InfoBites, sponsored by Supermicro and NVIDIA | JULY 2025

Powering AI Factories: Scaling GenAI with Direct-to-Chip Liquid-Cooling

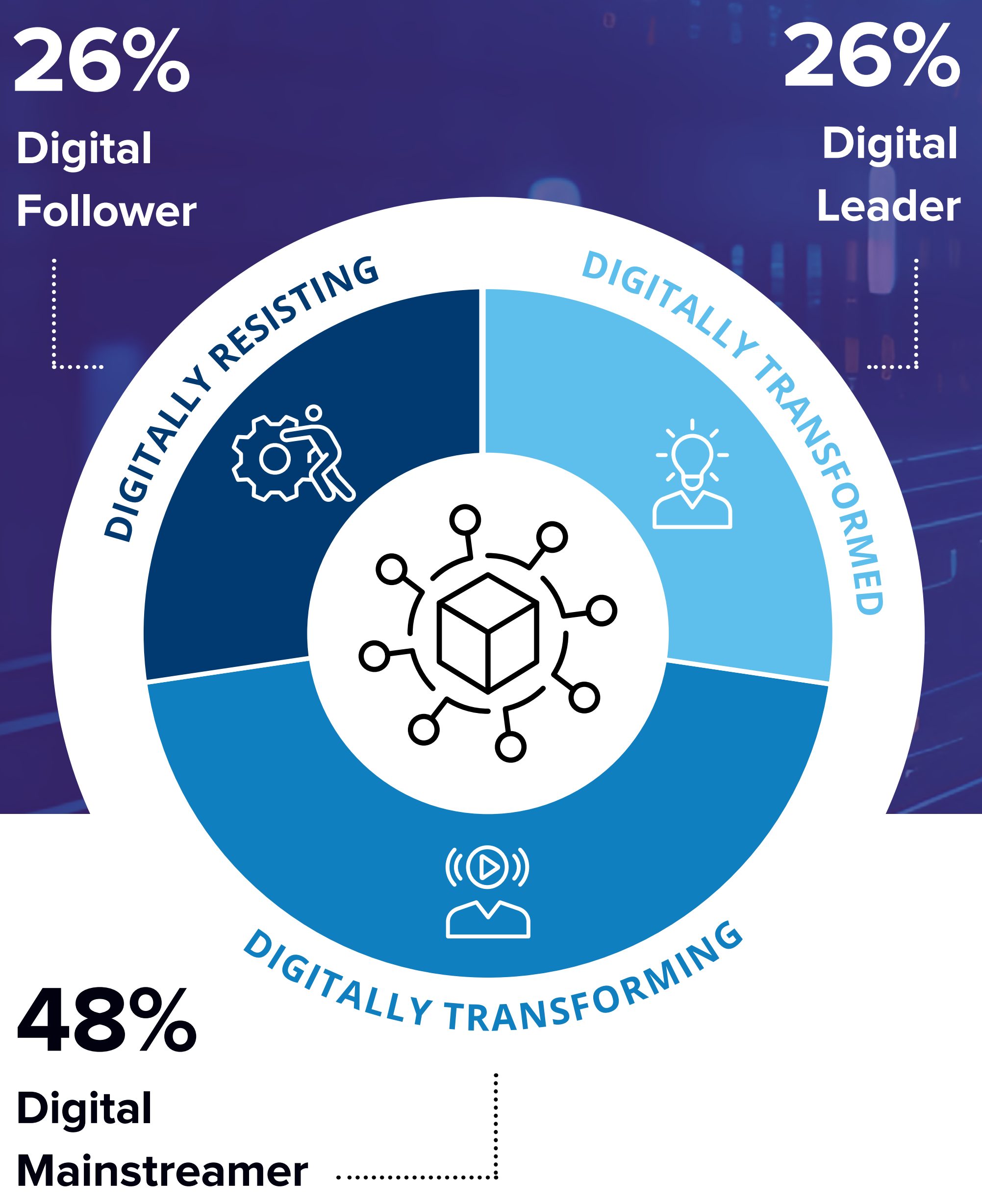
Transforming Datacenters from Cost Centers to Engines of Intelligence

Andrew Buss
IDC IT infrastructure EMEA

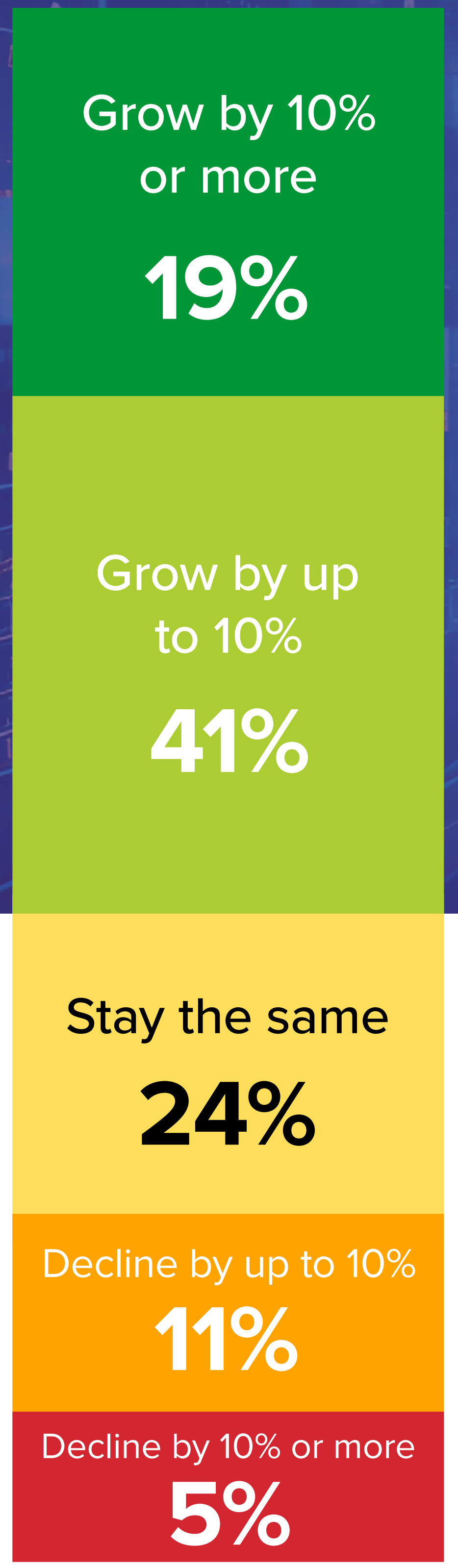
Luis Fernandes
IDC IT infrastructure EMEA

The fastest growing companies are investing in automation and data to drive GenAI readiness.

Digital Leaders are investing heavily in their IT infrastructure to drive **competitive advantage and differentiation** and to significantly outgrow the competition.



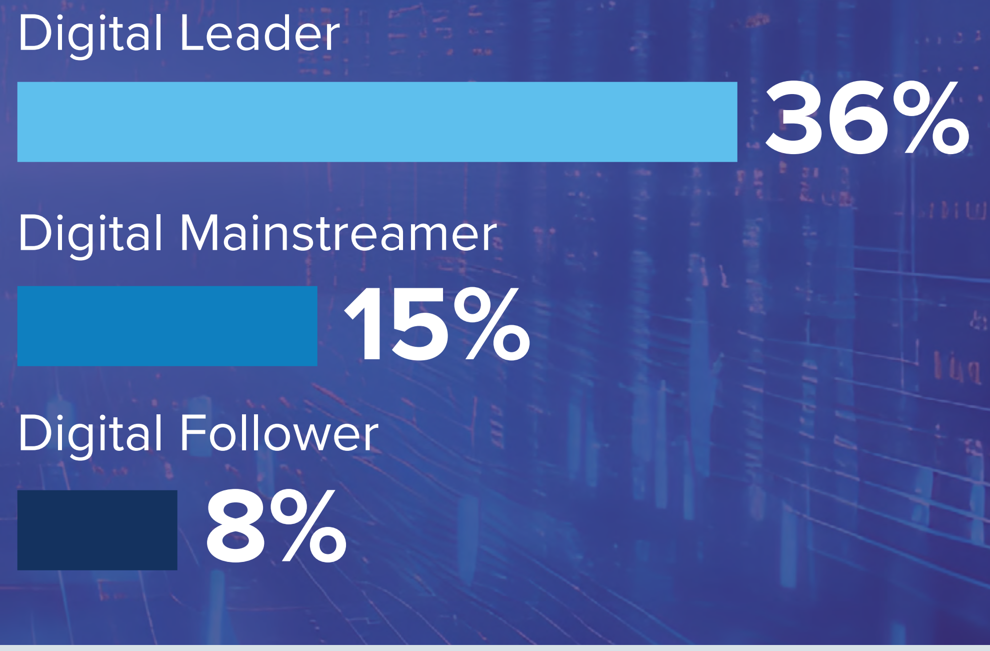
Companies in the mainstream of IT focus on using their IT investments and capabilities to become **more efficient**.



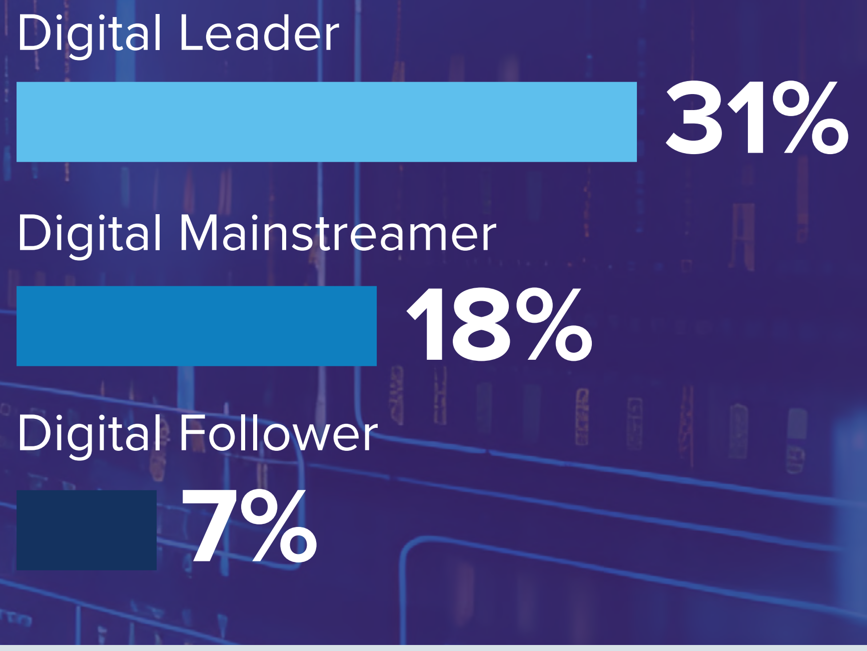
Company revenue growth
— latest financial year vs. prior year

Only a fifth of companies achieve the top **tier of revenue growth** of 10% of more

Proportion with **Revenue Growth** of 10% of more



Proportion with **IT Infrastructure Budget Growth** of 10% of more



Digital Leaders have extensively adopted a range of advanced infrastructure technologies to enable their success. This has helped them to build **AI factories** — datacenters that generate intelligence and revenue.

- AIOps
- Automation and orchestration
- Dynamic workload management
- API-centric development
- DevOps
- Containers and Kubernetes
- Digital trust
- Formal risk management
- Hyper-converged infrastructure
- Multi-cloud services
- Real-time data
- AI or GenAI in a business use case

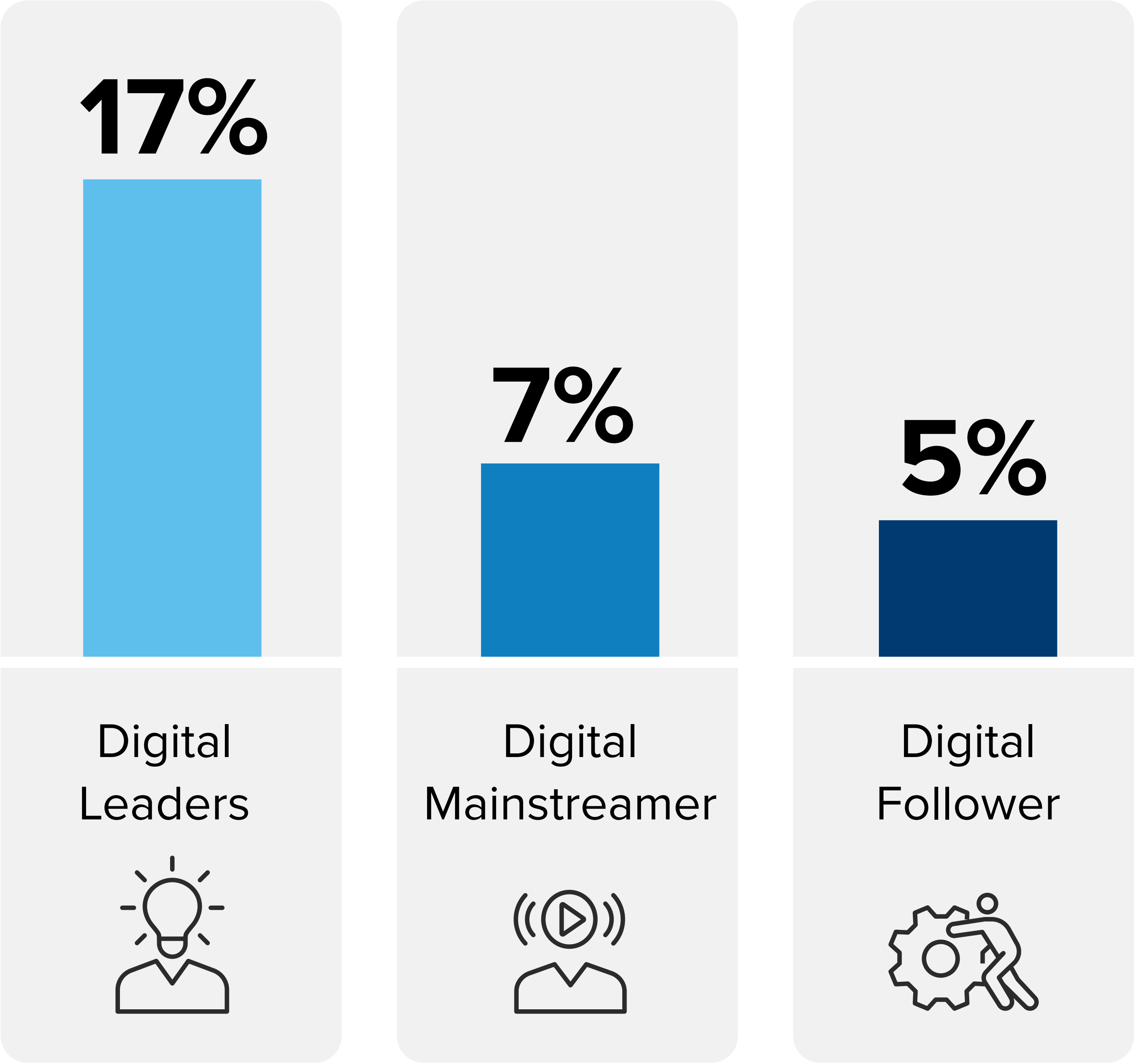
Digital leaders are building AI factories — datacenters that generate intelligence and revenue.

Source: IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590)

InfoBites, sponsored by SUPERMICRO NVIDIA.

AI infrastructure is no longer a backend cost; it is increasingly a front-end value engine.

GenAI is already having disruptive impacts in the digital economy.



Significant disruption seen to competitive position or business operating model **because of GenAI**

GenAI buildout has rapidly become a top IT priority for Digital Leaders.

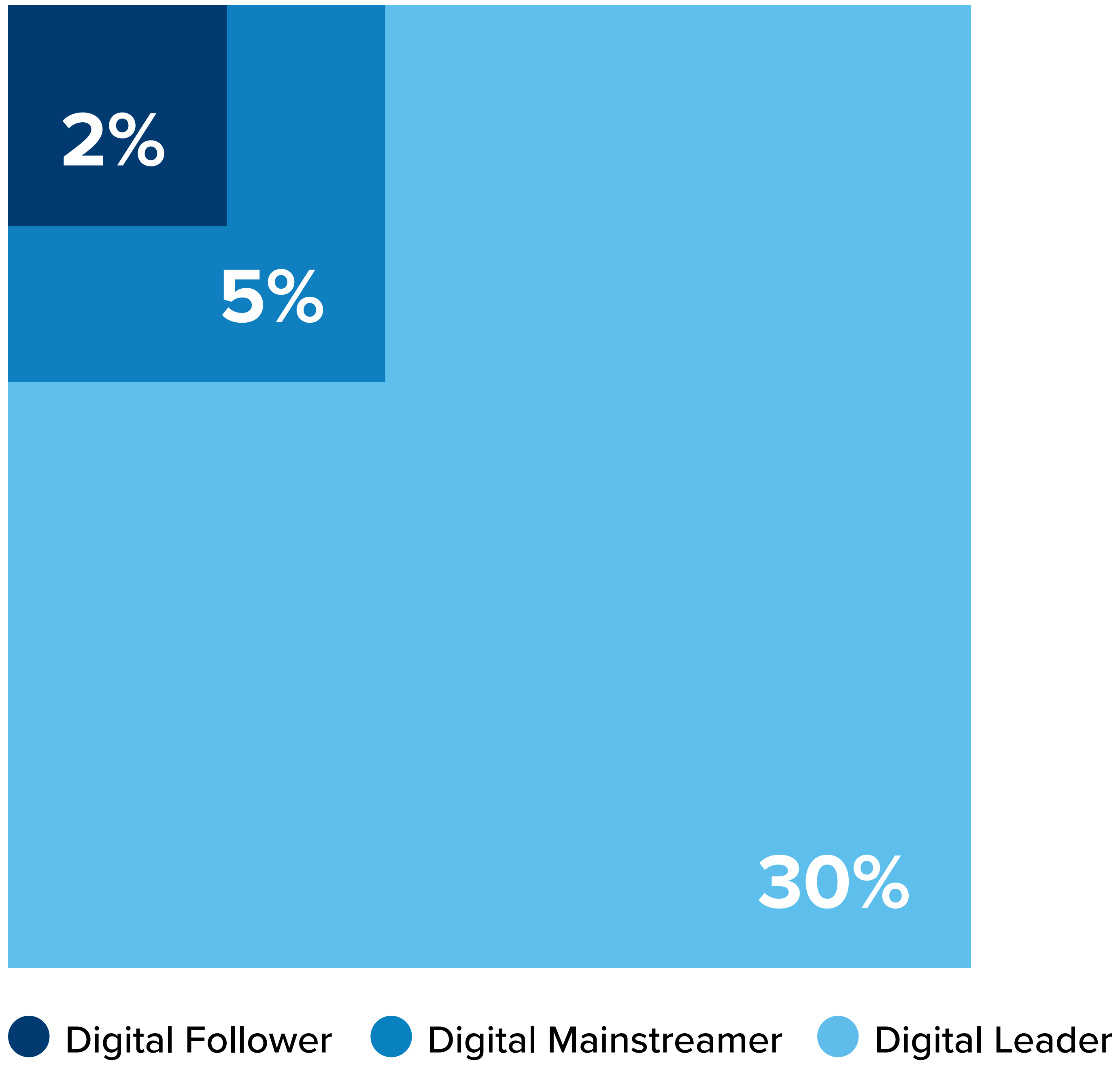
Top IT infrastructure Priorities for Digital Leaders for 2024

#1 Improve security and compliance

#2 **Build out GenAI infrastructure**

#3 Enhance business resilience

Digital Leaders have embraced GenAI extensively to drive better revenues and improve business efficiency.

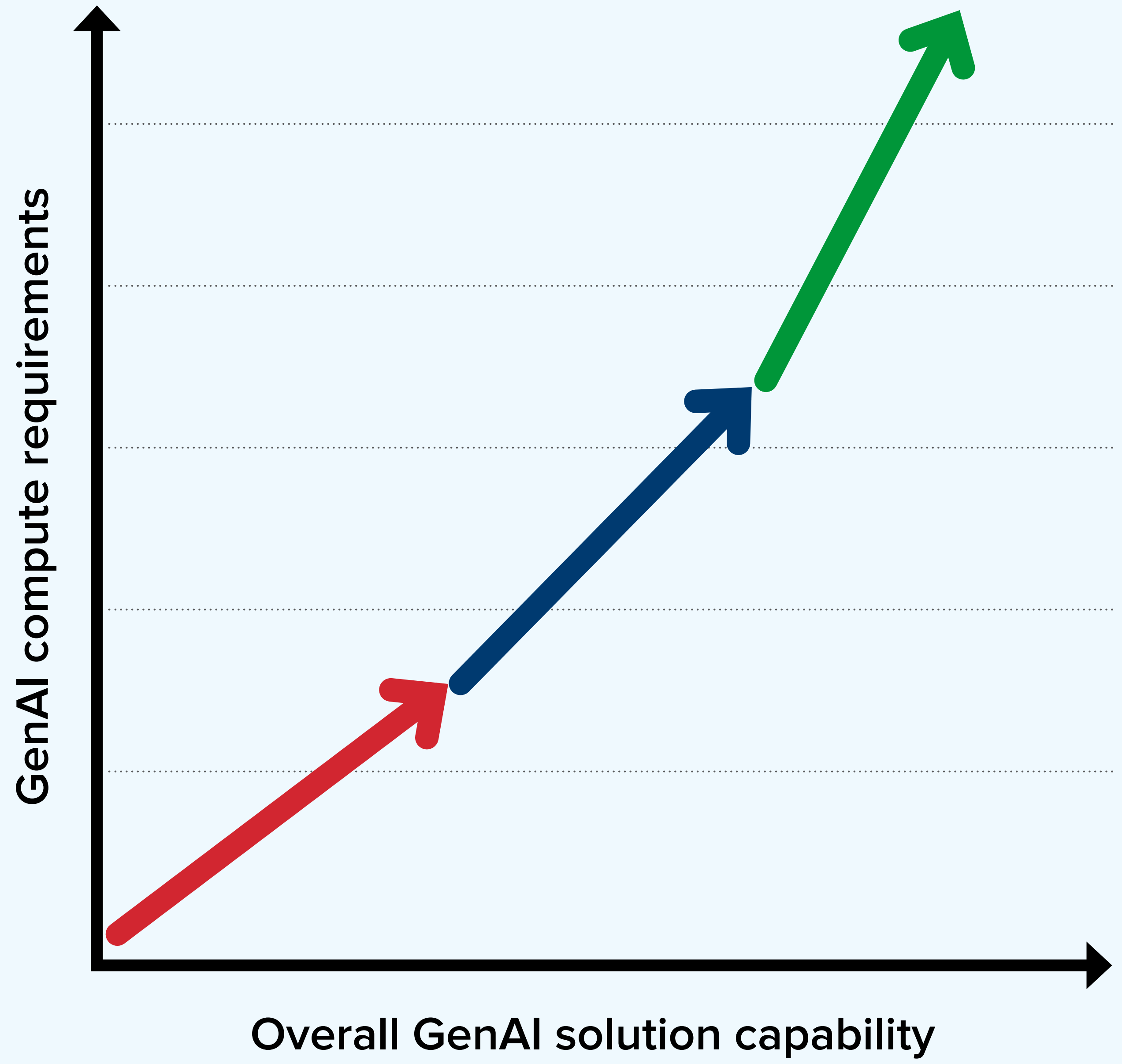


Extensive adoption of GenAI to support business use cases in production

Source: IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590)
Digital Leaders represent the top 26% most transformed IT organizations. Digital Mainstreamers are the middle 48% of transforming IT organizations.
Digital Followers represent the 26% least transformed IT organizations.

InfoBites, sponsored by SUPERMICRO NVIDIA.

Advances in GenAI reasoning will dramatically increase the need for accelerated computing.



Test-time scaling: Instead of processing a one-time answer, models allocate extra computational effort during inference, reasoning through multiple responses before arriving at an optimized answer.

Post-training scaling: The performance of a pretrained model can be improved by using techniques such as fine-tuning, distillation, pruning, quantization, reinforcement learning, and synthetic data augmentation.

Pre-training scaling: Increasing training dataset size, model parameter count, and computational resources results in predictable improvements in model intelligence and accuracy.

AI scaling laws are **driving exponential compute demand**. Establishing AI factories is essential in enabling this emerging model, in much the same way that foundational infrastructure was once required for the widespread adoption of electricity and the internet.

To support AI reasoning and agentic AI, test-time scaling can require up to 100 times more compute than standard inference. This is already having a dramatic impact with significant increases for power and cooling requirements for AI datacenters and infrastructure.

Source: IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590)

InfoBites, sponsored by SUPERMICRO NVIDIA

IDC

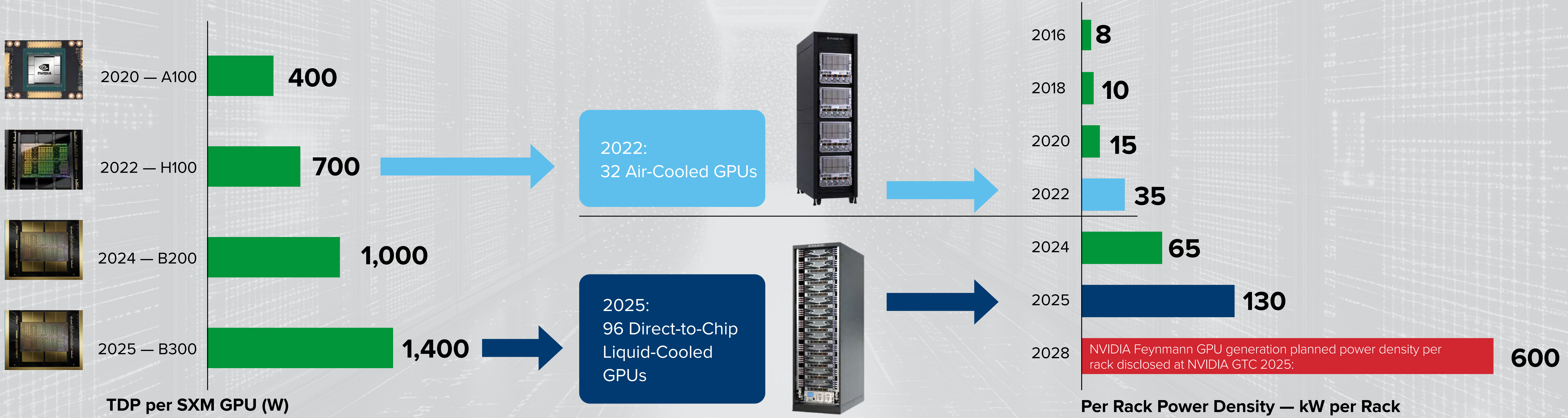
Direct-to-chip liquid-cooling enables the leap from traditional datacenters to high-throughput AI factories.

The performance needs of GenAI mean that rack scale systems are being engineered for maximum compute performance and density. Even with significant increases in performance per watt, increasing demand for scalable GenAI compute still means that overall GenAI compute power density is rising rapidly.

NVIDIA Datacenter GPUs TDP per SXM GPU (W)

Datacenter GPUs per Rack

kW per Rack for GPU-Accelerated GenAI Compute



2022:
32 Air-Cooled GPUs

2025:
96 Direct-to-Chip Liquid-Cooled GPUs

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

Datacenters are limited by the available power from the utility provider. That means that revenues are becoming power limited.
Power efficiency is a central metric for operational and revenue success; every watt not used for AI inference or training is lost revenue.

Sources: NVIDIA product specifications; NVIDA “GTC” keynote, March 2025; Wade Vincent’s “Data Center World” keynote, April 2025; IDC

InfoBites, sponsored by SUPERMICRO NVIDIA.

GenAI datacenters will need new approaches to maximize efficiency.

Key infrastructure-related challenges in building out GenAI infrastructure seen by Digital Leaders:[1]

35%
Deploying GenAI to remote datacenters for power or ESG reasons

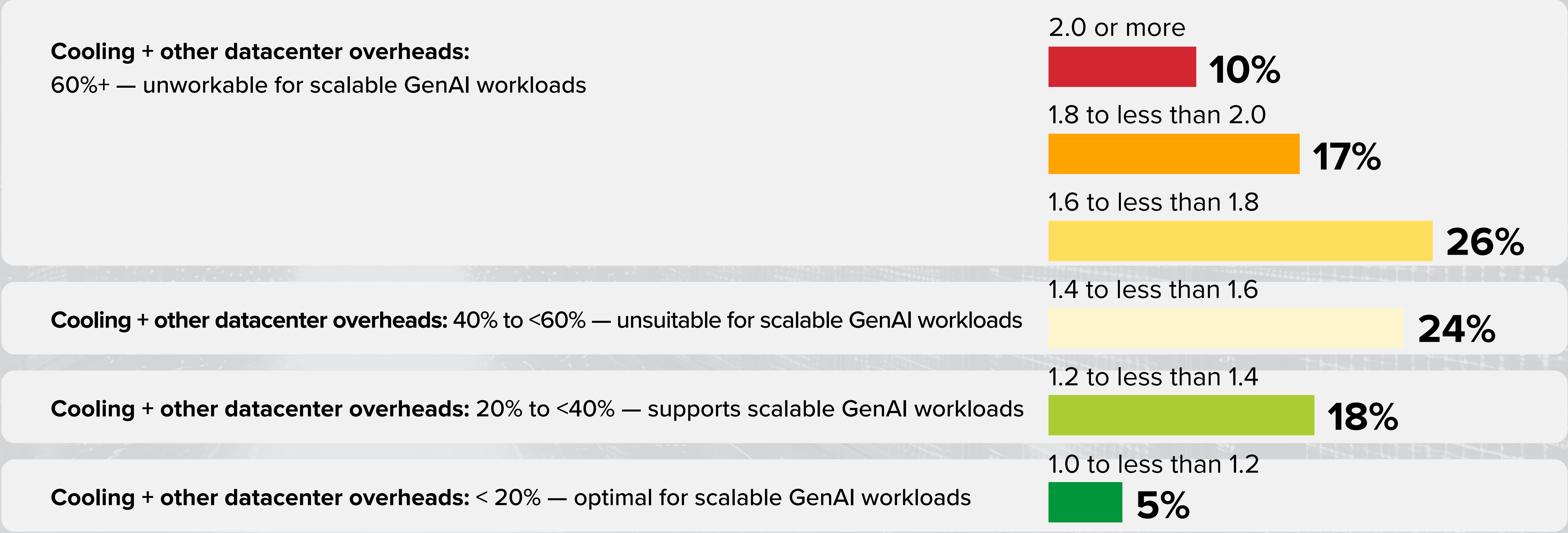
34%
Significantly increased power consumption within IT infrastructure

29%
The rising cost of power required to run GenAI workloads

17%
Not getting enough power to run GenAI workloads at the scale we need

The majority of air-cooled datacenters in operation today waste too much energy on cooling to run GenAI infrastructure effectively at scale.

The PUE of an Organization's Most Efficient Datacenter in 2023



Cooling + other datacenter overheads:
60%+ — unworkable for scalable GenAI workloads

Cooling + other datacenter overheads: 40% to <60% — unsuitable for scalable GenAI workloads

Cooling + other datacenter overheads: 20% to <40% — supports scalable GenAI workloads

Cooling + other datacenter overheads: < 20% — optimal for scalable GenAI workloads

To support **scalable and sustainable GenAI solutions**, we need to **maximize the energy directed toward powering GPUs and AI accelerators** while **minimizing the energy consumed by cooling, power distribution**, and other datacenter functions that constitute overhead rather than contributing direct value.

Sources: 1) IDC's *AI-Ready Infrastructure Survey, 2024*, EMEA (n = 1,590); 2) IDC's *Edge to Cloud Infrastructure Survey, 2023*, EMEA (n = 1,626)

InfoBites, sponsored by SUPERMICRO NVIDIA.

Direct-to-chip liquid-cooling is foundational to AI factory design. It supports 100kW+ racks and enables scalable and sustainable intelligence production.

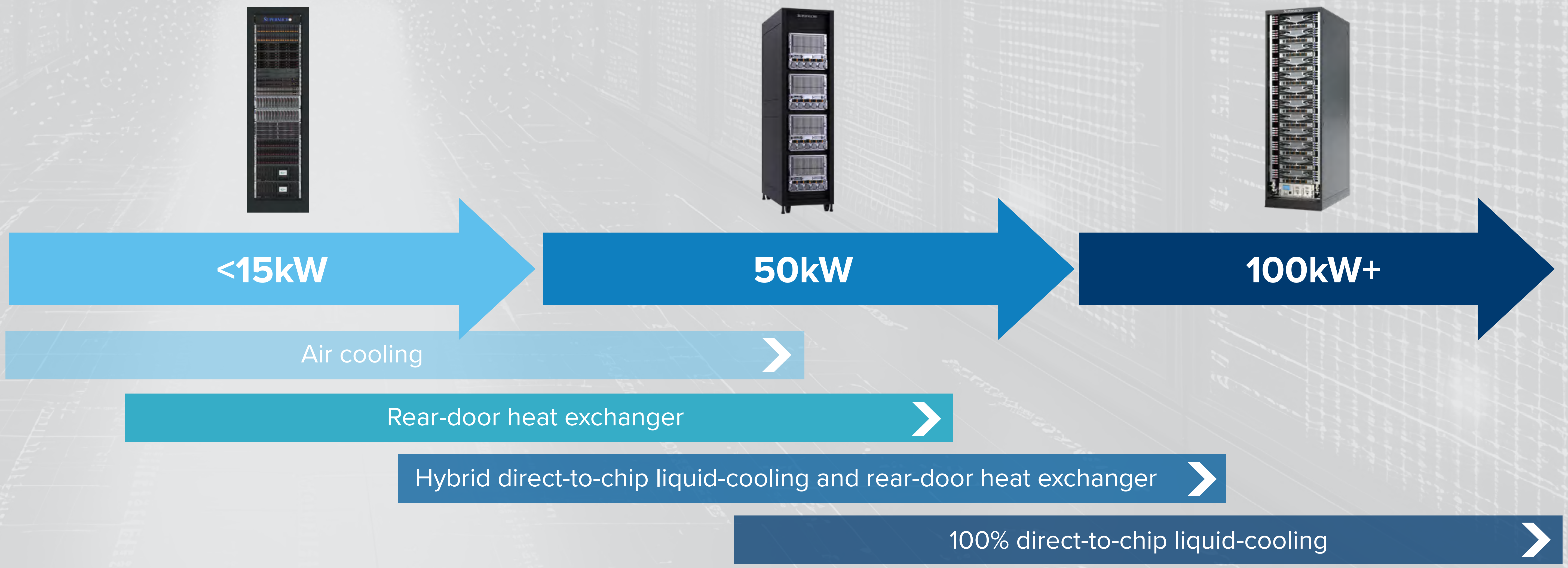
Elements of a direct-to-chip liquid-cooled datacenter that turn it into an AI factory:

- 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

The key benefits of direct-to-chip liquid-cooling:

- Dramatically lowers cooling overhead and enhances TCO
- Maximum system performance due to the high heat capacity of the technical cooling liquid
- More efficient datacenter use of water
- Helps meet ESG commitments and reporting requirements
- Supports significantly higher GenAI compute density
- Quieter operation, improving workplace comfort
- Prevents hotspots and thermal throttling
- Processors that run more consistently at peak “boost” clocks



<15kW
50kW
100kW+
Air cooling
Rear-door heat exchanger
Hybrid direct-to-chip liquid-cooling and rear-door heat exchanger
100% direct-to-chip liquid-cooling

InfoBites, sponsored by SUPERMICRO NVIDIA

Message from the sponsor

Supermicro and NVIDIA are redefining the economics of deploying AI factories. We offer state-of-the-art infrastructure solutions that address increased power and cooling challenges in modern AI datacenters. Additionally, significant savings can be achieved with direct liquid-cooling (DLC-2) for highly efficient generative AI datacenters.

For more information

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets.

With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.

Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.), the world's leading tech media, data, and marketing services company.

IDC Custom Solutions

This publication was produced by Custom Solutions. IDC's Custom Solutions group helps clients plan, market, sell, and succeed in the global marketplace. We create actionable market intelligence and influential content marketing programs that yield measurable results.

IDC

IDC UK
1st floor, Whitfield Street, London, W1T 2RE, United Kingdom
T 44.208.987.7100

@idc   @idc   idc.com