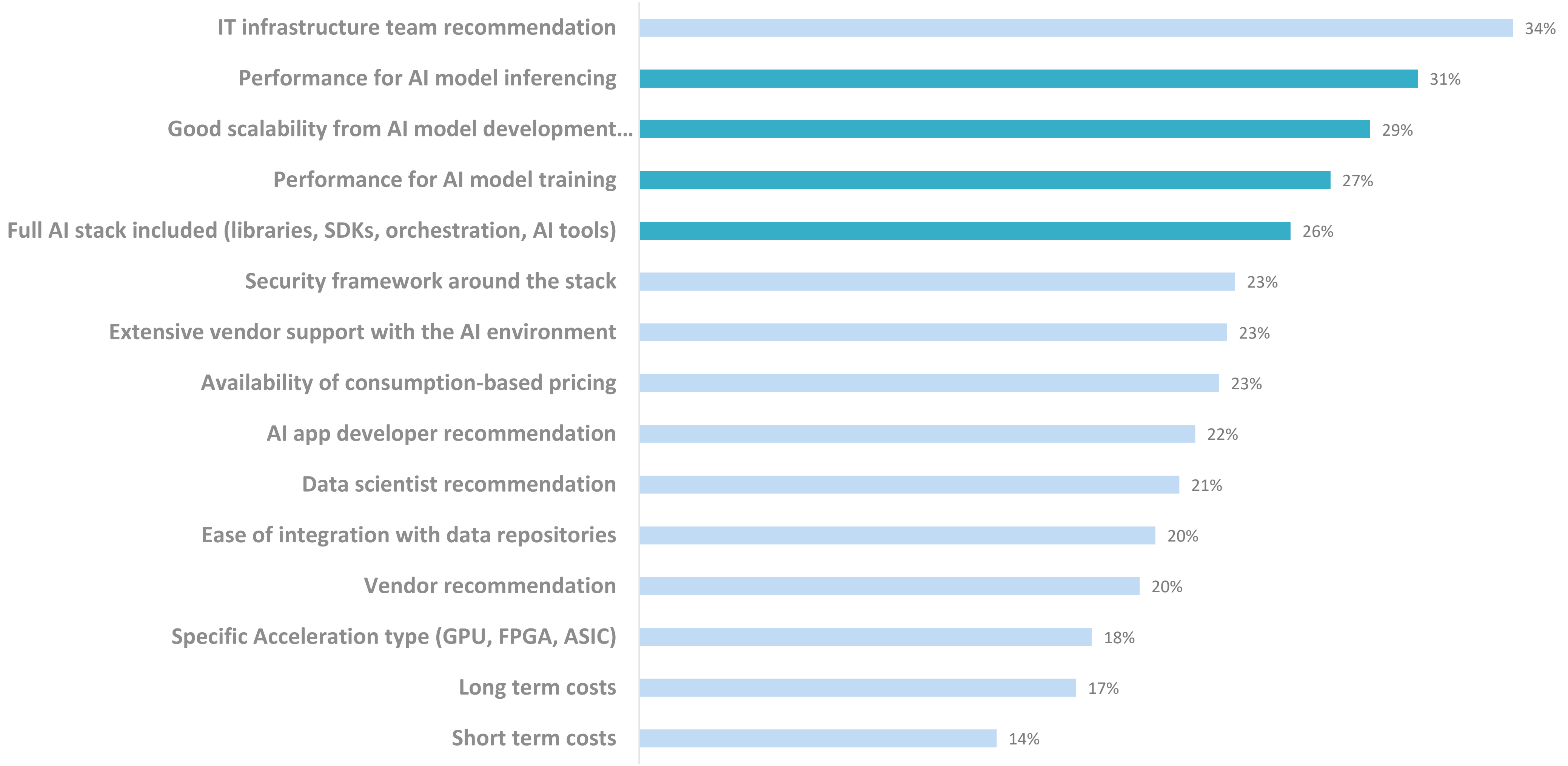
High performance
Compute and storage services optimized for HPC and AI

Edge computing
Extending cloud services to smaller, remote locations

Multicloud integration
Tools and commercial terms to facilitate management and security

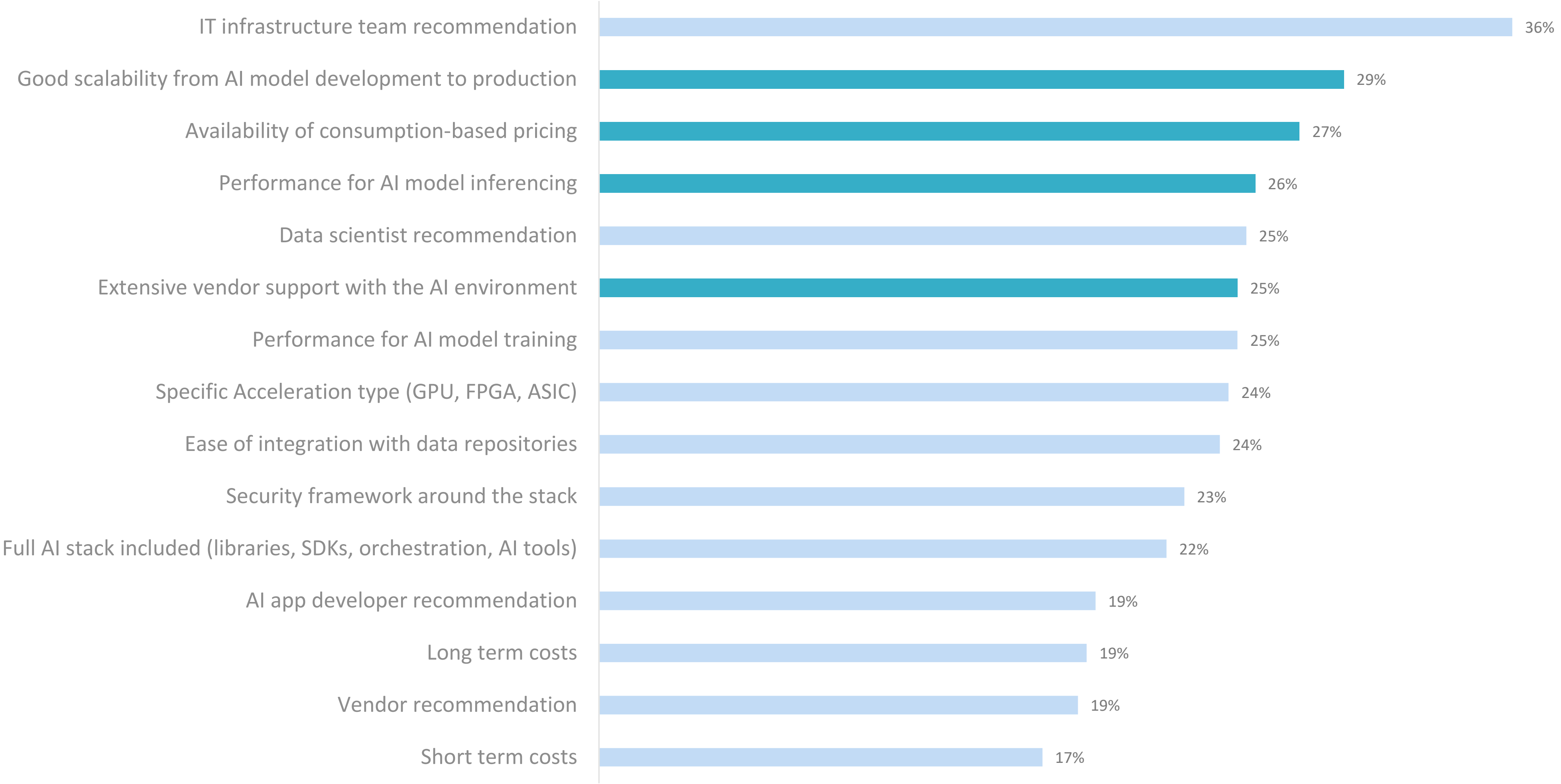
Deeper partnerships
Creating bridges with traditional enterprise OEMs and ISVs.

Key requirements when selecting compute infrastructure for AI in datacenters, at colocation providers and edge locations



Source: 2023 AI View, September 2023, IDC; N=400

Key requirements when selecting storage infrastructure for AI in datacenters, at colocation providers and edge locations



Source: 2023 AI View, September 2023, IDC; N=400

A typical Software Stack

Storage File System	🟦🟧▩ BPFS	🟦🟧▩ CephFS	🟦🟧▩ CHFS	DAOS	🟦🟧▩ Dell Poweredge	🟩 DDN GPFS/Lustre
	🟦🟧▩ Dell Powerstore	🟦🟧▩ FlashFS	🟦🟧▩ GekkoFS	🟦🟧▩ IME	🟦🟧▩ IBM Spectrum Scale	🟦🟧▩ Kapok
	🟦🟧▩ MadFS	🟦🟧 NetApp	🟦🟧▩ OceanFS	🟦 Panasas	🟦🟧▩ Pure Storage	🟦🟧▩ Quobyte
	🟦🟧▩ Vast	🟦🟧 WekaIO	🟦🟧▩ Qumulo	🟩 Accelero		
Schedulers, Workload Management & Orchestration	🟦🟧▩ Adaptive Moab Cluster Suite	🟦🟧▩ Altair PBS Pro	🟩 Altair Grid Engine (SGE)	Containers: Docker, Singularity	Oracle Grid Engine (SGE)	🟦🟧▩ Runai
	🟦🟧▩ Spectrum LSF	🟦🟧▩ SLURM				
System Management	🟦🟧▩ Ansible	🟦🟧 Apache Mesos	Aspen Cluster Management	Bright Cluster Manager®	Cycle Computing	Collectd
	🟦🟧▩ Ganglia	🟦🟧 Grafana	HPE Performance Cluster Manager	🟩 OpenHPC	🟦🟧▩ OneSIS	🟦🟧▩ Puppet Enterprise
	🟦🟧▩ Prometheus	🟩 Scyld Clusterware	🟩 Xcat	Zabbix		
Remote Visualization	NICE DCV, EnginFrame	Infovision VCollab	EXceed			
Network Fabric	🟩 Mellanox® Unified Fabric Manager™	Ethernet 10/20/40/100GigE	HPE Slingshot Ethernet	Nvidia NVlink	Rockport Network	OSU INAM
Storage Benchmark	🟦🟧 FIO	Iozone	IOR	Iometer	Cross Platform Disk	Vdbench
Operating System	🟦🟧 RedHat Enterprise	⬛ Centos/Rocky Linux	🟦🟧▩ FreeBSD	🟦🟧▩ Ubuntu Enterprise	🟦🟧▩ Suse	

🟦 HPE Apollo, Cray, SGI 🟧 NVIDIA ▩ ARM 🟩 Groq 🟪 Cerebras 🟥 Graphcore

A typical AI Applications Stack

Application Software Development Ecosystem

Programming Environment

🟦🟧▨ AOCL	🟦🟧▨ Boost	🟦🟧▨ CMake	🟦🟧▨ Cuda	🟦🟧▨ DDT, MAP Performance Report	🟦🟧▨ Deal.ii
🟦🟧▨ FFTW	🟦🟧 GNU Compiler Collection	🟦🟧▨ Gurobi	🟩 Java, Perl. Python (numpy, scipy)	🟦🟧▨ Julia	🟦🟧▨ MKL
🟦🟧▨ Octave	🟦🟧▨ OpenBLAS	🟦🟧▨ OpenCV	🟦🟧▨ OpenBLAS	🟦🟧▨ R	🟦🟧▨ Tensorflow
🟦🟧 Tensorflow	🟦🟧▨ Theano	🟦🟧▨ oneAPI	🟦🟧▨ VTune	🟦🟧 PGI	🟦🟧 NAG

Parallel Programming Environment

🟦🟧 MVAPICH2	🟦🟧 MVAPICH2-Azure	MVAPICH2-X/AWS	MVAPICH2-GDR	MVAPICH2-Virt	MVAPICH2-EA
OMB	OpenMPI	Platform MPI	Intel MPI	HPE MPI	Cray MPI

Data Management

🟦🟧▨ Aspera	HDF5/PHDF5	Md5deep Hashdeep	Ncdu	OnDemand	Rclone	Spark	Visidata	Zoltan

Visualization

🟦🟧▨ Ascinema	🟦🟧▨ Circos	🟦🟧▨ OpenCV	Exceed	🟦🟧▨ Infovision Vcollab	🟦🟧▨ Lowchart
🟦🟧▨ Nice DCV Enginframe	Xfig Fig2dev	Visit			

Productivity

🟦🟧▨ Git	🟦🟧▨ GNU Parallel	Jupyter Notebook	Lazygit	🟦🟧▨ OnDemand	🟦🟧 Tmux	OEMT

AI/Deep Learning

🟦🟧▨ Anaconda	🟦🟧 Jupyter	🟦🟧 Caffe	🟦🟧 CUDA	🟦🟧 CuDNN	🟦🟧 DeepGraph
🟦🟧 Deepstream	🟦🟧 Gensynth	🟦🟧 Keras	🟩 MXNet	🟦🟧 PaddlePaddle	🟦🟧 Pytorch
🟦🟧 Runai	🟦🟧 Singularity	🟦🟧 Scikit-learn	🟦🟧 TensorRT	🟦🟧 Tensorflow	🟦🟧 Theano

🟦 HPE Apollo, Cray, SGI 🟧 NVIDIA ▨ Intel 🟨 AMD 🟩 ARM

Partnering for success

Urgency to respond quickly to business disruption at a corporate level is influencing partner selection

Strategic Generative AI technology partners in the next 18 months

41% Cloud Providers

aws

Google Cloud Platform

Alibaba Cloud

Azure

IBM Cloud

33% IT Consulting Providers

accenture

Deloitte.

pwc

30% GenAI Tool Providers

OpenAI

ANTHROP\C

AI21labs

co:here

Source: GenAI ARC Survey, IDC, August 2023, WW N=1,363 Data weighted by IT spend

Investing in the right partner when building an AI infrastructure stack

Hardware

Silicon

intel, NVIDIA, AMD, AMPERE, Google, aws, IBM

GRAPHCORE, SambaNova Systems, cerebras, SAMSUNG, Qualcomm

Compute

DELL Technologies, Hewlett Packard Enterprise, Lenovo, inspur 浪潮

SUPERMICRO, cisco, FUJITSU, NEC, HUAWEI

Storage

NetApp, PURESTORAGE, DELL Technologies, Hewlett Packard Enterprise

IBM, HITACHI Inspire the Next, FUJITSU, HUAWEI, inspur 浪潮

Networking

ERICSSON, NOKIA, cisco, JUNIPER NETWORKS, HUAWEI, ciena

Netcracker, Adtran, ZTE, NEC, FUJITSU, Hewlett Packard Enterprise

Power and Cooling

Schneider Electric, EATON, VERTIV, submer, EMERSON

ICEOTOPE, ABB, SIEMENS, MITSUBISHI, liquidstack

Software

Security

ARTHUR

AIShield Powered by Bosch

CRANIUM

ROBUST INTELLIGENCE

MINDGARD

Data Logistics and Protection

Deasie, TONIC, BROADCOM, PROTOPIA, VeraData The Dense Science Company, VERITAS, COMMVAULT, veeam, COHESITY

Software Defined Compute and Storage

Red Hat, vmware, SUSE, Microsoft, cisco, IBM, BROADCOM, NUTANIX, VAST

Systems Management

cisco, bmc, BROADCOM, Microsoft, splunk>, vmware, IBM, HITACHI Inspire the Next, Red Hat, servicenow, amdocs

Services

Cloud Infrastructure

aws, Microsoft, Google, ORACLE, Paperspace

Alibaba.com, Tencent Cloud, IBM, CoreWeave, Lambda

Colocation

EQUINIX, DIGITAL REALTY, NTT, QTS, CyrusOne

Cyxtera, FLEXENTIAL, switch, cologix

Connectivity

NTT, AT&T, verizon, vodafone, T, orange, BT, Telefónica

TATA COMMUNICATIONS, Telstra, américa móvil, etisalat e&, COMCAST, sky

Communications Platforms

twilio, RingCentral, VONAGE, sinch, infobip, Microsoft, zoom

webex by cisco, gupshup, 8x8, tanla, bandwidth, Alibaba.com

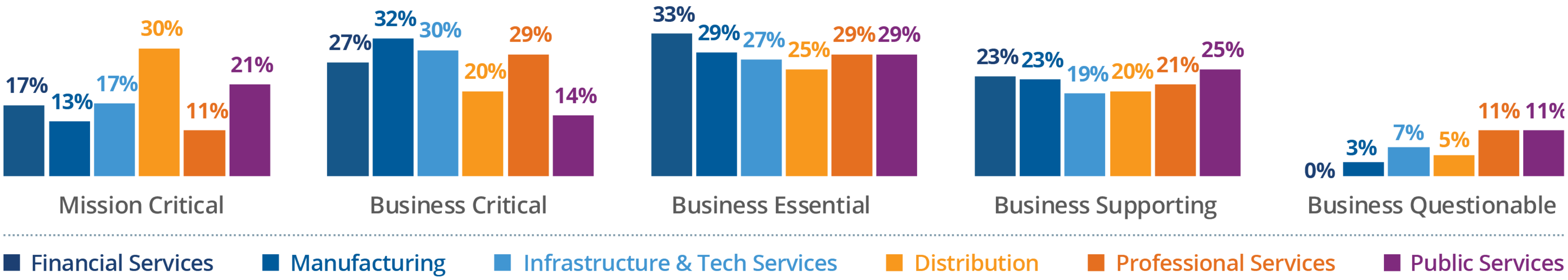
Source: IDC, 4Q23

For areas on which IDC publishes market share data, the top 3-5 market share leaders are represented. For areas on which IDC does not publish market share data, vendor selection is up to analyst discretion.

Consider using AI software services

Pre-trained standalone services that provide capabilities based on machine learning, deep learning, and other AI/ML technologies for applications and workflows to help improve business outcomes.
Used to build AI-powered applications.

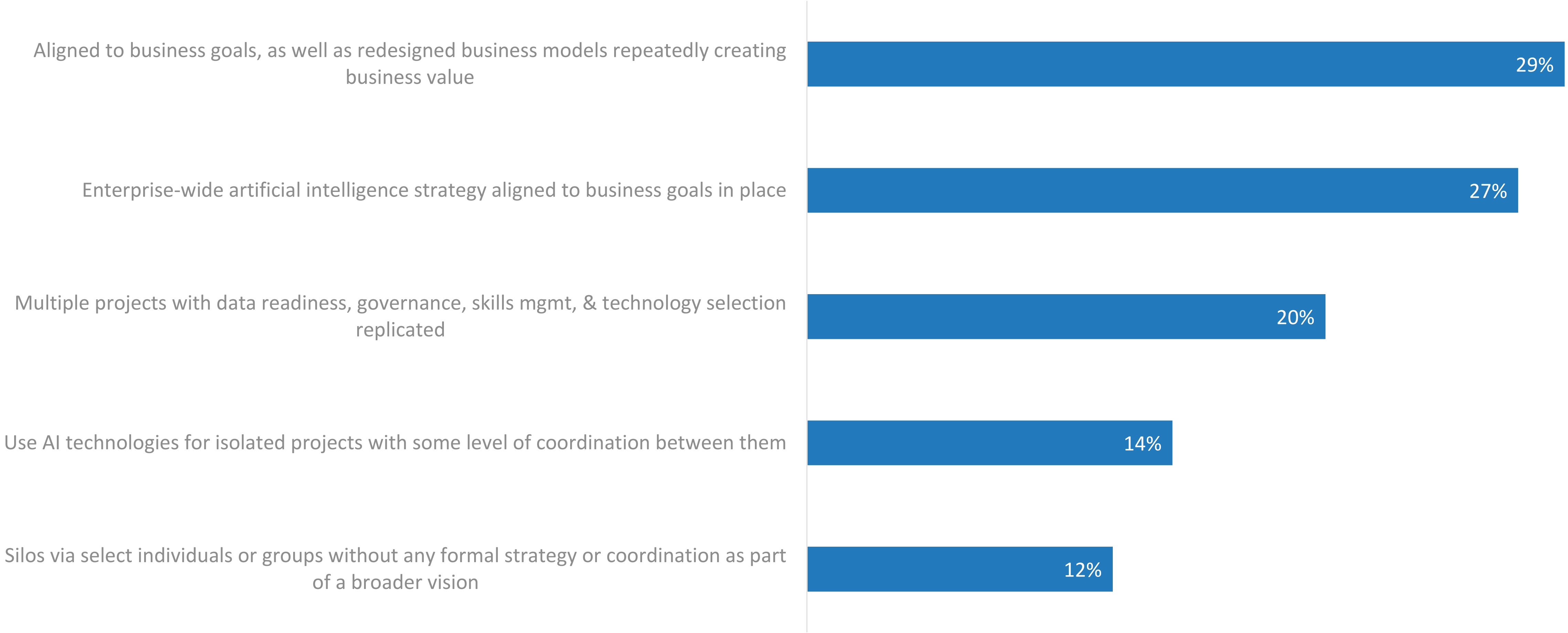
How would you describe AI Software Services in relation to the application's impact on your business today?



N = 277 Source: QD12 2Q23 Cloud Pulse Survey, August 2023 IDC

In Closing...

Consider an AI Center for Excellence to accelerate maturity of AI adoption



Source: 2023 AI View, September 2023, IDC; N=400

Thank you!

Ashish Nadkarni
ANadkarni@idc.com