IDC Directions

Delivering agentic services requires inference at scale

Rick Villars
Group Vice President, Worldwide Research

The AI economy in 2030

Global impact

$22.3 Trillion

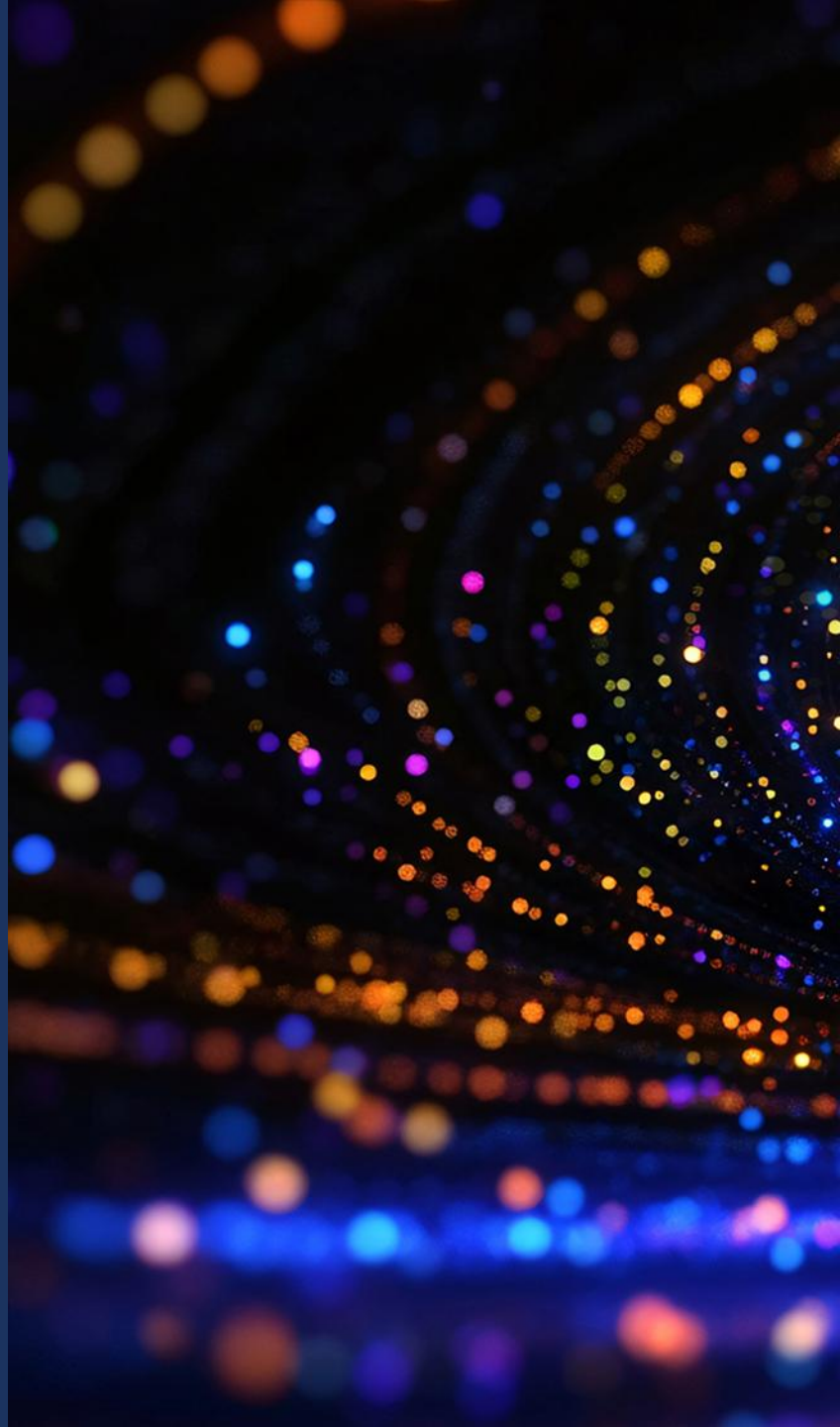
3.7% of GDP

Service providers account for 75% of infrastructure spend in 2025 with growing focus on agentic workloads

AI agents are biggest drivers of software & services growth as well as biggest business disruptors

Source: IDC's Macroeconomic Center of Excellence, March 2025

Completing the agentic transformation



Challenges in productizing agentic AI

Despite their promise, AI agents currently present a few challenges:

Slower performance

Due to the complexity of operations, AI agents can be slower.

Higher costs

Running AI agents is expensive, whether due to per-token charges from third-party providers or the need for powerful GPU clusters. amplify the situation.

Increased risk of errors

The added complexity (distributed systems consistency) introduces a higher likelihood of mistakes, requiring robust validation and monitoring.

Lack of skills

Not enough AI expertise and aging population creates a mismatch between demand and supply of talent. Job displacement concerns amplify the situation.

IDC

Challenges in productizing agentic AI

Despite their promise, AI agents currently present a few challenges:

Slower performance

Due to the complexity of operations, AI agents can be slower.

Higher costs

Running AI agents is expensive, whether due to per-token charges from third-party providers or the need for powerful GPU clusters. amplify the situation.

IDC

The inference delivery challenge

Agents = delivering the right answer

...at the right time...in the right context

Inferences required per answer delivered

How many inferences? Agents & apps

Timing and Distribution of Adoption in Applications

	2025	2027	2029	2031	2033	2035	2037
Agents as Apps	0%	0%	5%	20%	35%	50%	65%
Agent Led	0%	5%	20%	40%	50%	35%	25%
Ag-Enhanced	3%	30%	40%	25%	10%	10%	5%
As/Ad-Enhanced, Ag Supplemented	17%	35%	20%	10%	5%	5%	5%
As/Ad-Enhanced	50%	20%	10%	5%	0%	0%	0%
Traditional	30%	10%	5%	0%	0%	0%	0%

Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)

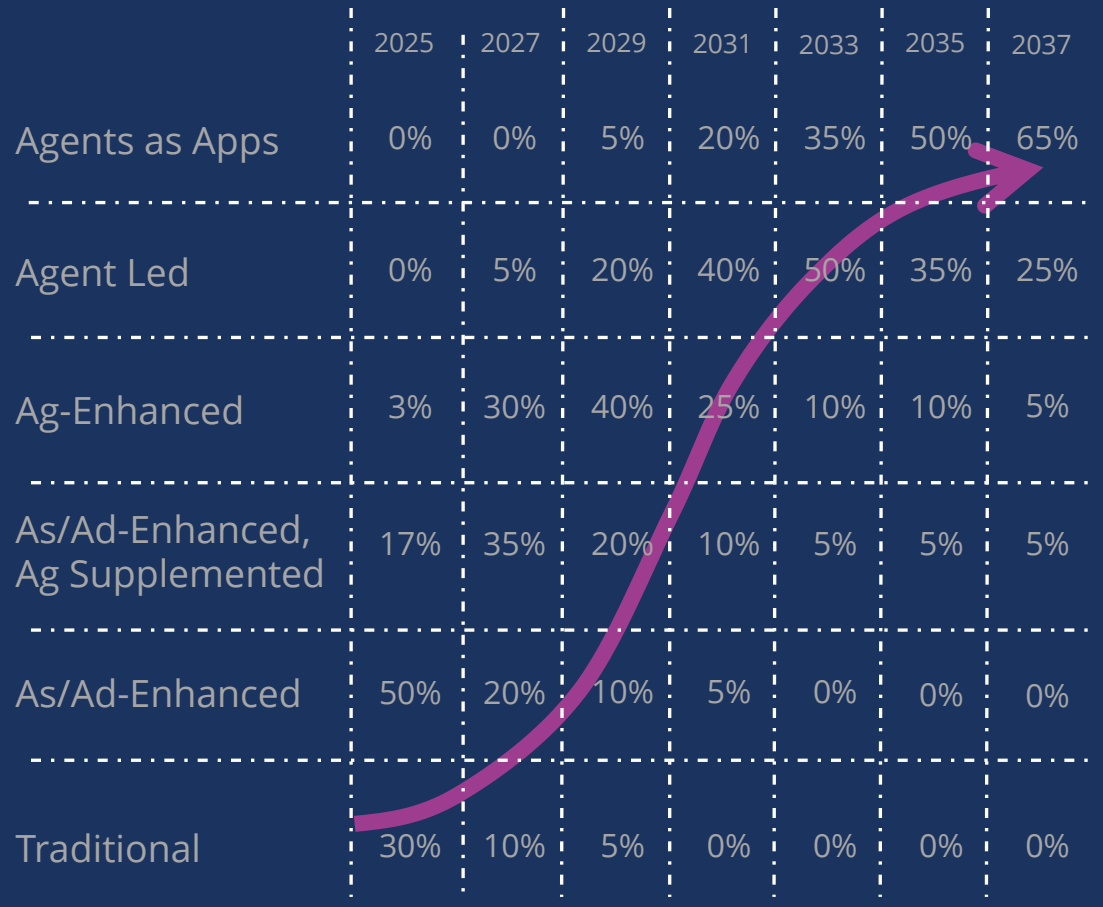
47%

AI contribution to total software value in 2029

Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

Inference load assumptions

Timing and Distribution of Adoption in Applications



	2025	2027	2029	2031	2033	2035	2037
Agents as Apps	0%	0%	5%	20%	35%	50%	65%
Agent Led	0%	5%	20%	40%	50%	35%	25%
Ag-Enhanced	3%	30%	40%	25%	10%	10%	5%
As/Ad-Enhanced, Ag Supplemented	17%	35%	20%	10%	5%	5%	5%
As/Ad-Enhanced	50%	20%	10%	5%	0%	0%	0%
Traditional	30%	10%	5%	0%	0%	0%	0%

Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)

Inference types

- GenAI requests
- Other AI system requests
- Non-AI systems (API/query)

Inference modes

- Single mode (text)
- Single mode (media)
- Multi-mode

Inference frequency

- Single Query
- Prompt-stream
- Event-triggered
- Continuous monitoring

Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

Inference load : Agent-less

Agent-less

Timing and Distribution of Adoption Applications

	2025	2027	2029	2031	2033	2035	2037
Agents as Apps	0%	0%	5%	20%	35%	50%	65%
Agent Led	0%	5%	20%	40%	50%	35%	25%
Ag-Enhanced	3%	30%	40%	25%	10%	10%	5%
As/Ad-Enhanced, Ag Supplemented	17%	35%	20%	10%	5%	5%	5%
As/Ad-Enhanced	50%	20%	10%	5%	0%	0%	0%
Traditional	30%	10%	5%	0%	0%	0%	0%

Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)

Inference types

- GenAI requests
- Other AI system requests
- Non-AI systems (API/query)

Inference modes

- Single mode (text)
- Single mode (media)
- Multi-mode

Inference frequency

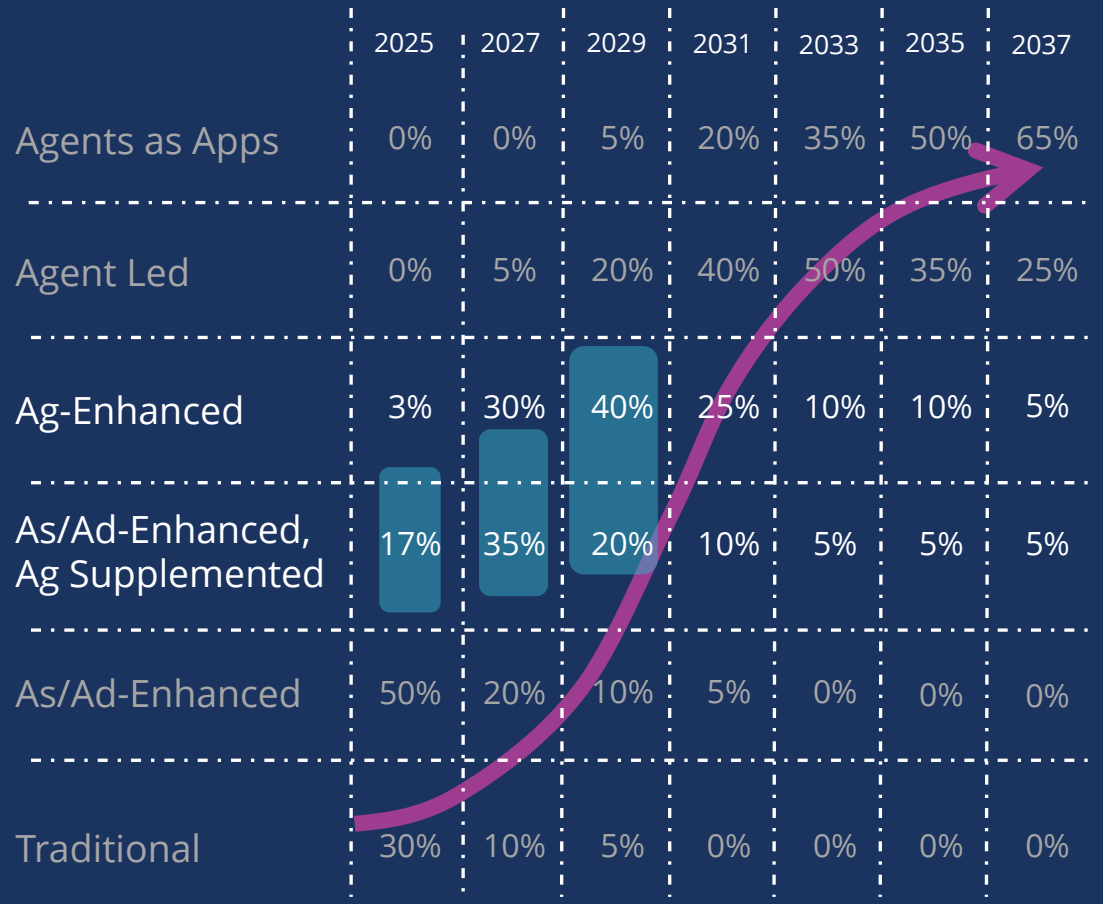
- Single Query
- Prompt-stream
- Event-triggered
- Continuous monitoring

Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

Inference load : Agent-enhanced

Agent-enhanced

Timing and Distribution of Adoption in Applications



Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)

Inference types

- GenAI requests
- Other AI system requests
- Non-AI systems (API/query)

Inference modes

- Single mode (text)
- Single mode (media)
- Multi-mode

Inference frequency

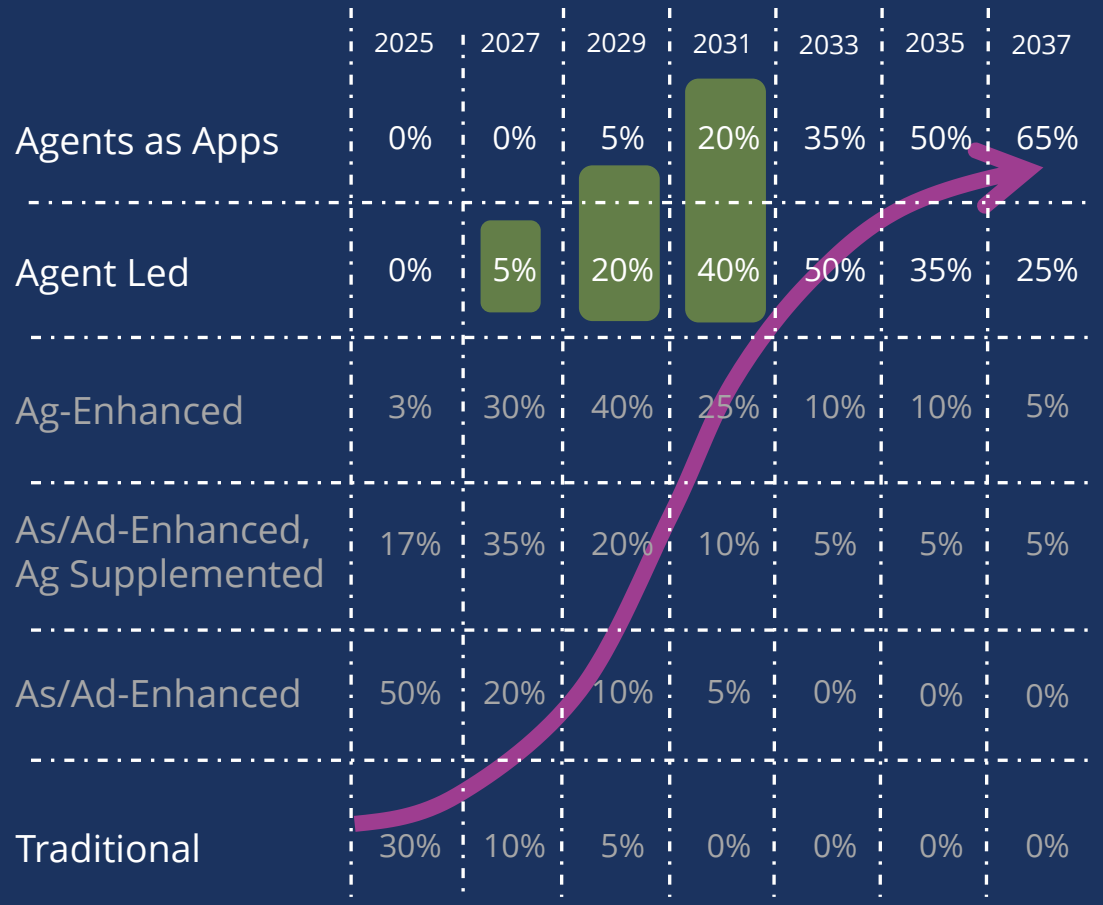
- Single Query
- Prompt-stream
- Event-triggered
- Continuous monitoring

Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

Inference load: Agent-based

Agent-based
(agent fleets)

Timing and Distribution of Adoption in Applications



Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)

Inference types

- GenAI requests
- Other AI system requests
- Non-AI systems (API/query)

Inference modes

- Single mode (text)
- Single mode (media)
- Multi-mode

Inference frequency

- Single Query
- Prompt-stream
- Event-triggered
- Continuous monitoring

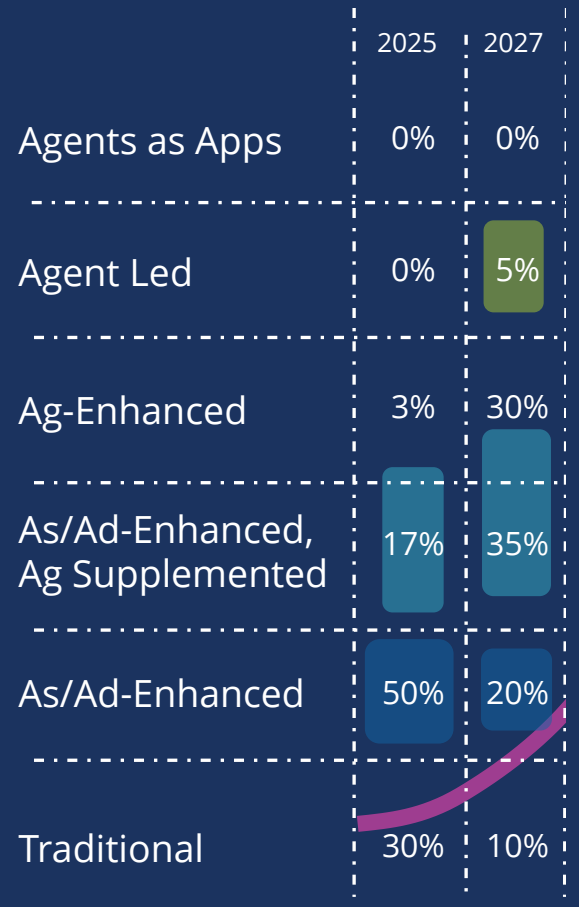
Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

The agentic transition has already started

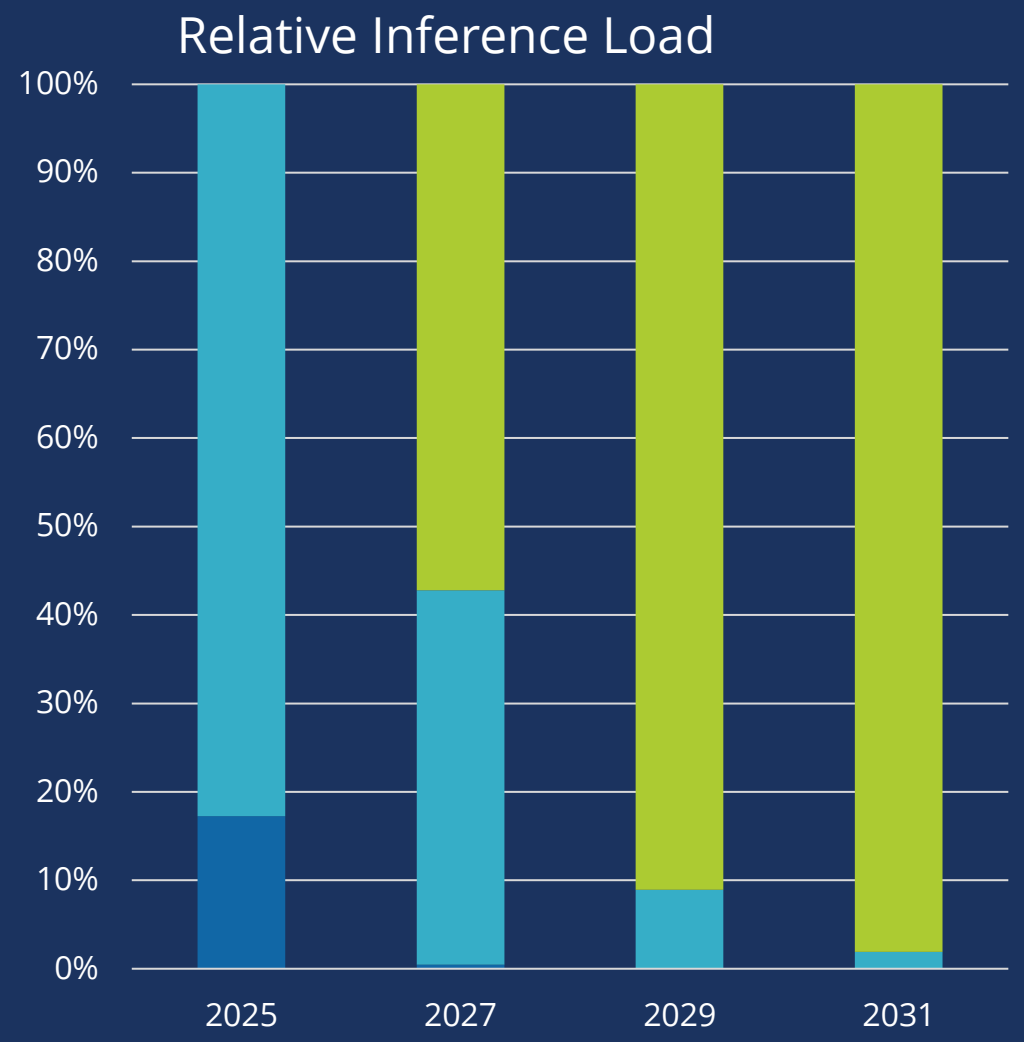
Agent-based
(agent fleets)

Agent-enhanced

Agent-less



Legend: Assistants (**As**), Advisors (**Ad**)

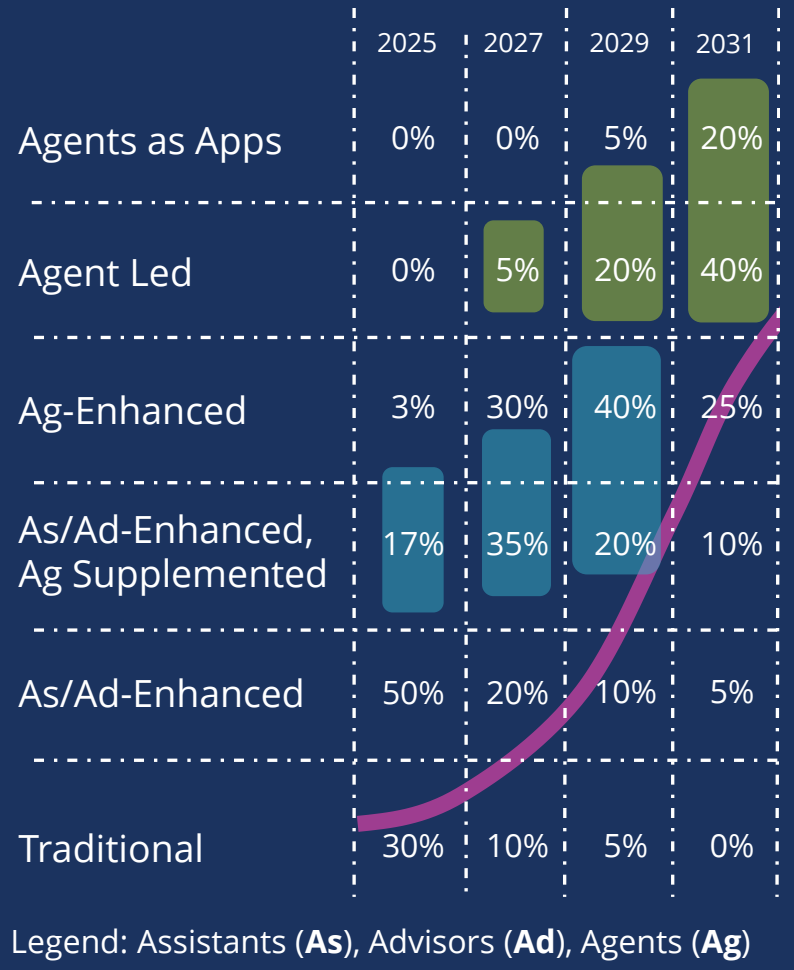


Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

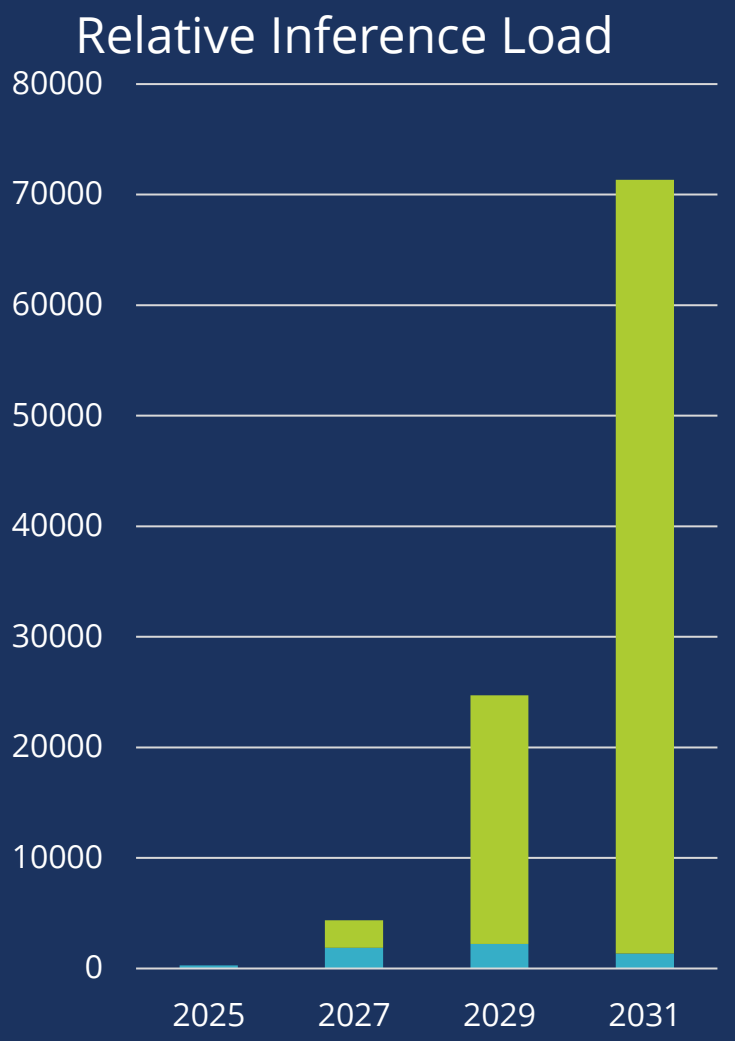
We haven’t seen anything yet

Agent-based
(agent fleets)

Agent-enhanced



Legend: Assistants (**As**), Advisors (**Ad**), Agents (**Ag**)



1/3rd

Agent contribution to AI SW value in 2029

Source: The Agentic Evolution of Enterprise Applications, February 2025, IDC #US53194625

Scaling the inference cliff

Inference loads will grow much faster than agentic AI contribution to software value

Reduce per inference response time and processing costs

Inference optimized hardware: Critical issues

- Focus of service providers on inference loads in current buildout
- Effective use of reduced precision floating-point solutions
- Trade-offs between centralized and distributed deployment

HPC and AI infrastructure index

Inference optimized software

- Boosting utilization and consistency for inference asset deployments
- Integrated platforms for agentic and inference resource management
- Extending distributed inferencing from clusters to dispersed edge

AI-ready infrastructure & data logistics

Essential guidance: The inference scale journey

Opportunistic
AI Pivot

The Agentic Bump (6-12 months)

- SW & Services Providers drive agent surge
- Cloud SP buildout and optimization at the core
- Data control & cost monitoring in enterprise

Repeatable
AI Alignment

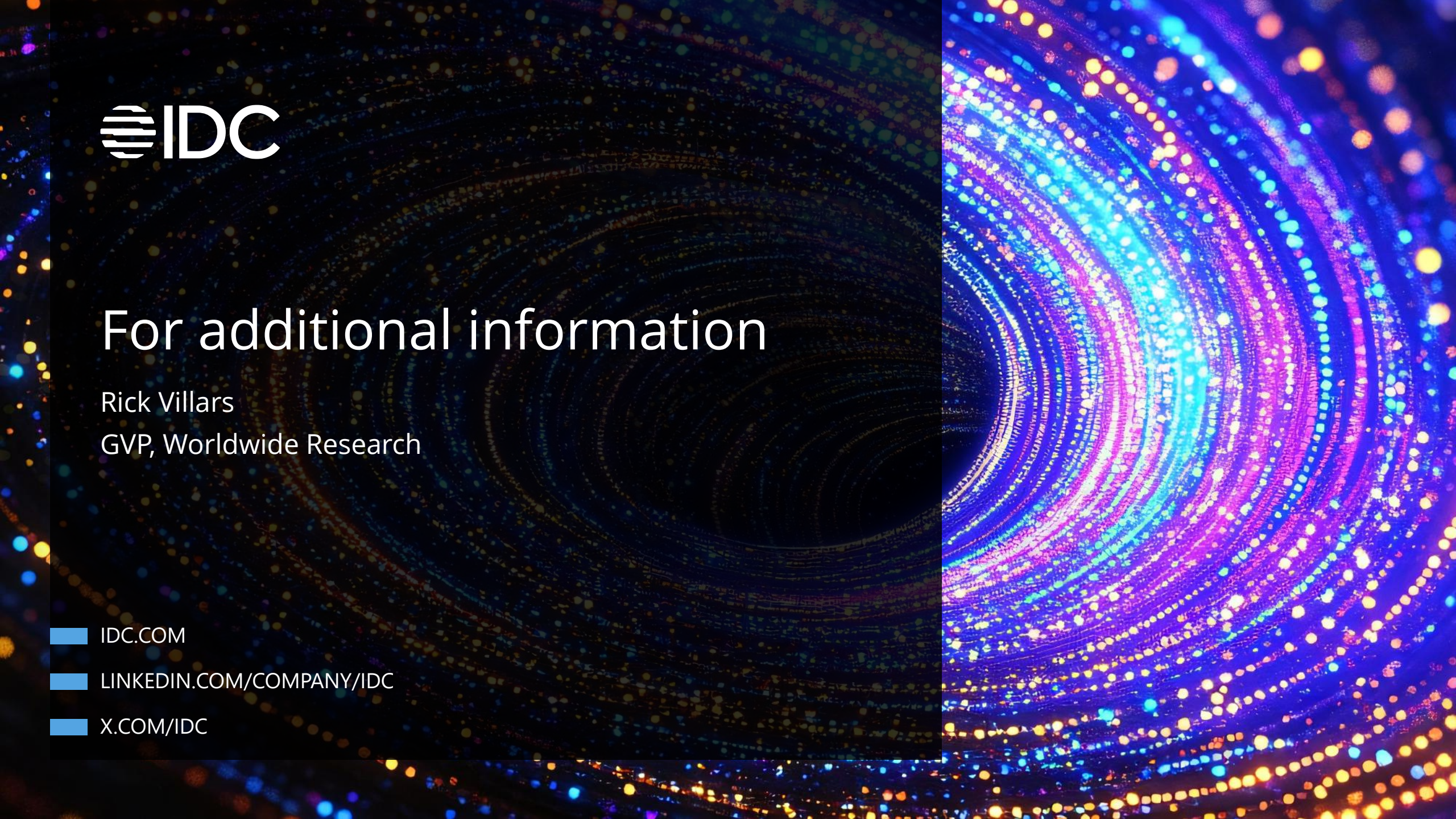
Agentic expansion (12–36 months)

- Agent building & interconnect accelerate
- Demand for standard inference platforms
- Major upgrades of core systems/apps

Managed
AI Transform

Agent-led (36+ months)

- Agentic fleets drive 2nd inference surge
- Demand for inference & answer caching options
- Preparing for physical/ agentic convergence



For additional information

Rick Villars
GVP, Worldwide Research

- IDC.COM
- LINKEDIN.COM/COMPANY/IDC
- X.COM/IDC


IDC Directions