Deep Impact of Advanced AI Technology Developments

– Government Regulatory Measures –*

Prof. Dr. V. David Sánchez A., Ph.D.
Brilliant Brains, Palo Alto, California
September 2023

Abstract

The closed-door AI Insight Forum organized by the U.S. Senate's AI working group gathered executives of AI hightech companies including Google, Microsoft, IBM, Meta, X, OpenAI, Nvidia among multiple others on September 13, 2023. The relevance of regulatory activities of the U.S. government including the underlying public policy on AI developments was the subject of the event, in particular how the U.S. Congress should approach the opportunities and risks posed by AI [1] and which if any evolving government regulations will impact and promote responsible AI R&D as well as maintain U.S. competitiveness. On the other hand, the congressionally hosted Hackathon was launched in 2011 as a launchpad for dialog and technological solutions in Congress. Times have changed rapidly since the first of a series of events, where inspiring ideas came from social media and other digital platforms to this year's fifth Hackathon on September 14, 2023 [2], co-hosted by the U.S. House of Representatives' Office of the Chief Administrative Officer (CAO) with major topics including Artificial Intelligence (AI), legislative workflow, constituent casework, and community engagement. In particular, AI has the potential to change the legislative branch's workflow, fears and opportunities are shared in the private and government sectors. Figure 1 shows from left to right: the senate's AI Insight Forum, declarations that regulatory legislation might come in a months-timeframe, and the Hackaton events from the 1st, inagural one in 2011 to the 5th one this year 2023.

To visualize some of the current AI technology developments with major impact, a new generation of AI supercomputer, the Nvidia's DGX GH200, is outlined. Figure 2 left shows an end-to-end, cloud-native suite of AI software called Nvidia AI Enterprise that accelerates the data science [3] pipeline and streamlines the development and deployment of predictive AI models [4]. The Nvidia AI Enterprise platform for production AI includes AI workflows, frameworks, and pretrained models to

*This abstract has been granted permission for public release. The author has been designing and building advanced scalable mission-critical concurrent computer systems and custom-chip-interconnects for AI and ML parallel supercomputers in the framework of Federal U.S., European, and German (NASA, ESA, DLR) and DoD classified programs as well as State of California advanced AI technology programs for next generation data centers.

Figure 1: The AI Insight Forum and the Hackaton organized by the U.S. Senate's Artificial Intelligence (AI) working group and the House on September 13&14, 2023, respectively

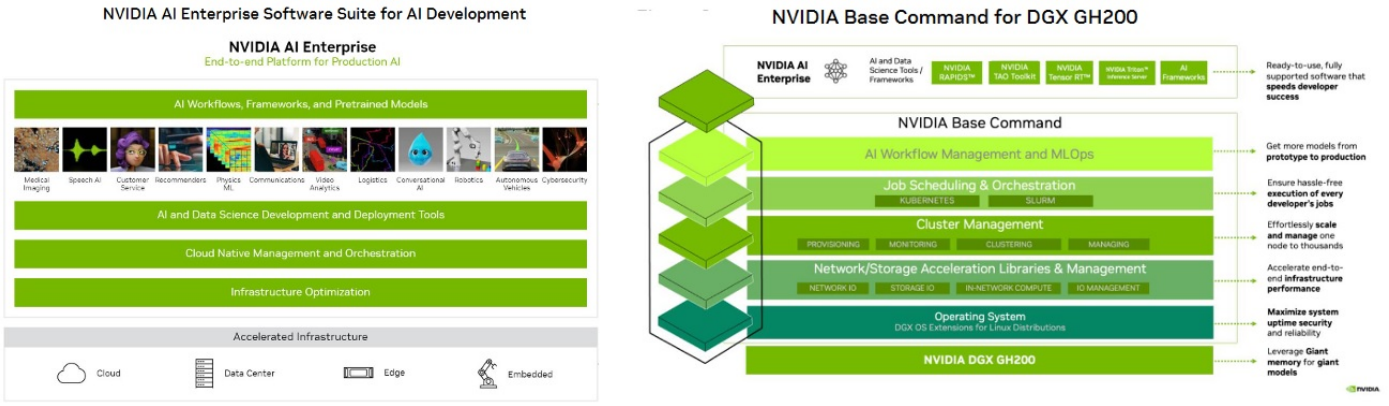

Figure 2: – Software architecture of next generation AI supercomputing

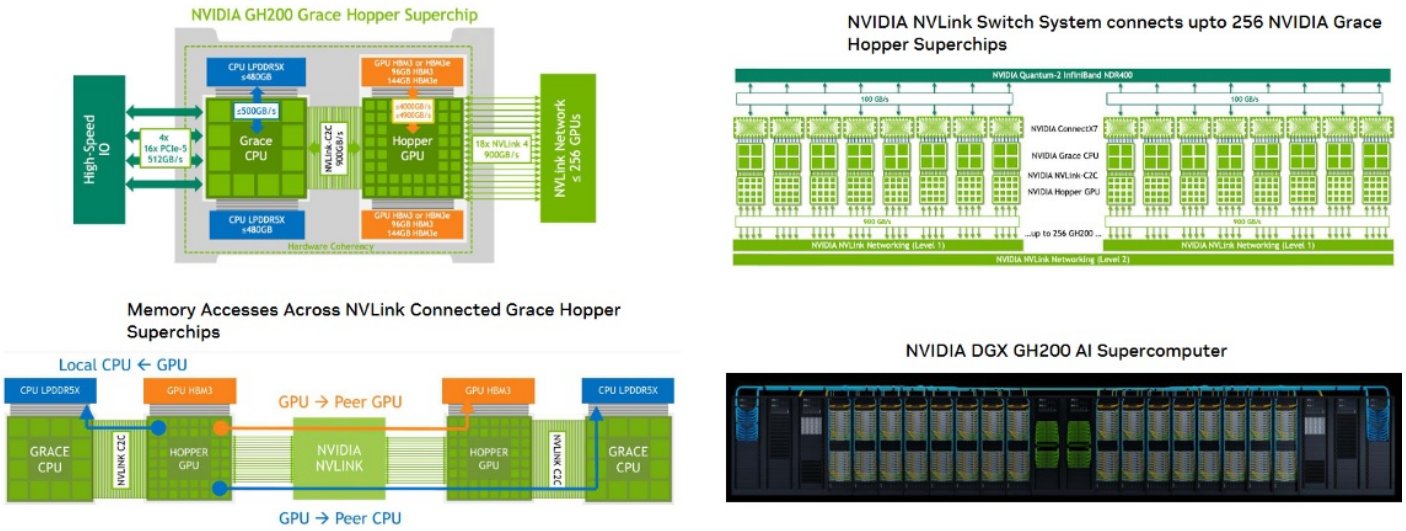

Figure 3: Next generation AI supercomputing – Hardware advances w.r.t. previous generation

AI and Data Science development and deployment tools, to cloud native management and orchestration, and to infrastructure optimization that run on an accelerated infrastructure (cloud, data center, edge, embedded) for diverse applications including medical imaging, speech AI, customer service, recommenders, physics ML, communications, video analytics, logistics, conversational AI, robotics, autonomous vehicles, and cybersecurity. Figure 2 right shows the system software stack for the DGX GH200, called the Nvidia base command based on a customized installation of the Ubuntu Linux OS together with AI workflow management and MLOps, job scheduling & orchestration, cluster management, and network/storage acceleration libraries & management.

Figure 3 left shows the hardware architecture of the DGX GH200 AI supercomputer. On the left side, top and bottom, the GH200 Grace Hopper superchip [5] and memory accesses across NVLink connected GH200 chips, respectively. On the right side, top and bottom, how a NVLink switch system can connect up to 256 GH200 chips and the entire DGX GH200 AI supercomputer, respectively. Those customized switch systems need to be developed when building parallel distributed systems [6]. Custom-designed neurocomputers as an integral part of real-time supercomputers based among others on neurochips were designed already decades ago, for example for national security, DoD classified programs [7] and foreseen for industrial applications as well [8]. Programming for parallel distributed architectures of learning algorithms and AI models in languages for concurrent parallel programming based on the actor model were designed and coded, e.g., in [9]. A vast number of algorithms, AI models, and applications were reported in detail in the author's 15-year tenure as Chief Scientist and Editor-in-Chief of an Elsevier Science's artificial intelligence, machine

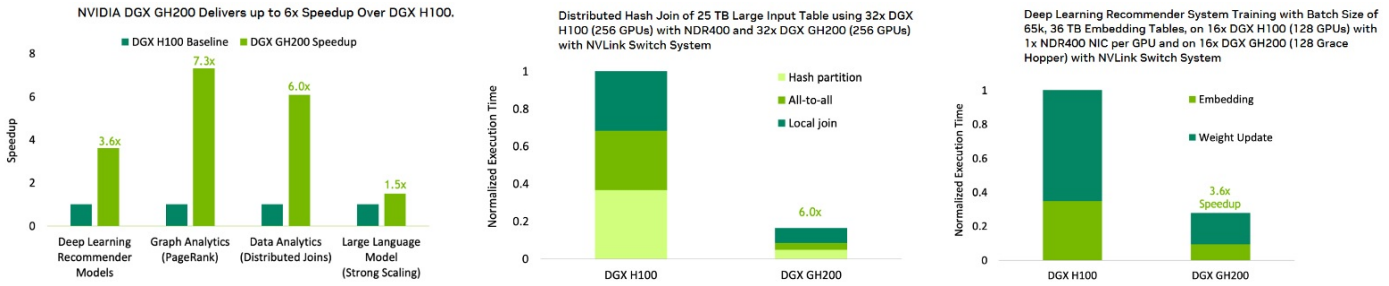

Figure 4: Speedup w.r.t. previous generation AI supercomputing

learning journal [10].

Figure 4 left shows the performance speedup of four different vertical applications when running the corresponding AI models on the next generation DGX GH200 versus on the previous generation DGX H100 AI supercomputer, which amounts for 3x to 7x approximately. Figure 4 center and right shows in further detail the performance speedup of distributed hash joins in databases and large recommender systems, respectively. The systematic and automatic design of artificial brains and building AI supercomputers to cope with the underlying complexity of learning was discussed in detail in [11], in particular providing bounds for its sample and computational complexity. That is the reason why we need to continue building more powerful AI supercomputers, which in some highly specific vertical applications already now surpass human capabilities. That way, when our infrastructure for learning from examples is more powerful, so is our capability to solve more complex problems in practice. The associated impact to society needs to be analyzed and if/when needed regulated as well, in parallel.

Government agencies and departments, whether small, local or part of a multi-agency board have common regulation demands and oversight requirements while attempting to streamline processes, improve the citizen experience, and safeguard the public well-being. Their tasks can go from local business licensing to financial institution regulation among multiple others. For example, to automate licensing and renewal at the state level, an state regulatory platform suite is necessary. Figure 5 shows Tyler's ETK Regulatory [12] that provides a flexible software with regulatory-specific features and functions for licensing, enforcement, and revenue management. Agencies can among others automate the entire regulatory lifecycle and enable online self-service for licenses and the public. Operational mandates expand beyond licensing and enforcement, thus an excepcionally flexible solution is required. Regulatory solutions can be tailored in the areas: professional, occupational, banking and securities, business, health, agriculture, alcohol, and education. Its entity-centric data model for maintaining complex relationships and the resulting database is shown togther with its shared business rules and workflows that include back office and public portal functions. Integrated modules include analytics-dashboards, interactive graphics, key performance indicators (KPIs), a reporter builder for standard and ad hoc reports, and customizable help. It has interfaces to exam and payment providers, accounting systems, domain- and state-specific imports and exports for its final utilization by agency users, agency admins, applicants and licensees as well as citizens in general.

ETK Regulatory is secure, configurable, scalable, extensible, easily maintainable based on entellitrak [13], a low-code application development platform for case management and business process management (BPM), see Figure 6 left for its modules, completely integrated, fully-featured solutions. It incorporates an agile, data-first approach to deliver an appropriate outcome for each case based on the information management of structured and unstructured data, collaboration, and guided decision making providing a continuous spectrum from case to business management solutions. Implementation acceleration can be achieved by starting with one of mutiple application accelerators, i.e., proven best practice solutions, provided. It offers a single, Section 508 compliant, web-based interface. Advanced application development and programming includes organizational and hierarchy modeling, rules management, security and permissions management, integrated de-

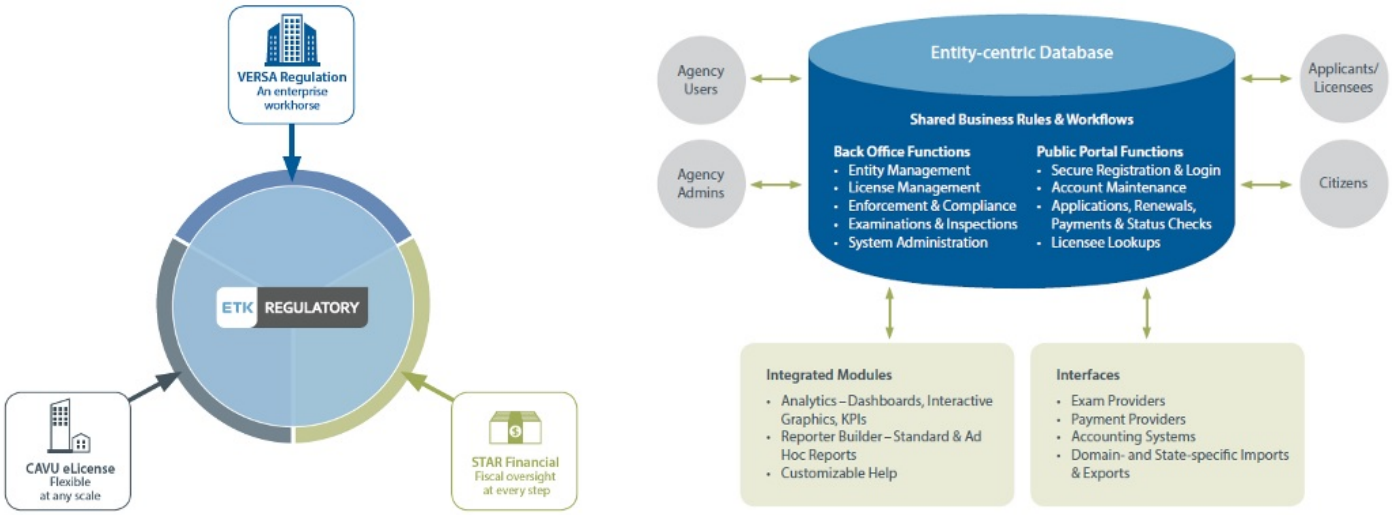

Figure 5: ETK Regulatory: VERSA Regulation, CAVU License, and STAR Financial

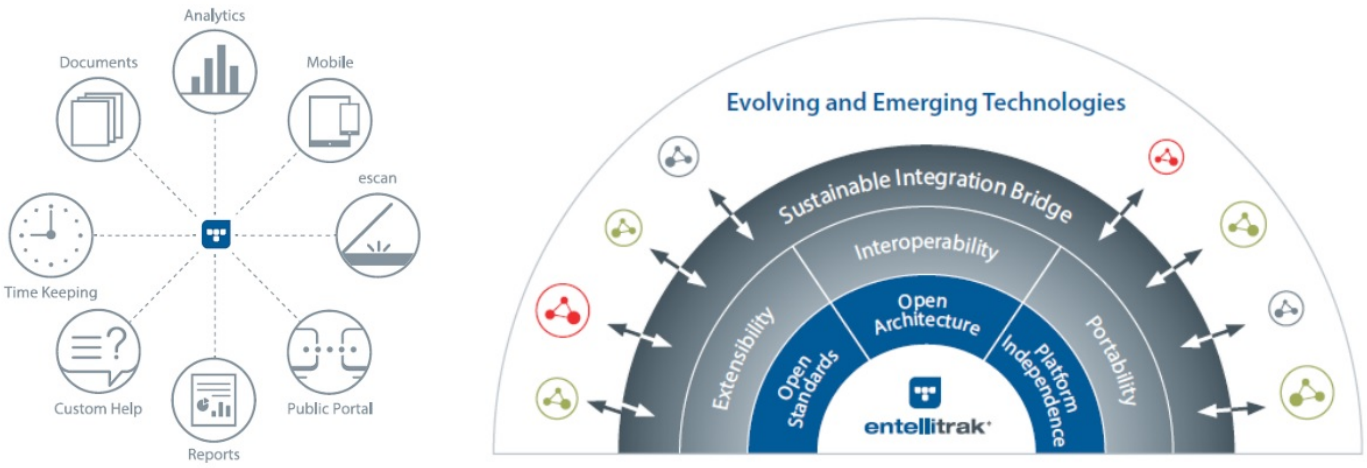

Figure 6: Entellitrak - A low-code application development platform

velopment environment (IDE), advanced search engine, forms and content management. Figure 6 right shows entellitrak's platform independent, open architecture to integrate and communicate with external systems to provide turnkey solutions. It is Java-based, compatible with virtually all relational database management systems (RDBMS) including SQL Server and Oracle. It can be hosted on-prem or in the cloud. Advanced security policies are built-in which include a role-based access, Security Sockets Layer (SSL)-based encryption of all data, completely configurable permissions through a Create/Read/Update/Delete user interface (CRUD UI), and single sign-on authentication via Lightweight Directory Access Protocol (LDAP), active directory, Rivest, Shamir, Adleman (RSA) SecurID tokens, authentication portals, as well as smart cards and Common Access Cards (CACs).

A state regulatory platform suite is shown in Figure 7 left as an example of a government regulatory solution. It can serve for from issuing business licenses to providing oversight of public service commissions. To work effectively, this solution meets the requirements for flexible work flows, multiple communication options, and efficient collaboration. Functionalities required can be developed precisely to protect the public and provide an efficient service to the industries that are being regulated. Services can be provided to the public, applicants and licensees regardless whether they are on a computer, tablet, or smart phone due to its mobile-responsive design, see Figure 7 right. In this report, the state of the art in Artificial Intelligence research and technology development as well as rationales and technologies for government regulation at all levels: local, state, and federal

Figure 7: Examples include a State Regulatory Platform Suite, mobile-responsive online portal

are summarized. Some initial planned standards and policy considerations are included in [16] and in [17], respectively. Focus on AI standards include concepts and terminology, data and knowledge, human interactions, metrics, networking, performance testing and reporting methodology, safety, risk management, and trustworthiness. The last area requires guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security. The Whitehouse has recently unveiled a so called AI bill of rights. Some argue a tougher AI regulation as the one outlined is required [18]. The Office of Science and Technology Policy (OSTP) at the White House recently published the blueprint for an AI Bill of Rights [19] that identified five principles to guide the design, use, and deployment of automated systems to protect the American public in the age of AI. They are called safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives, consideration, and fallback.

Pros and cons of underying principles and concrete potential regulatory measures are discussed towards the potential near-term implementation of AI regulation that protects AI innovation and provide safeguards in this exciting field with so many applications and further promise and on the other hand, severe fears already existent in the general population and tech practitioners and executives as the technology allows for more and more intelligent capabilities to be performed by machines which were thought of until very recently as only being able to be performed by humans. Fears, dangers to the public are based on the irregulated use of the citizens' data to train AI models taken from the Internet, the potential mass layoffs that the further introduction of AI technology may cause, the proliferation of AI-enabled fraud, and the dissemination of huge volumes of misinformation including for political purposes, among others. As is already partially evidenced in practice in all aforementioned areas, a challenging appropriate liegislation appears to be due for which the U.S. Congress might not be well prepared yet.

References

[1] M. Rounds. *Rounds moderates historic forum with major tech leaders*. U.S. Senate, Artificial Intelligence (AI) working group, September 13, 2023. https://www.rounds.senate.gov/newsroom/press-releases/rounds-moderates-historic-forum-with-major-tech-leaders.

[2] K. McCarthy and H. Jeffries. *Speaker McCarthy and leader Jeffries co-host fifth congressional Hackaton*. U.S. House of Representatives, September 14, 2023. https://www.speaker.gov/speaker-mccarthy-and-leader-jeffries-co-host-fifth-congressional-hackathon/

[3] V.D. Sánchez. *Computational Data Science Research and Technology Development – State of the Art –*. January 2023.
https://profdrvdsaphd.lima-city.de/documents/ComputationalDataScience.pdf

[4] Nvidia. *NVIDIA DGX GH200 AI Supercomputer – AI Supercomputer for the Generative AI Era –*. White Paper, WP-11400-001_v01, June 2023.
https://resources.nvidia.com/en-us-dgx-gh200/technical-white-paper

[5] Nvidia. *NVIDIA GH200 Grace Hopper Superchip*. Datasheet, August 2023.

[6] V.D. Sánchez. *A Parallel Distributed Architecture for Vision and Telerobotics (in German)*, in M. Baumann and R. Grebe (Eds.), Parallele Datenverarbeitung mit dem Transputer, Berlin, Springer-Verlag, Reihe Informatik aktuell, 295-303, 1993.

[7] V.D. Sánchez et al. *The Design of a Real-Time Neurocomputer Based on RBF Networks*. Neurocomputing, 20, 111-114, 1998.

[8] V.D. Sánchez. *Neurocomputers in Industrial Applications (in German)*. Technische Rundschau, 82 [25], 60-65, 1990.

[9] V.D. Sánchez. *ANSpec – A Specification Language (in German) –*. CHIP Plus, 7, 17-22, 1990.

[10] V.D. Sánchez. *Neurocomputing 50th volume anniversary*. Neurocomputing, 50, ix, 2003.

[11] V.D. Sánchez. *Searching for a solution to the automatic RBF network design problem*. Neurocomputing, 42 [1-4], 147-170, 2002.

[12] Tyler Technologies, MicroPact. *ETK Regulatory – The most flexible regulatory licensing and enforcement software on the market–*.
https://www.tylertech.com/products/ETK-Regulatory/ETK-Regulatory-Brochure.pdf

[13] Tyler Technologies, MicroPact. *entellitrak – A low-code application development platform for case management and BPM*.
https://www.tylertech.com/Products/TylerEntellitrak/Tyler-Entellitrak-Brochure.pdf

[14] Tyler Technologies. *Enterprise State Regulatory*.
https://www.tylertech.com/resources/resource-downloads/enterprise-state-regulatory-application-brochure

[15] Tyler Technologies. *State Regulatory Platform Suite*.
https://www.tylertech.com/resources/resource-downloads/State-Regulatory-Platform-Suite-Brochure

[16] National Institute of Standards and Technology. *U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools*. August 9, 2019.

[17] Congressional Research Service. *Artificial Intelligence: Background, Selected Issues, and Policy Considerations*. CRS Report R46795, May 19, 2021.

[18] M. Heikkilä. *The White House just unveiled a new AI Bill of Rights – It's the first big step to hold AI to account*. MIT Technology Review, October 4, 2022.

[19] The White House, Office of Science and Technology Policy (OSTP). *Blueprint for an AI Bill of Rights*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/