

Defining a Practical Path to Artificial Intelligence

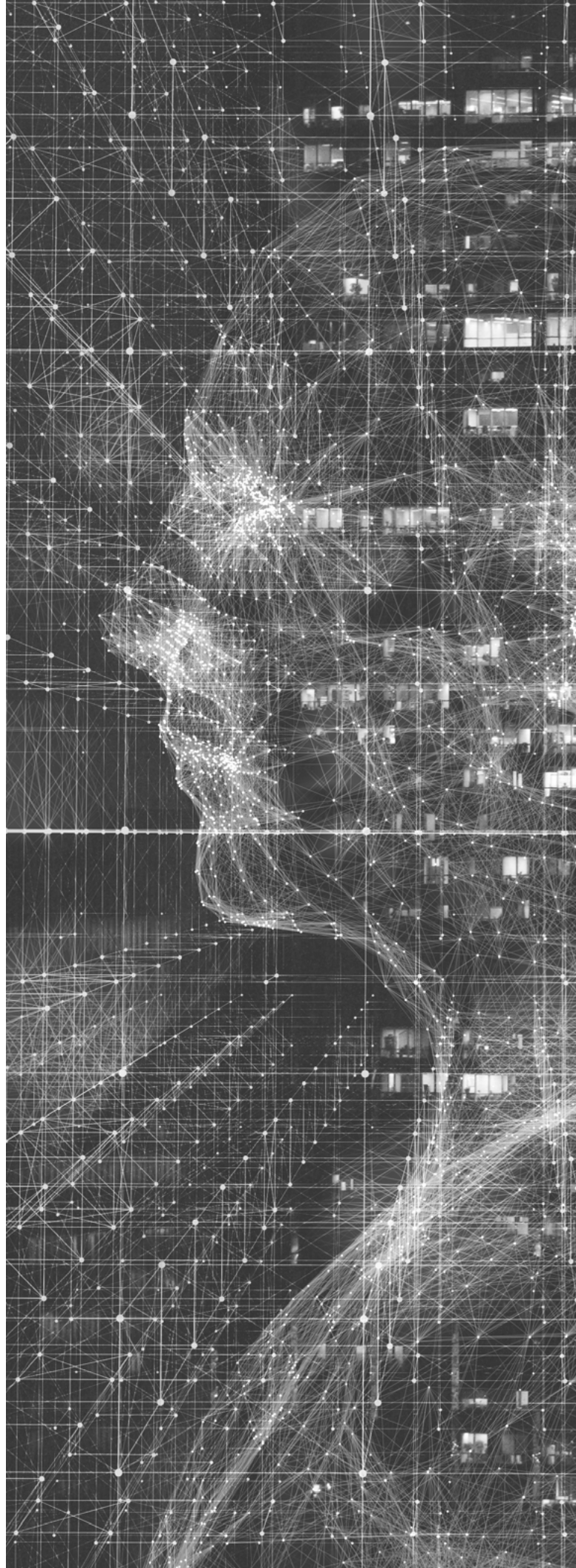


Foreword

Extracting actionable insights from raw data is critical to the success of virtually any business or organization in any industry – from finance, to healthcare, government and manufacturing. The data sets that are used to glean these insights are getting exponentially larger and being created exponentially faster. The pace of data growth will only continue to accelerate. Since the term “big data” was coined in the 1990’s we have made great strides in deriving intelligence from these massive pools of information, but the reality is, we are just scratching the surface of what is possible. Artificial Intelligence (AI) is able to discover deeper insights and identify correlations between different data sets far faster and more completely than the human brain could possibly process. The steady rise of AI as a business tool marks the beginning of a new technology wave. Is AI accessible for organizations of all sizes? According to Gartner’s 2018 CIO Agenda Survey, **only four percent of CIOs have implemented AI**, while a further 46 percent have developed plans to do so. This is typical with every new technology trend – a slow increase of early adopters emerge until something comes along to change the dynamic and make access and implementation more practical.

As the Gartner report points out, there is a perceived steep learning curve with AI. While many enterprises want to kickstart their AI initiatives, the challenges to building an AI-optimized infrastructure typically hold them back. But like the big data revolution 10 years ago, AI is actually more approachable and simpler than many CIOs may think. In fact, with the evolution of purpose-built AI Infrastructures and the advancement of Graphics Processing Units (GPUs) that enable massively parallel, deep analysis in real-time; cognitive computing may be the norm in data centers in record time. But how?

In this paper we will evaluate the path to AI for public sector agencies and commercial organizations, identify key questions and options to consider, and determine how to accelerate implementation and management within existing budgets.



Chapter 1: Addressing Roadblocks to AI

AI is the next evolution of big data and data analytics initiatives that have been in play across organizations for a decade. What's holding enterprises back from moving forward in adopting AI? It boils down to three main issues:

- ▶ A struggle to “do it yourself” (DIY), build an AI practice
- ▶ Finding the talent to build and support an AI practice
- ▶ Funding AI

The Burden of DIY

As new technologies come into the market, enterprises are often stuck with building their own solutions. The same holds true for AI. When “DIY” is the only option, AI initiatives stall out.

- ▶ **Software** Users are forced to take open source code, often with limited documentation, then compile and tune the code. In tuning, there are countless knobs that they need to tweak to get more out of their training runs.
- ▶ **Hardware** Users often spend months researching various options and running Proofs of Concept (POC) to see which delivers the best performance. Even after purchasing the system, constant maintenance is required.
- ▶ **Slow Workloads** Even when all of this is done, users find that workloads are slow. It's because the system is built with legacy components, which are cobbled together, creating bottlenecks.

An Evolving Workforce

Like the evolution of every other technology, early trepidation often stems around having a workforce with the skills and talent to support IT. The understanding gained over the past 50 years can become irrelevant - from human skills to tools to infrastructure - in the past everything has changed dramatically. But in this case, the AI evolution doesn't replace the architectural foundation of the approach that already exists in - it augments it - in effect allowing employees to make more informed decisions and pursue more strategic oversight of data.

According to Whit Andrews, Research Vice President and Gartner analyst, the most transformational benefits of AI in the near term will arise from using it to enable employees to pursue higher-value activities.




The DIY Dilemma:

- ▶ Never-ending cycles of compiling and customizing open source software
- ▶ Months of system building and tuning, constant maintenance
- ▶ Legacy solutions full of data bottlenecks, from storage to GPU to apps

What's Changing? The Algorithm is the Same, but the Technology is Different

AI is based on neural networks and at its core is mathematical. It's the application of statistical data modeling to formalize relationships in data variables based on equations. This modeling relies on data and hypotheses. Computational modeling has been in play since the computer entered the world, but using this statistical modeling to solve bigger challenges, not just math problems, is where AI takes the stage.

The new AI concept is still centered on neural networks, feeding data and continually adjusting the coefficients to minimize the margin of error. Therefore, the process that the workforce has come to understand with early data analysis is the same. In the past, traditional CPUs could process these coefficients ten at a time and now thanks to the processing power that has emerged from the gaming world, a GPU can filter thousands to millions of coefficients and dramatically speed up analysis of data.



Gartner predicts that by 2020, 20 percent of organizations will dedicate workers to monitoring and guiding neural networks.

Chapter 2: A Technology Collision with Perfect Timing – Purpose-Built AI Infrastructures

The DIY dilemma becomes obsolete, thanks to a perfect collision of technology. Now it's practical rather than theory. What happened?

- ▶ **Hardware Improvements** - Hardware evolution in recent years has dramatically changed the volume of data that can be processed and the speed at which it can happen. Thanks to the gaming and research industries, GPUs are ubiquitous and companies like NVIDIA now create purpose-built GPUs for AI, which process data 45 times faster than traditional CPUs. GPUs also provide thousands of cores per socket, making them exponentially more dense than traditional CPUs, and provide scalability unachievable in legacy architectures. Modern storage technology from manufacturers, like Pure Storage, now support extremely high IOPs and bus speeds using NVMe. PCIe and NVLink can now transfer many GB per second for these massively parallel and concurrent workloads.
- ▶ **Software** - Early adopters on the frontier of AI have now built user-friendly software solutions designed for

neural networking. NVIDIA's CUDA, Apache's MXNet and Google's Tensorflow-GPU are all robust software options that make AI easier to execute. Dataiku, which leverages these aforementioned technologies, has gone so far as to make AI a click-and-drag exercise, accessible by any level of user.

- ▶ **Containerization** - This is an alternative to full machine virtualization that involves encapsulating an application in a container with its own operating environment. The adoption of containerization allows IT architects to leverage software, data science and this standardized approach to the algorithm that has catapulted the industry forward. Containerization enables rapid innovation and development of algorithms, dramatically reducing the time from initial development to deployment in production. It also facilitates rapid adaptation if the underlying datasets change or evolve. We see this evidence in threat detection and the growing strength of our nation's cybersecurity profile every day.

CPU vs. GPU Comparison

Total Vertices	Size of Adjacency Matrix	CPU Time(s)	GPU Time(s)	Speedup
1000	1,000,000	3.9s	0.103s	37.86x
2000	4,000,000	30.90s	0.698s	44.34x
4000	16,000,000	244.22s	5.09s	47.98x
8000	64,000,000	1941.0s	39.1s	49.64x
10000	100,000,000	3778.1s	77.8s	48.56
11111	123,454,321	5179.2s	108.01s	47.95

Table 1: Simple implementation of the Floyd-Warshall all-pairs-shortest-path algorithm written in two versions, a standard serial CPU version and a CUDA GPU version¹.

On average the GPU time is 45x faster!

As data volume grows at an epic pace, these massive data sets are meeting this hardware, software and computing power, creating the perfect collision of technology that can provide a distinct advantage to organizations. Where in the past it has been really difficult for small agencies

to get compute, network and storage that is built off the protocols they understand already, today, with the advent of infrastructures designed for AI – you no longer need a PhD to leverage AI anymore.

¹Stephen D. Boyles. User Equilibrium and System Optimum. <https://www.slideshare.net/VishalSingh405/cpu-vs-gpu-presentation-54700475>

What is AI Ready Infrastructure?

Deep learning requires more than fast compute and high-bandwidth interconnects. When designing a full-stack platform for large-scale deep learning, the system architect's goal is to provide a well-integrated, adaptable, scalable set of compute, storage, networking and software capabilities that seamlessly work together providing peak efficiency and effectiveness. This requires provisioning as many GPUs as possible, while ensuring linearity of performance as the environment is scaled, all the while keeping the GPUs fed with data.

Keeping the GPUs fed requires a high-performance data pipeline all the way from storage to GPUs. AIRI™ supplies the right balance of storage, performance and scalability. When defining storage for deep-learning systems, architects must consider three requirements:

- ▶ **Diverse Performance** - Deep learning often requires multi-gigabytes-per-second I/O rates but isn't restricted to a single data type or I/O size. Training deep neural network models for applications as diverse as machine vision, natural-language processing, and anomaly detection requires different data types and dataset sizes. Storage systems that fail to deliver the performance required during neural network training will starve the GPU tier for data, and prolong the length of the run, inhibiting developer productivity and efficiency. Providing consistency of performance at various IO sizes and profiles at a capacity scale will ensure success.

- ▶ **Scalable Capacity** - Successful machine learning projects often have ongoing data acquisition and continuous training requirements, resulting in a continued growth of data over time. Furthermore, enterprises that succeed with one AI project find ways to apply these powerful techniques to new application areas, resulting in further data expansion to support multiple use cases. Storage platforms with inflexible capacity limits result in challenging administration overheads to federate disparate pools.
- ▶ **Strong Resiliency** - As the value of AI grows within an organization, so does the value of the infrastructure supporting its delivery. Storage systems that result in excessive downtime or require extensive administrative outages can cause costly project delays or service disruptions.
- ▶ **AI-Ready Infrastructure (AIRI)** provides enterprises a simple solution to hit the ground running, and is the product of the partnership between storage experts Pure Storage, and high-performance computing experts NVIDIA. AIRI aims to solve infrastructure complexities. It eliminates the difficulties of building an AI data center – while neatly and compactly delivering all the necessary components in a small form factor. Now every enterprise can finally start to explore what AI can do with their most important asset, data.

Cost Concerns Answered

Even as AI-ready infrastructure becomes accessible and deployable for all the reasons described, the elephant in the room is often still: How do we pay for it and how can we implement it quickly and scale for the future? This is where the perfect collision of needs and technology continues and gains additional momentum with the advent of modern acquisition models that provide greater financial flexibility.

Just as the democratization of high performance computing has catapulted AI, the as-a-Service model has simplified acquisition of this technology. Acquiring AI infrastructure as-a-Service allows users to manage and maintain it from a monthly operating budget versus investing CapEx dollars up-front. With as-a-Service, costs are more predictable and organizations can scale as needs change.

Chapter 3: Deploying AI: Evaluating Operational Impact

Ready for AI? Key Questions to Consider

According to Gartner researcher, Whit Andrews, the main goal in the AI revolution should be to “normalize AI planning and development for the whole organization, including leaders of data and analytics, applications and lines of business.” These simple steps help bring together those leaders around a common purpose:

- ▶ Assess which business outcomes would benefit most from AI
- ▶ Evaluate AI as simply the latest advanced analytical technology available that might help achieve those outcomes

As that evaluation of technology becomes more granular, speed with data begins to take center stage. The faster teams can iterate on a model, the better. And that means doing everything possible to eliminate data transfer time and other such variables, so models can iterate dramatically faster. AI needs will continue to evolve and force infrastructure to keep pace with pushes for speed and performance. In that evaluation process, organizations should give great value to what will serve their current needs but also acknowledge outgrowing them is inevitable. Data infrastructure has to be future-proof to deliver high performance and increase the productivity of data scientists or else it too will become obsolete.

Here are key questions to consider whether your organization can benefit from a purpose-built AI infrastructure:

- ▶ Do you have data silos where data has to be copied from one to another?
- ▶ Do you have a data science team working with terabytes of data?
- ▶ Does it take too much time to go from cleaning data to getting feedback about a model?
- ▶ Do you spend a significant amount of time managing and tracking versions of training data sets?
- ▶ Do you spend significant resources on keeping your current infrastructure online, replacing components or keeping up with workflow changes?

A Timeline to AI: How AI Ready Infrastructure as-a-Service is Changing the Game

Federal government has been at the forefront of the AI revolution since the beginning, but outside of the agencies that are early adopters, the notion of implementing AI on a budget has appeared to be a tedious and time intensive task.

That's where AI as-a-Service offers a tremendous solution – not just for federal agencies, but organizations of all types. For smaller or mid-size agencies or any organization struggling with juggling resources, AI as-a-Service can be the way to get started.

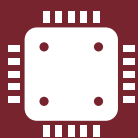
AIRI in conjunction with the as-a-Service model can dramatically impact timelines:



DIY

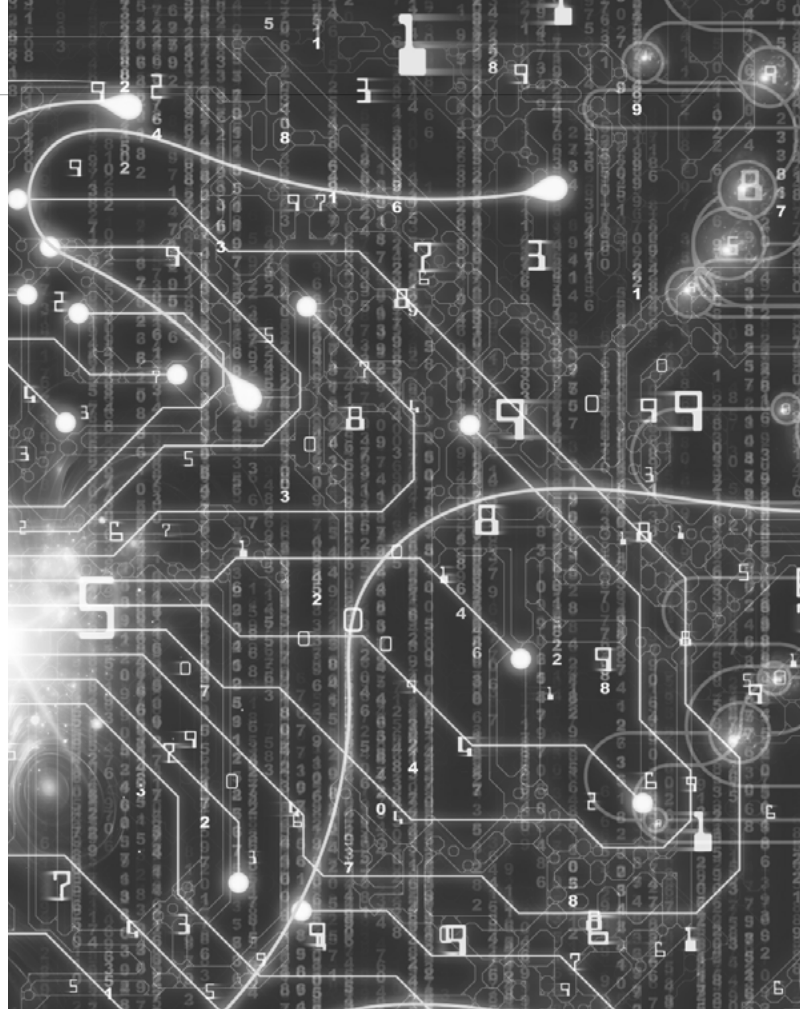
There is typically an **18 month procurement** and acquisition process in Federal organizations and an 18 month deployment process = 36 months to realize any benefit from a typical AI solution acquired via traditional structures.

VS.



AIRI
as-a-Service

AIRI is ready to deploy. It only takes a day to stand up the capability –and within 30 days of migrating the data, organizations can reap the results.



Cost & Space Impact of AIRI: 50 Racks Under 50 Inches

AIRI reduces racks of complexity into a complete solution – delivering the performance of 50 racks of legacy technologies under 50 inches. It's less than 20 rack units, which provides the performance of an entire data center building in a form factor that's a little taller than waist-high. The power and space savings here are dramatic for organizations. The system is also run on Ethernet, not InfiniBand, which also reduces complexity and cost.

This streamlined infrastructure also simplifies “time to insight” and the productivity of data scientists. Most data scientists run AI workloads from a single server. Scaling a project across many servers is difficult to do, and multi-node scaling was reserved for very few AI experts. With AIRI, any data scientist can do multi-node training and produce actionable results faster. Imagine the potential time savings and the ability to make better decisions – this is possible when you have the production bed to act faster.

Use Case: Quickly Determine Security Threats

A large government agency wanted to perform image classification and object detection on live video feeds to better access security threats. They had an existing model, but they wanted to build on and improve it to keep their process streamlined.

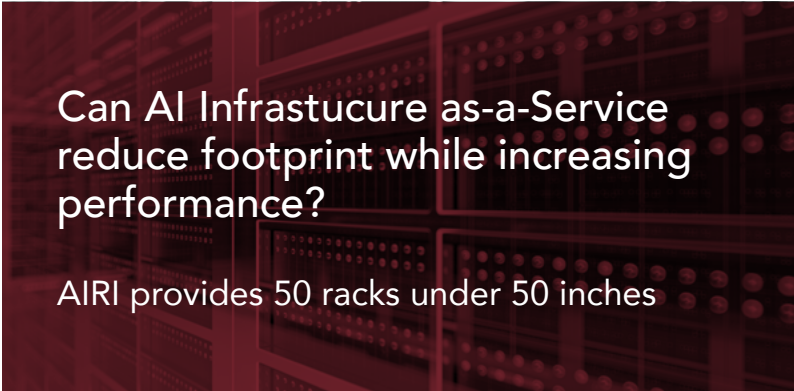
The Challenge

The time, cost and process to develop and approve a new model for object detection is tedious and can require many rounds of review. The organization needed the ability to update an existing model with new images that have been labeled, correctly or incorrectly, to deepen learning and ultimately produce better quality intelligence on which to base decision-making.

The Business Impact of AIRI

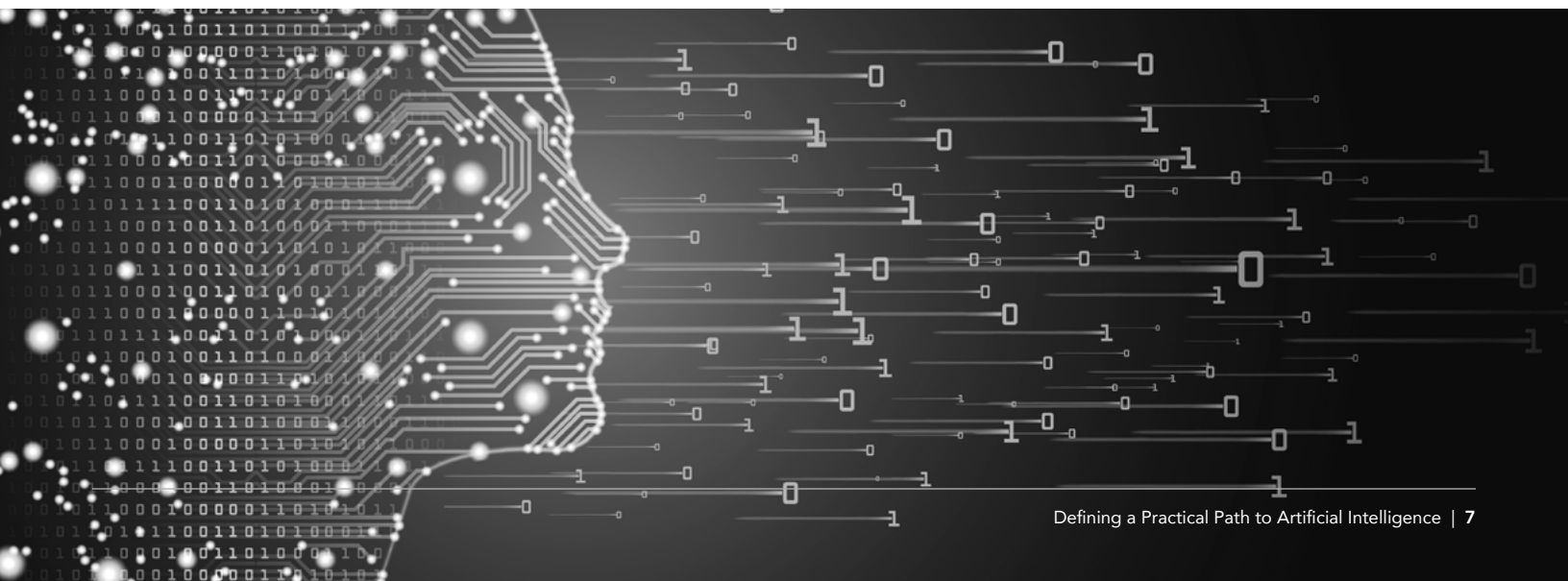
AIRI enables a more efficient approach to modeling that has many positive impacts to the organization including: improved speed, accuracy and performance of data analysis as well as greater independence and dexterity for the analysts.

The GPU power of the AIRI dramatically improves the feedback loop to improve learning and speed the process, thereby reducing the time data scientists spend retraining their models. The cost savings compound over time as a result of deploying more accurate models and reducing the strain on resources. AIRI allows for faster experimentation, data analysis and accuracy supporting stronger outcomes and decision-making for the organization.



Can AI Infrastructure as-a-Service
reduce footprint while increasing
performance?

AIRI provides 50 racks under 50 inches



How ViON, Pure Storage and NVIDIA Work Together

Simply put, AIRI is the technical accelerator with its high-performance infrastructure and ViON is the delivery accelerator with the as-a-Service model. ViON brings direct partnerships and in-house support with cleared resources that provide expertise from hardware to software. With AIRI as-a-Service, ViON enables organizations to have the AI solution on-site behind the customer firewall.

Many AI development teams have limited access to elite GPU systems due to recurring public cloud shortages and the expense of dedicated systems. ViON's secure in-house lab moves extreme high-performance testing into the agile process by default.

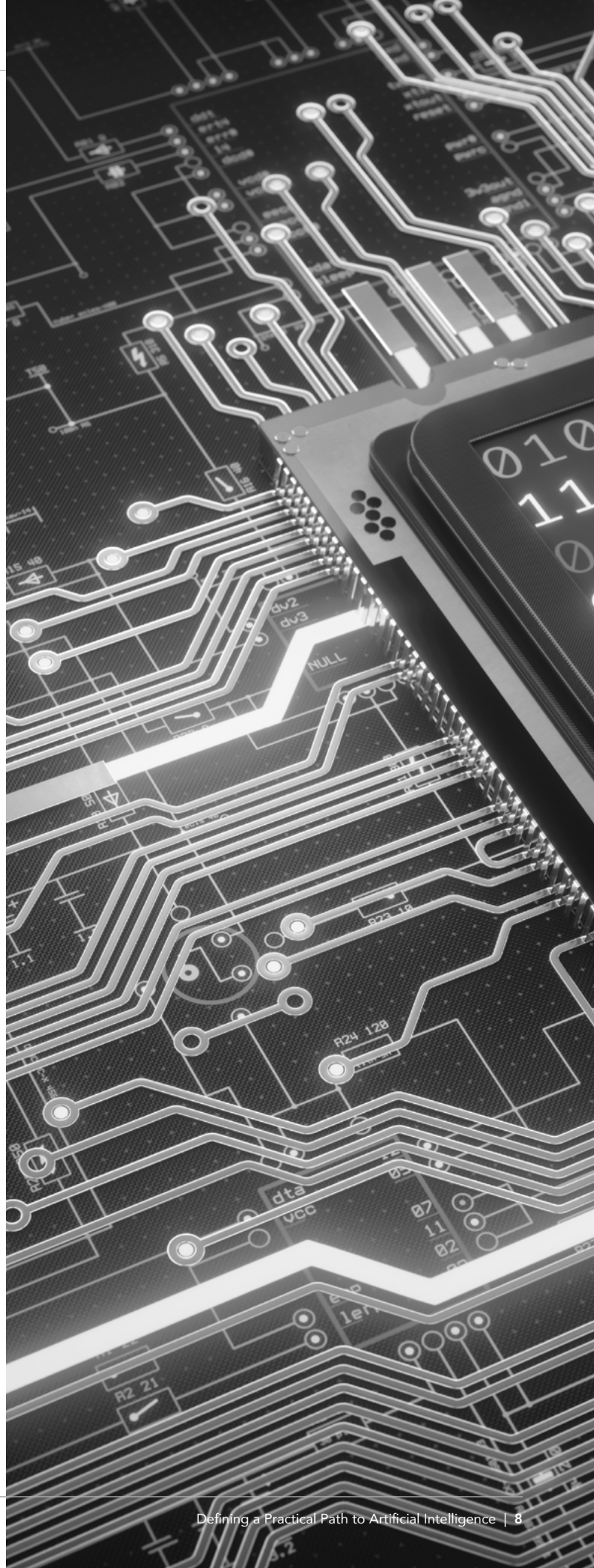
About Pure Storage

Pure Storage (NYSE: PSTG) helps innovators build a better world with data. Pure's data solutions enable SaaS companies, cloud service providers, and enterprise and public sector customers to deliver real-time, secure data to power their mission-critical production, DevOps, and modern analytics environments in a multi-cloud environment. One of the fastest growing enterprise IT companies in history, Pure Storage enables customers to quickly adopt next-generation technologies, including artificial intelligence and machine learning, to help maximize the value of their data for competitive advantage. And with a Satmetrix-certified NPS customer satisfaction score in the top one percent of B2B companies, Pure's ever-expanding list of customers are among the happiest in the world.

About NVIDIA

NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at <http://nvidianews.nvidia.com/>.

For more information contact:
William Hill | billy.hill@vion.com
Lead Data Scientist, ViON





About ViON

ViON Corporation is a cloud service provider with over 37 years' experience designing and delivering enterprise data center solutions to government agencies and commercial businesses. The company provides IT as-a-Service solutions including on-premise public cloud capabilities to simplify the challenges facing business leaders and agency executives. Focused on supporting the customer's evolution to the next generation data center, ViON's Data Center as-a-Service offering provides innovative solutions from OEMs and disruptive technology providers via a consumption-based model. The complete range of as-a-Service solutions are available to research, compare, procure and manage via a single portal, ViON Marketplace. ViON delivers expertise and an outstanding customer experience at every step with professional and managed services, backed by highly-trained, cleared resources. A veteran-owned company based in Herndon, Virginia, the company has field offices throughout the U.S. (www.vion.com).

