NVIDIA | Deloitte.

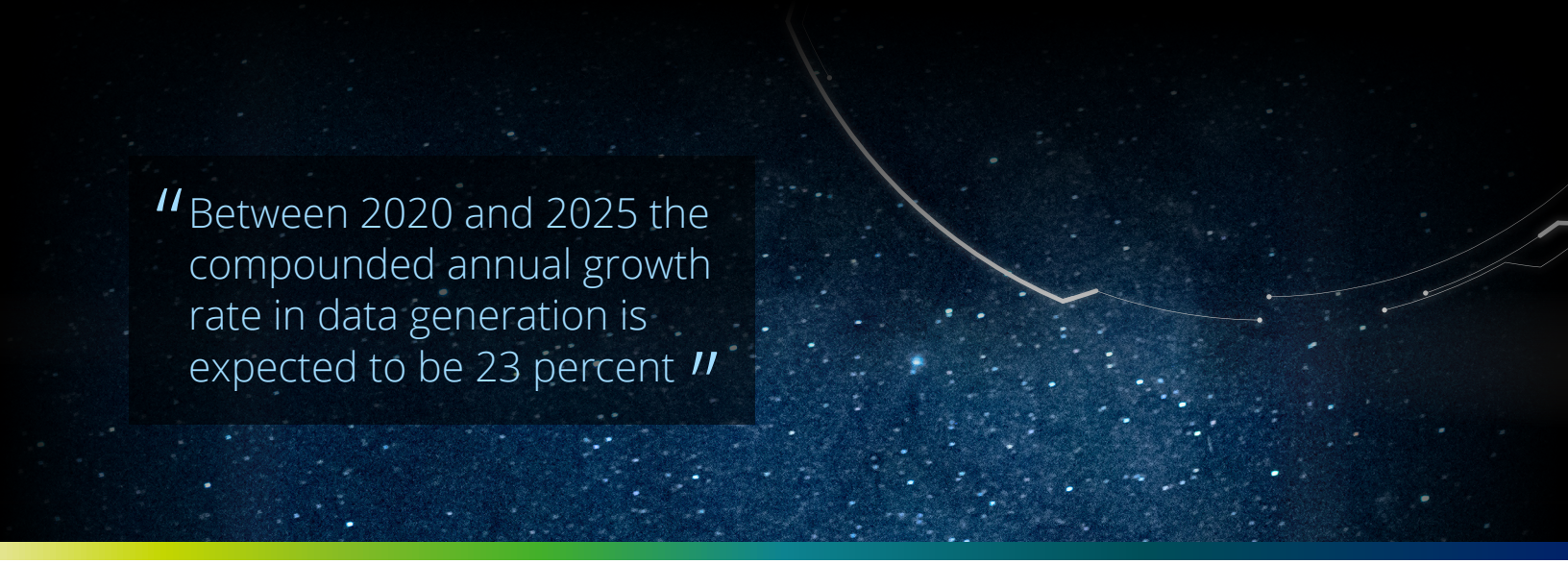
Accelerated Computing for AI in Government

The government is building momentum in adopting artificial intelligence (AI) and is seeing more use cases go into production. Between 2020 and 2021, federal agency budget requests for AI research and development increased by more than 50 percent[1]. Major legislation, including the 2020 National AI Initiative Act, the 2019 AI in Government Act, and the 2018 OPEN Government Data Act are laying the groundwork for widespread integration of AI across government agency operations. To further coordinate and advise federal agency AI efforts, the National AI Advisory Committee was launched in September 2021 and the White House created a National AI Initiative Office in January 2021.

Despite these advances, many agencies on the AI journey will soon confront substantial challenges and obstacles – if they haven't already. At scale, AI often requires access to large data sets, demanding compute resources, speed, and complex networking. As AI implementation gathers pace across government, agencies may find that they do not have ready access to such capabilities and assets.

The challenges arise because the unique computational needs of AI quickly outpace the capabilities of traditional data center architectures. Many AI computations can be done faster and more effectively with GPU-based and other accelerated computing architectures. As a result, designing a plan for an accelerated infrastructure—from the beginning—will help agency leaders maximize the impact of their AI investments.

[1] NITRD, "Artificial Intelligence R&D Investments," Office of Science and Technology Policy, The White House, accessed October 11, 2021, https://www.nitrd.gov/apps/itdashboard/ai-rd-investments/



AI brings needed benefits but demands careful planning

The latest AI applications adopted by government agencies generate substantial benefits in terms of their ability to predict, simulate, and automate agency operations and mission-oriented outcomes. However, they also create new challenges around data, modeling performance, and security planning.

The appetite for data to train newer and higher performing AI algorithms continues to increase dramatically. For example, the popular natural language processing (NLP) model GPT-3 released in 2020 was trained on 570 gigabytes of text. Its predecessor, GPT-2, was released just one year prior but was trained on only 40 gigabytes of text. As NLP and other modeling techniques further evolve, their training and data requirements will continue a similarly rapid rise.

Additionally, technologies such as edge computing and the Internet of Things have led to an exponential rise in data generation. Between 2020 and 2025 the compounded annual growth rate in data generation is expected to be 23 percent, outstripping the forecasted growth in the data storage capacity[2]. As a result, government entities have an increasing need to analyze, manipulate, and capitalize on growing data sets efficiently and effectively.

Advances in modeling techniques are also presenting new demands on organizations. The deployment of AI applications to address complex policy problems and scenarios now requires very high speed and accuracy, including real-time, low-latency processing and rapid training cycles. This is especially relevant for complex modeling and simulation, such as agent-based simulations for traffic systems, war-gaming for space systems, molecular drug discovery, and model-based systems engineering.

Other challenges include security concerns for government clients (e.g., the Department of Defense and national security organizations) that require extensive control over data, server space limitations that restrict scale, and infrastructure costs. Additionally, niche use cases, such as autonomous machines and vehicles, often require edge computing in areas where large batches or streams of data cannot be transmitted for processing in near real-time.

Various architectural solutions can address these challenges, but balancing product design, workload needs, security concerns, and costs requires careful assessment.

[2] IDC, "Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts" (International Data Corporation, March 2021), https://www.idc.com/getdoc.jsp?containerId=prUS47560321

Accelerated Computing to the rescue

Many AI applications rely on central processing units (CPUs) for computations. In contrast to CPUs, graphics processing units (GPUs) use parallel processing to break down complex computational problems into pieces that are then executed simultaneously. This capability makes GPUs valuable for compute-intensive subroutines, such as matrix multiplication, deep learning, and other demanding processes that power advanced AI use cases. In addition, newer classes of network adapters, such as SmartNICs and Data Processing Units (DPUs), can move data from storage and between servers at a high transfer rate, and also have the ability to accelerate a broad range of advanced networking, storage, and security services. These together with GPUs provide the basis of an accelerated computing infrastructure.

The additional power can provide many potential benefits. A case in point is cost savings over time. Because models generally take less time to train on accelerated computing infrastructure, resource consumption is often reduced. Along with faster training, accelerated computing results in faster predictions and decision making.

Government entities can access accelerated computing on premises or in the cloud, whether with general-purpose or supercomputers, customizing and scaling the approach to meet business needs. Agencies should consider several factors, such as the size of the project and data sets, project duration, budget, and future growth, in determining the right approach.

> "(GPUs) use parallel processing to break down complex computational problems into pieces that are then executed simultaneously"

On premises

Government entities can purchase individual workstations for data scientist power-users or set up enterprise data centers. Enterprise servers are ideal for organizations building scalable AI that demands the pinnacle of performance. They offer greater control and flexibility of the infrastructure but require more management and maintenance from IT teams. Although purchasing hardware entails higher up-front fixed costs, there are no limitations on the timespan or size of projects other than what the hardware can handle.

Good candidates for on-premises accelerated computing include government entities with strict data security regulations, especially those that handle sensitive or confidential information. Examples include the Department of Defense and national security agencies, federal health entities, and certain civilian sector entities, such as the Internal Revenue Service.

Two examples illustrate the current uses of on-premises accelerated systems.

Mass General Hospital and Brigham and Women's Hospital Center for Clinical Data Science use on-premises NVIDIA GPUs and underlying software to power generative adversarial networks that create synthetic brain MRI images. These synthetic images improve the training of deep learning models that classify MRI scans with a higher degree of confidence, which in turn better serves doctors and clinicians by assisting their diagnosis productivity. Higher diagnostic efficiency leads to improved doctor performance. This naturally cascades to much better patient outcomes, which is the critical factor for healthcare providers[3].

Swiss Federal Railway has 15,000 trains that provide 1.2 million rides per day. A NVIDIA DGX system powers fault detection in railway tracks, reduces time for onsite inspections, and supports simulations and deep reinforcement learning to optimize train schedules and dispatching[4]. With the DGX, engineers can simulate the daily scheduled movement of all its trains in just 3 seconds, which helps reduce delays and supports increasingly complex train dispatching[5].

[3] "AI Can Generate Synthetic MRIs to Advance Medical Research," NVIDIA Developer (blog), September 16, 2018, https://developer.nvidia.com/blog/ai-can-generate-synthetic-mris-to-advance-medical-research/

[4] Scott Martin, "AI and GPUs Could Lead to Autonomous Trains," August 15, 2018, https://blogs.nvidia.com/blog/2018/08/15/autonomous-trains-deep-learning-dgx-drive/

[5] NVIDIA, SBB Improves Train Management with NVIDIA DGX-1, 2019, https://www.youtube.com/watch?v=byarUcd58Ug

Cloud–based

Cloud computing offers several benefits over traditional on-premises architectures, namely on-demand access to services such as compute power and storage, allowing users to "pay as you go" instead of absorbing up-front costs and trying to forecast future demand patterns. These advantages, if constructed properly, can lower the organization's capital expenditure and allow it to scale operations in an agile, resilient, and reliable manner. Providers offer flexible pricing and deployment options that are scalable based on project size.

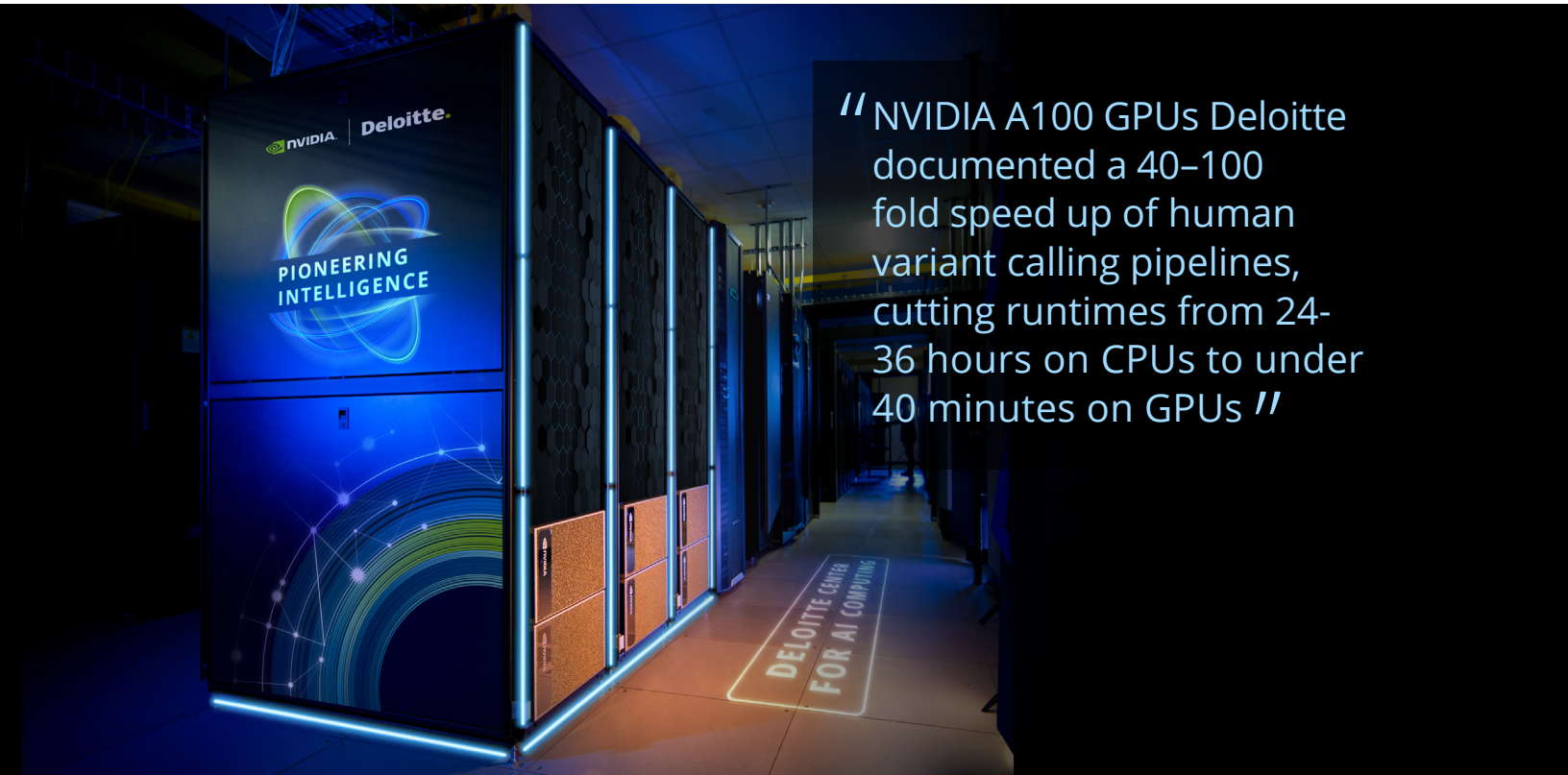
Cloud service providers that handle sensitive government data are required to go through the Federal Risk Authorization Management Process (FedRAMP), which applies standards set by the National Institute of Standards and Technology. Providers that initially fail to meet the standards have subsequently been bringing their infrastructure to code.

Deloitte uses cloud-based accelerator environments for the testing and validation of genomic, proteomic, and imaging data. This has resulted in the cloud deployment of NVIDIA/ NIH prebuilt Clara diagnostic and segmentation models using the MONAI framework with CT and MRI scans. This framework expedited the training of models and analysis of images for COVID-19 detection and brain tumor segmentation. In addition, due to the heightened awareness and urgency around genomic sequencing and biosurveillance related to Covid-19, Deloitte has been benchmarking NVIDIA's's GPU accelerated genomic sequencing and variant calling workflows. Working with NVIDIA's Parabricks genomic analysis package and running on the latest NVIDIA A100 GPUs Deloitte documented a 40–100 fold speed up of human variant calling pipelines, cutting runtimes from 24-36 hours on CPUs to under 40 minutes on GPUs. This time saving becomes critical to provide real-time biosurveillance or to enable genomics as a component of diagnostic precision medicine.

Supercomputers

Supercomputers send data quickly between processors, which speeds up training toward a specific task and produces faster results important for real-time applications. They are highly effective when the volume of data and scale of a model require computing power only possible through distributed high-performance computing. Additionally, GPU-based supercomputers offer high level of performance with the largest computing memory capacity and processing speed. They excel in power, performance, and efficiency for AI workloads by enabling parallelized processing of data and models at scale.

Use cases that involve consistently large volumes of data generation from IoT or edge systems may be able to best leverage the comparative advantages of supercomputers. For example, sensor data that is constantly collected from autonomous vehicles could be relayed to super computers where underlying AI algorithms for autonomous driving are retrained and improved. Given the volume of data and complexity, computing at the edge or in the cloud may be less practical. A supercomputer would be better suited for these tasks before updates are uploaded to edge systems.



"NVIDIA A100 GPUs Deloitte documented a 40–100 fold speed up of human variant calling pipelines, cutting runtimes from 24-36 hours on CPUs to under 40 minutes on GPUs"

Plan for scale and success

The tradeoffs and dilemmas posed by AI needs and computational considerations are rarely obvious and always evolving. Agencies need to design a flexible plan for incorporating accelerated computing in the early stages of AI adoption. Moreover, AI needs and applications will evolve over time. Accordingly, the following steps are important considerations:

> "Agencies need to design a flexible plan for incorporating accelerated computing in the early stages of AI adoption"

- Build a clear AI strategy that identifies how an agency's mission and operations can be reinforced and scaled up through the prediction, simulation, automation, and other core functions of artificial intelligence.
- Identify within your existing AI strategy those use cases that will be most aided by accelerated computing, specifically ones that involve distributed data collection and flows, real-time analysis and speed, and involve complex dynamics, among other factors.
- Be mindful of user perspectives and interests. Existing workflows and workforce practices will have a major impact on the potential benefits of AI as well as the shape, location, and distribution of accelerated computing architectures. Significant gains could be left unrealized if human design considerations are treated as an afterthought.
- Identify a multi-disciplinary team that can help define an execution plan. Team members need extensive industry knowledge, AI expertise, and a demonstrated understanding of accelerated computing and solution development.
- Consider future needs. Issues to weigh in selecting the right projects include long-term relevance; relative costs and benefits of an AI solution; scaling up pilot initiatives; and security and confidentiality.
- Consider dedicated computing options to realize the inherent benefits of greater computing power and speed as usage increases. This is especially relevant as usage and data processing requirements increase and needs and inference become more nuanced.

Agency leaders have a great opportunity to apply AI to help their organizations. As usage gets more widespread and complex, they are going to need to utilize accelerated computing architecture solutions, such as GPUs and optimized software, to support their AI journey. Planning for these capabilities early will allow for seamless expansion and sophistication of AI use cases.

Get in touch

Christine Ahn
Principal
Deloitte Consulting LLP
chrisahn@deloitte.com

Anthony Abbattista
Principal
Deloitte Consulting LLP
aabbattista@deloitte.com

AUTHORS

By Anthony Robbins
Vice President
NVIDIA Federal

Ed Van Buren
Principal Government and Public Sector AI Leader
Deloitte Consulting LLP