

Real-world Applications of Planning Patterns

Created by:

Eleni Verteouri

Gen AI Tech Lead @ UBS

Created & Narrated by:

Dipanjana Sarkar

Head of Community & Principal AI Scientist @ Analytics Vidhya

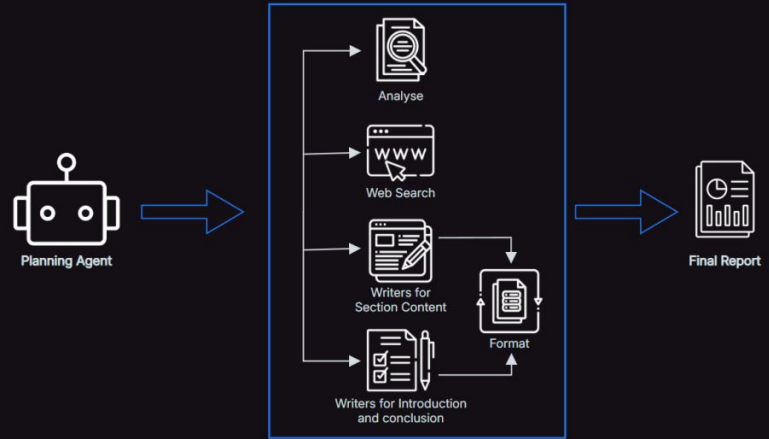
Google Developer Expert - ML & Cloud Champion Innovator

Published Author



Deep Research & Structured Report Generation

- It analyzes user-defined topics and creates a structured report outline.
- Parallel web searches fetch fresh content for each section.
- Parallel writers draft each section independently.
- Completed sections are formatted for consistency.
- Introduction and conclusion are generated based on the full content.
- Finally, everything is compiled into a polished final report.



What is Deep Research?

February 2, 2025 Release

Introducing deep research

An agent that uses reasoning to synthesize large amounts of online information and complete multi-step research tasks for you. Available to Pro users today, Plus and Team next.

Try on ChatGPT ↗

- **OpenAI** launched **Deep Research on February, 2025**, as an additional capability in their ChatGPT platform.
- Deep Research can do multi-step research on the internet for complex tasks or queries given by the user.
- They claim it can accomplish in minutes what would take a human several hours or days or even weeks.

Sample Use Case of Deep Research

What are you researching?

I need a detailed report on the current strategy Nvidia is following to win the majority markets in AI, do your own research and generate the report without asking me any followup questions



Search

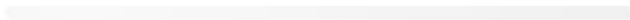


Deep research



Got it. I'll generate a detailed report on Nvidia's current strategy to dominate the AI market, covering key elements such as their product offerings, ecosystem, partnerships, and competitive moves. I'll get back to you shortly with a comprehensive breakdown.

Starting research



Output of the Use Case of Deep Research

Research completed in 8m · 21 sources

Nvidia's AI Market Domination Strategy (2025)

Introduction

Nvidia has emerged as the undisputed leader in artificial intelligence (AI) hardware, powering a majority of AI workloads worldwide. Various estimates indicate Nvidia holds between 80% to 95% of the AI accelerator market [patentpc.com](#) [athoday.com](#), thanks to a holistic strategy spanning cutting-edge hardware, a robust software ecosystem, deep industry partnerships, strategic acquisitions/investments, and expansion across global markets. This report provides a comprehensive analysis of Nvidia's market dominance in AI, covering its hardware strategy, software ecosystem, collaborations (with OpenAI, Google, etc.), competitive initiatives (cloud data centers, partnerships), and global expansion in emerging markets.

Hardware Strategy: GPUs, Data Center Products, and Custom Silicon

Nvidia's hardware strategy centers on relentless GPU innovation and specialized processors for AI:

- Flagship GPUs for AI:** Nvidia's latest data center GPUs like the A100 (2020, Ampere architecture) and H100 (2022, Hopper architecture) have become industry standards for training and deploying AI models. The H100 in particular introduced a Transformer Engine and FP8 precision, offering up to 9x faster training and 30x faster inference on large language models compared to its predecessor A100 [nvidiaresearch.com](#). These GPUs are highly sought after – an H100 unit can cost \$25,000–\$40,000 due to surging demand [patentpc.com](#). Nvidia is already pushing forward with next-generation GPUs (Blackwell architecture slated for 2024) to sustain performance leadership, with the upcoming “GB200” chip reportedly delivering up to 30x H100's performance on LLM inference at far lower energy/cost [investopedia.com](#).
- Grace Hopper “Superchip” (GH200):** To handle ever-larger AI models, Nvidia introduced the Grace Hopper (GH200) Superchip, which tightly couples a 72-core Arm-based Grace CPU with an H100 GPU via high-speed NVLink. This CPU-GPU fusion provides massive memory capacity and bandwidth – in dual-chip configuration, GH200 offers 144 CPU cores, 1.2 TB of unified memory, and 10 TB/s bandwidth, allowing generative AI models 3.5x larger to be trained/inferenced than previous systems [nvidiaresearch.com](#) [nvidiaresearch.com](#). The GH200 is optimized for giant-scale AI and HPC workloads, effectively a custom silicon platform for data centers in the AI era.
- Integrated AI Systems (DGX & HGX):** Nvidia packages its chips into full systems like the DGX series – ready-to-use AI supercomputers. For example, the DGX H100 integrates 8x H100 GPUs with high-speed NVLink and InfiniBand interconnects, delivering 32 petaFLOPs of AI compute in one node [nvidiaresearch.com](#). Such systems come with Nvidia's complete software stack, enabling enterprises to plug-and-play AI at scale [nvidiaresearch.com](#). Nvidia also provides the HGX server boards to OEMs, standardizing AI server designs. This vertical integration (from silicon to systems) ensures optimal performance and quicker adoption in data centers.

Software Ecosystem: CUDA, Frameworks, and Developer Platforms

A critical pillar of Nvidia's AI dominance is its expansive software ecosystem. CUDA, Nvidia's parallel computing platform introduced in 2006, is often cited as the *main reason* competitors struggle to catch up [patentpc.com](#). CUDA provides a common language (APIs and libraries) that lets developers harness GPU acceleration for general-purpose computing. Over years, Nvidia built a rich stack of AI software on top of CUDA:

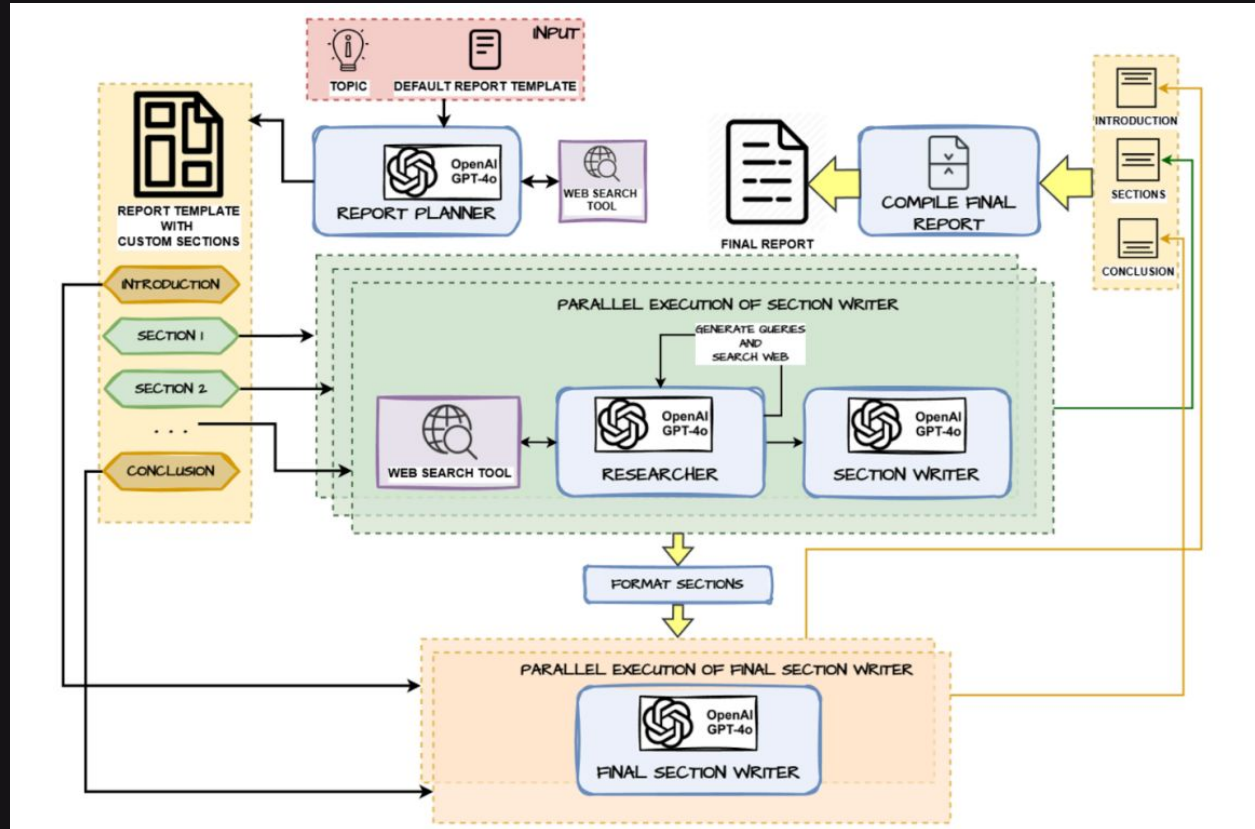
- CUDA and Developer Libraries:** The CUDA toolkit and drivers enable GPUs to accelerate popular programming languages (C++, Python via CUDA Python, etc.) [nvidiaresearch.com](#) [nvidiaresearch.com](#). CUDA Deep Neural Network (library) for training neural nets, TensorRT (for inference optimization), and countless other libraries (cuBLAS, NCCL for communication) ensure that AI models run efficiently on Nvidia hardware out-of-the-box, having invested millions of hours in CUDA code – forming a moat that competitors find platforms costly.
- AI Framework Integrations:** Nvidia works closely with major AI frameworks. PyTorch (via TorchVision or TensorFlow ops) use Nvidia's CUDA libraries under the hood. Docker containers (NGC) provide ready-to-run environments for machine learning deployment. The result is that data scientists can train models on Nvidia hardware whereas rival hardware often requires custom code or less mature ecosystems. The synergy between hardware and software has made Nvidia's platform the dominant force in AI development [patentpc.com](#).

Partnerships and Collaborations Driving Dominance

Nvidia amplifies its reach through strategic partnerships with leading AI players and tech companies. These collaborations ensure Nvidia's technology is deeply embedded in the AI infrastructure of cloud providers, research labs, and enterprises:

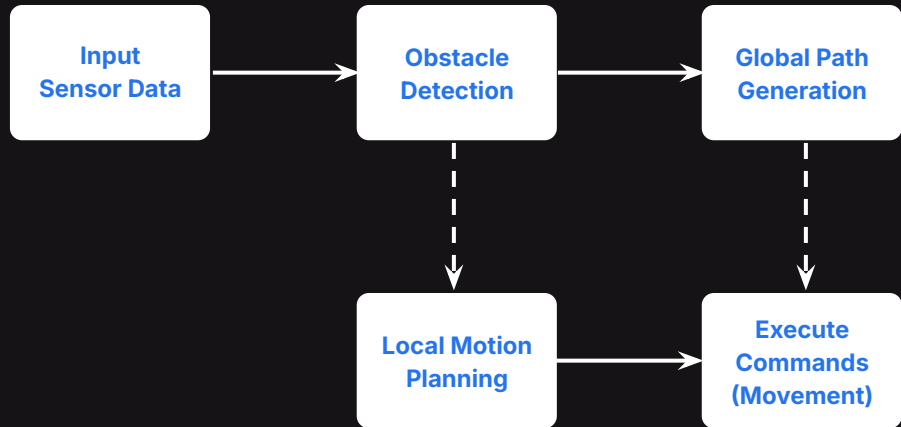
- Cloud Hyperscalers (Microsoft, AWS, Google, Oracle):** All major cloud providers rely on Nvidia GPUs to power their AI services:
 - Microsoft (Azure):** Nvidia and Microsoft announced a multi-year collaboration to build a massive AI supercomputer in Azure, combining tens of thousands of Nvidia GPUs with Nvidia's high-speed networking and software stack [nvidiaresearch.com](#). Azure's latest GPU instances (ND H100 v5) are based on Nvidia H100 NVL platforms [azure.microsoft.com](#). Notably, OpenAI's models (like ChatGPT and GPT-4) are trained on Azure, using Nvidia A100/H100 GPUs in large clusters [nvidiaresearch.com](#). Microsoft's partnership with Nvidia ensures Azure remains a top choice for AI workloads, offering state-of-the-art Nvidia hardware to its customers.
 - Amazon Web Services (AWS):** AWS has long offered Nvidia GPU instances (P3, P4, P5 instances, etc.) for AI. In 2023 AWS announced EC2 UltraClusters with up to 20,000 H100 GPUs interconnected for extreme-scale training [nvidiaresearch.com](#). Nvidia works with AWS to integrate features like Nvidia's Quantum-2 InfiniBand in these clusters, enabling near-supercomputer performance in the cloud. While AWS has developed its own AI chips (Trainium for training, Inferentia for inference) to reduce cost, the AWS CEO characterized their in-house chips as a “supplement” to Nvidia GPUs, not a full replacement, given Nvidia's broad capabilities [nvidiaresearch.com](#). In practice, Nvidia GPUs still power many of the highest-end AI jobs on AWS, and AWS's support of the Nvidia ecosystem (e.g. selling Nvidia's DGX systems via AWS Marketplace) shows a cooperative stance.
 - Google Cloud:** Google famously designed TPUs (Tensor Processing Units) for its internal AI needs, yet Nvidia remains an important partner. Google Cloud offers Nvidia A100 and H100 instances to customers, and even Google's own AI research uses Nvidia hardware for certain projects. For instance, Google's next-gen Gemini AI model is reportedly being trained on Nvidia GPUs, despite

Architecture: Deep Research & Structured Report Generation



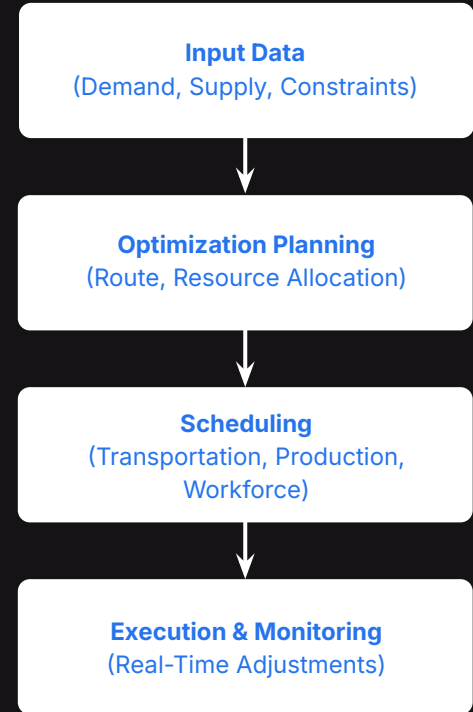
Autonomous Vehicles (Robotics & Navigation)

- In autonomous vehicles, planning patterns are crucial for **real-time decision-making** and **ensuring safe navigation** through dynamic environments.
- **Step 1:** Input sensor data (cameras, LiDAR, etc.) about the vehicle's environment.
- **Step 2:** Process data to detect obstacles, road signs, traffic, etc.
- **Step 3:** Generate a **global path** (based on GPS, maps).
- **Step 4:** Use **local motion planning** (for obstacle avoidance, immediate road conditions).
- **Step 5:** **Execute commands** (accelerate, brake, steer) based on processed information.



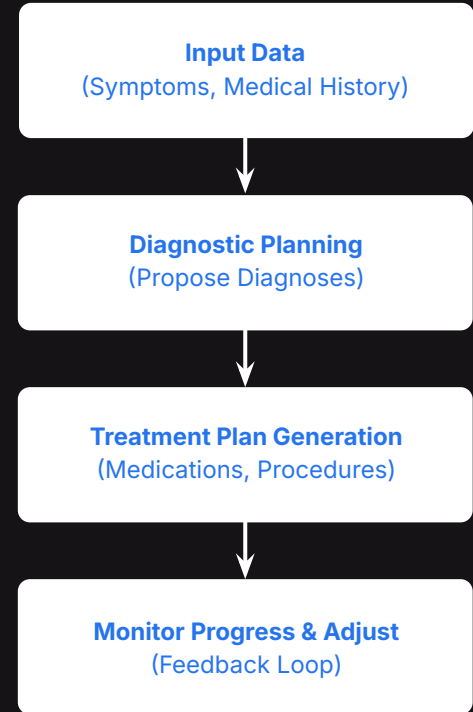
Supply Chain Management

- In supply chain management, planning patterns **optimize the flow of goods and resources**, reducing costs and improving delivery efficiency.
- **Step 1:** Input demand forecasts, supply conditions, and constraints.
- **Step 2: Optimization Planning** to calculate the most efficient supply chain paths.
- **Step 3: Generate scheduling plans** for transportation, production, and staffing.
- **Step 4: Execute and monitor real-time adjustments** based on new information (weather, inventory status).



Healthcare & Medical Diagnostics

- In healthcare, planning patterns assist in **diagnosing conditions** and **formulating personalized treatment plans**, improving patient care.
- **Step 1:** Input patient data (symptoms, medical history).
- **Step 2:** **Diagnostic Planning** to propose possible diagnoses.
- **Step 3:** **Treatment Plan Generation** based on diagnosis.
- **Step 4:** **Monitor** patient's progress and adjust treatment as needed (feedback loop).



Manufacturing Automation

- In manufacturing automation, planning patterns **optimize production schedules** and **resource usage**, for cost savings and efficiency.
- **Step 1:** Input production requirements (product specifications, deadlines).
- **Step 2:** **Job-Shop Scheduling** to optimize machine use and task order.
- **Step 3:** **Real-Time Adjustments** for any changes in machine status or raw material availability.
- **Step 4:** **Final Production Execution** based on optimized schedule.



Thanks!