UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF TEXAS
MARSHALL DIVISION

PARTEC AG and BF EXAQC AG, Plaintiffs, vs. MICROSOFT CORPORATION, Defendant.	Civil Action No. 2:24cv433 **JURY TRIAL DEMANDED**

COMPLAINT FOR PATENT INFRINGEMENT

This is an action for patent infringement in which Plaintiffs ParTec AG ("ParTec") and BF exaQC AG ("BFX") (collectively "Plaintiffs") make the following allegations against Defendant Microsoft Corporation ("Microsoft" or "Defendant") for infringing the Patents asserted in this matter.

PARTIES

1. Plaintiff ParTec AG is a German corporation with its principal place of business at Possartstraße 20, 81679 München, Germany.

2. Plaintiff BF exaQC AG is a German corporation with its principal place of business at Südliche Münchner Straße 56, 82031 Grünwald, Germany.

3. Defendant Microsoft is a Delaware corporation with a principal place of business at One Microsoft Way, Redmond, WA 98052. Microsoft has been registered to do business in the State of Texas since March 13, 1995, and may be served with process via its registered agent: Corporation Service Company d/b/a CSC - Lawyers Incorporating Service Company, 211 E. 7th Street, Suite 620, Austin, TX 78701.

JURISDICTION AND VENUE

4. This Court has subject matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the patent laws of the United States, 35 U.S.C. §§ 1 *et seq.*

5. This Court has personal jurisdiction over Microsoft because Microsoft conducts business in and has committed acts of patent infringement in this District and the State of Texas and has established minimum contacts with this forum state such that the exercise of jurisdiction over Microsoft would not offend traditional notions of fair play and substantial justice. Venue is also proper in this District pursuant to 28 U.S.C. § 1400(b) because Microsoft has regular and established physical places of business in this District and has committed acts of patent infringement in the District.

6. Among other things, Microsoft has seven corporate offices in the State of Texas, employing hundreds of persons. Microsoft represents that one of those offices is in Frisco, Texas, within this District.

Microsoft U.S. office locations

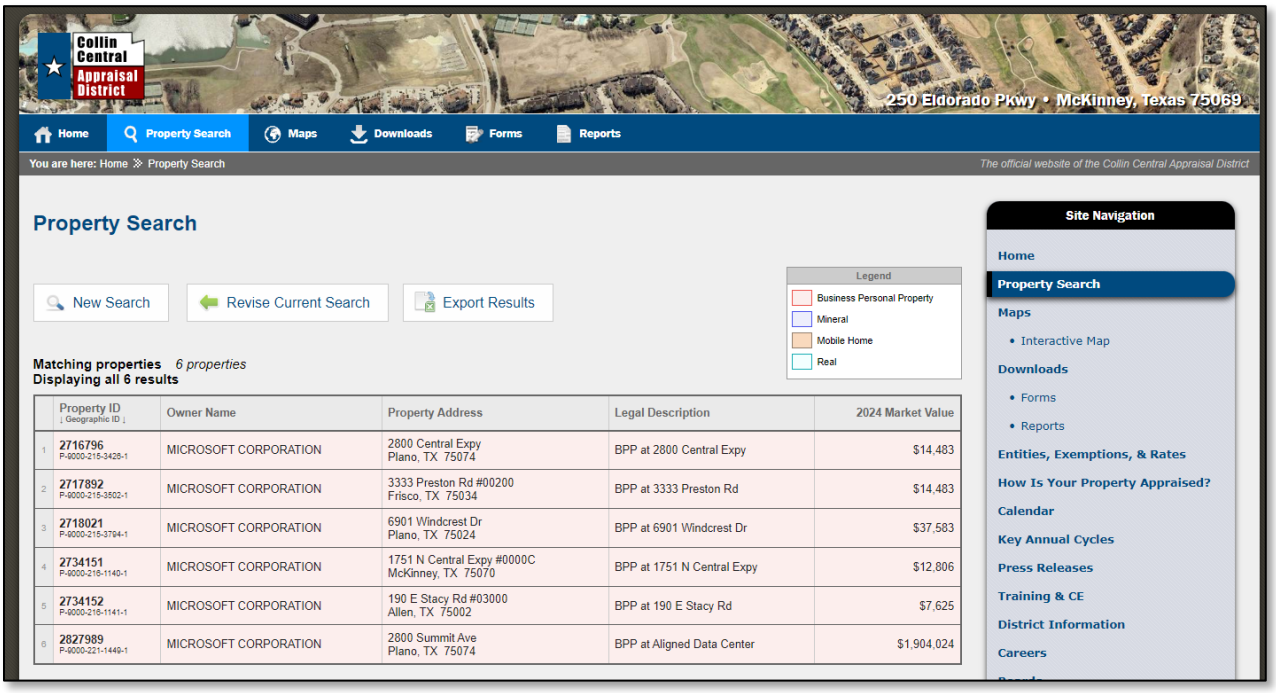
Microsoft reaches customers at sales offices, support centers and technology centers throughout the country. Use the clickable map or the location links for more information.

Texas

Austin
Houston
San Antonio
Dallas
Friendswood
Frisco
The Woodlands

Source: Microsoft, *About*, https://www.microsoft.com/en-us/about/officelocator/all-offices (last accessed June 7, 2014)

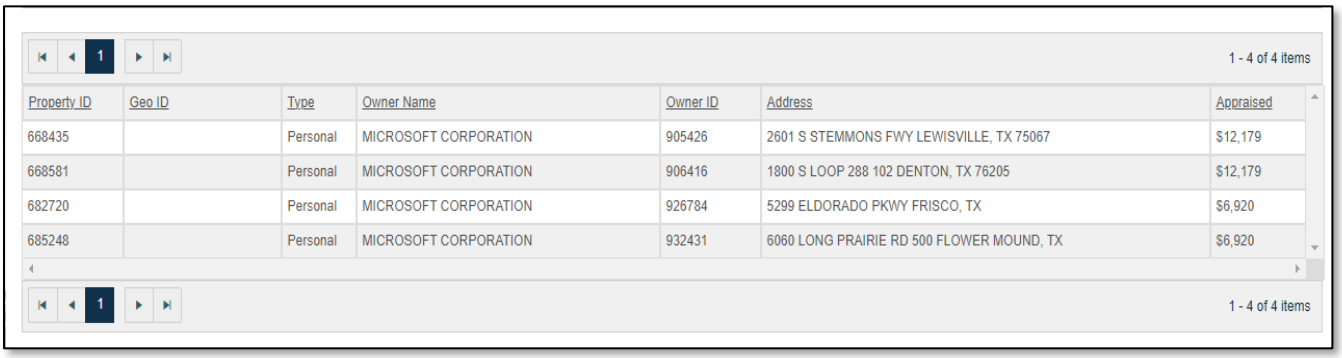
7. In addition, Microsoft maintains millions of dollars of business personal property in Collin County, within this District:



	Property ID ↓ Geographic ID ↓	Owner Name	Property Address	Legal Description	2024 Market Value
1	2716796 P-9000-215-3426-1	MICROSOFT CORPORATION	2800 Central Expy Plano, TX 75074	BPP at 2800 Central Expy	$14,483
2	2717892 P-9000-215-3502-1	MICROSOFT CORPORATION	3333 Preston Rd #00200 Frisco, TX 75034	BPP at 3333 Preston Rd	$14,483
3	2718021 P-9000-215-3794-1	MICROSOFT CORPORATION	6901 Windcrest Dr Plano, TX 75024	BPP at 6901 Windcrest Dr	$37,583
4	2734151 P-9000-216-1140-1	MICROSOFT CORPORATION	1751 N Central Expy #0000C McKinney, TX 75070	BPP at 1751 N Central Expy	$12,806
5	2734152 P-9000-216-1141-1	MICROSOFT CORPORATION	190 E Stacy Rd #03000 Allen, TX 75002	BPP at 190 E Stacy Rd	$7,625
6	2827989 P-9000-221-1449-1	MICROSOFT CORPORATION	2800 Summit Ave Plano, TX 75074	BPP at Aligned Data Center	$1,904,024

Source: Collin Central Appraisal District, *Property Search*, https://www.collincad.org/propertysearch (last accessed June 7, 2024) (search results for "Microsoft")

8. Microsoft similarly maintains significant business personal property in Denton County, within this District:



1 - 4 of 4 items

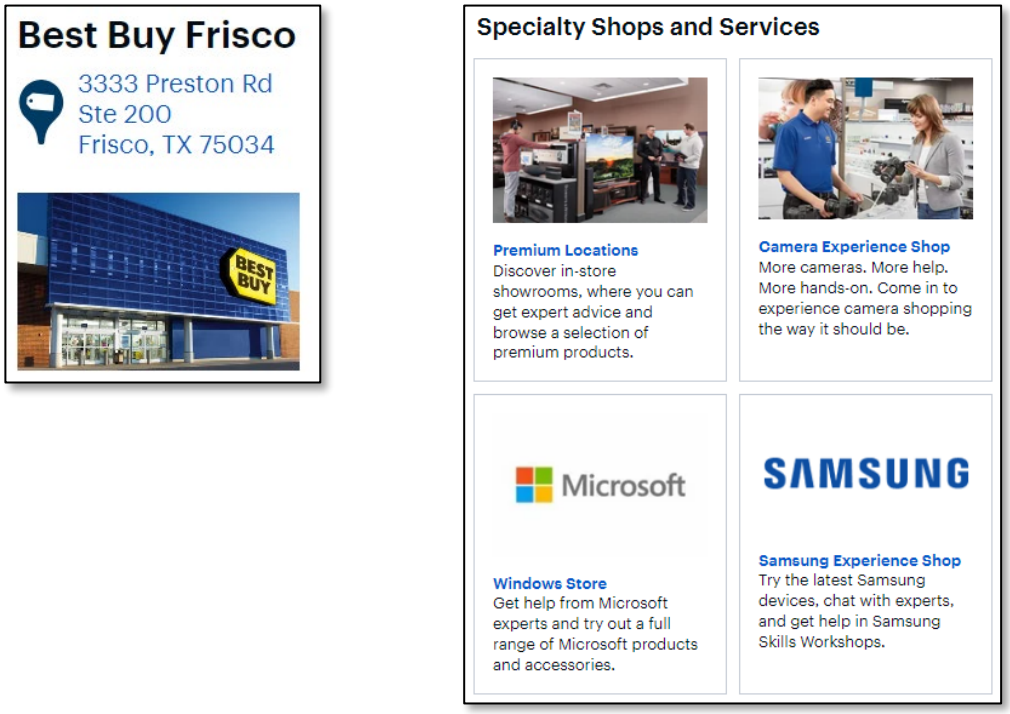
Property ID	Geo ID	Type	Owner Name	Owner ID	Address	Appraised
668435		Personal	MICROSOFT CORPORATION	905426	2601 S STEMMONS FWY LEWISVILLE, TX 75067	$12,179
668581		Personal	MICROSOFT CORPORATION	906416	1800 S LOOP 288 102 DENTON, TX 76205	$12,179
682720		Personal	MICROSOFT CORPORATION	926784	5299 ELDORADO PKWY FRISCO, TX	$6,920
685248		Personal	MICROSOFT CORPORATION	932431	6060 LONG PRAIRIE RD 500 FLOWER MOUND, TX	$6,920

1 - 4 of 4 items

Source: Denton CAD, *Property Search*, https://esearch.dentoncad.com/ (last accessed June 7, 2024) (search results for "Microsoft")

9. For example, Microsoft operates Microsoft Windows Stores within Best Buy retail locations located throughout this District. The following are three examples of such stores: 2800 N Central Expy, Plano, TX 75074; 3333 Preston Rd Suite 200, Frisco, TX 75034; and 2601 S Stemmons Fwy, Ste 300, Lewisville, TX 75067.

Source: Best Buy, *Locations*, https://stores.bestbuy.com/tx/plano/2800-n-central-expy-202.html (last accessed June 7, 2024) (showing "Windows Store" at Plano Best Buy)



Source: Best Buy, *Locations*, (last accessed June 7, 2024) https://stores.bestbuy.com/tx/frisco/3333-preston-rd-180.html (showing "Windows Store" at Frisco Best Buy)

Source: Best Buy, *Locations*, https://stores.bestbuy.com/tx/lewisville/2601-s-stemmons-fwy-258.html (last accessed June 7, 2024) (showing "Windows Store" at Lewisville Best Buy)

10. The Microsoft Windows Stores operated by Microsoft within Best Buy stores are regular and established places of business for Microsoft. Microsoft rents the space. They are, as Microsoft itself touts, Microsoft stores within Best Buy, or a "store-within-a-store." *See* Microsoft, *Talking Retail: The New Windows Store Only at Best Buy* (June 13, 2013), https://blogs.windows.com/windowsexperience/2013/06/13/talking-retail-the-new-windows-store-only-at-best-buy/ ("Today, we announced a strategic partnership to create the Windows Store only at Best Buy, a comprehensive store-within-a-store in 500 Best Buy locations across the United States and more than 100 Best Buy and Future Shop locations in Canada. The stores within Best Buy will range in size from 1,500 square feet to 2,200 square feet and will be the premier destination for consumers to see, try, compare and purchase a range of products and accessories . . . ."); Thomas Lee, *Store Within a Store Concept*, Minneapolis Star Tribune (July 14, 2013), https://www.startribune.com/best-buy-bets-big-on-store-within-store-concepts/215301161/ ("Microsoft and Samsung are essentially leasing their spaces from Best Buy . . . .").

11. Microsoft is responsible for and controls the day-to-day operations of such stores. Microsoft is responsible, *inter alia*, for its "own pricing and merchandise." Thomas Lee, *Store*

Within a Store Concept, Minneapolis Star Tribune (July 14, 2013), https://www.startribune.com/best-buy-bets-big-on-store-within-store-concepts/215301161/. Microsoft employs Microsoft "Specialists" to "manage and support the training, merchandising, events, and operations of the Microsoft product ecosystem within Best Buy." Microsoft, *Careers*, https://jobs.careers.microsoft.com/global/en/job/1622416/Partner-Stores-Specialist (last accessed June 7, 2024); Microsoft, *Careers*, https://jobs.careers.microsoft.com/us/en/job/1385093/ (last accessed June 7, 2024). They "[m]aintain Microsoft merchandising standards in accordance with Microsoft brand guidelines." *Id.* Stated differently, they "[s]upport and manage the Microsoft business for up to 5 [Best Buy] stores; including aligning training and other store business needs." *Id.* Microsoft also employs "Partner Activations & Readiness Leads" who "support[] the in-store Windows Store Specialists." Microsoft, *Careers*, https://jobs.careers.microsoft.com/us/en/job/1417570/Partner-Activations-Readiness-Lead (last accessed June 7, 2024). These individuals "[d]eliver store design updates," "[e]nsure proper planning, prototype, shakedown, and training steps are taken to deliver near-flawless execution for large-scale transformations," and "[p]rovide operational support to field team with store list management, ordering, and replenishment of supplies, training sessions and mentorship." *Id.*

12. Besides maintaining Microsoft Windows Stores within Best Buy retail locations, Microsoft has approximately $2 million of property at Aligned Data Center, at 2800 Summit Ave, Plano, TX 75074, within this District.

Source: Google Street View of 2800 Summit Ave, Plano, TX 75074

13. Microsoft maintains data servers at this location.

Source: Dave Montgomery, *Texas Lures Data Centers, Not for Jobs but for Revenue*, N.Y. Times (April 26, 2016) ("a customizable customer pod containing servers at Aligned Data Centers in Plano, Tex.")

14. As detailed in later sections, this case accuses the Microsoft Azure AI system of infringement. There is a Microsoft point of presence (POP) location[1] for the Azure network in Plano, Texas, within this District (likely at the data center discussed in the preceding paragraphs).

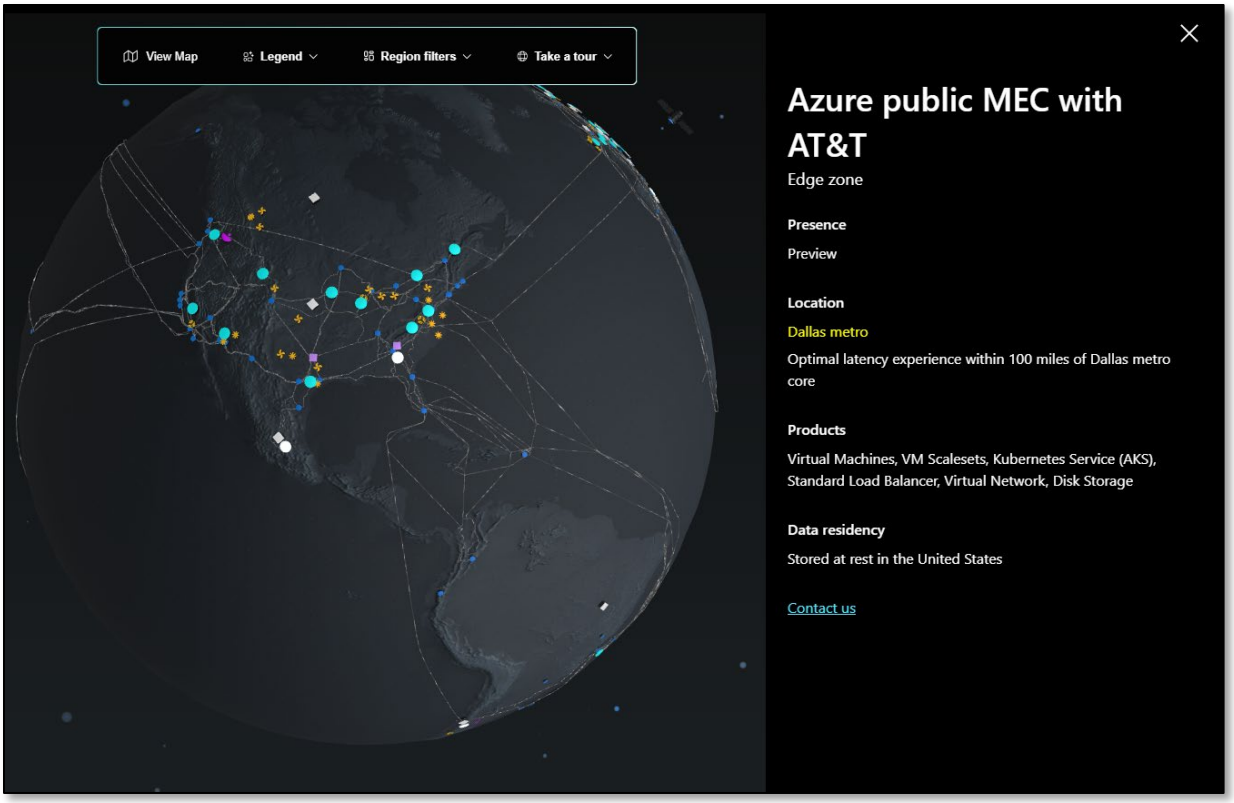
[1] POPs are part of content delivery networks—"a distributed network of servers that can efficiently deliver web content to users. A content delivery network store[s] cached content on edge servers in point of presence (POP) locations that are close to end users, to minimize latency." Microsoft, *What is a Content Delivery Network on Azure?*, https://learn.microsoft.com/en-us/azure/cdn/cdn-overview, (last accessed June 9, 2024).

Expand table

Region	Cities
North America	Etobicoke, Canada (2) Montreal, Canada Vancouver, Canada (2) Querétaro, Mexico (2) Atlanta, GA, USA (3) Boydton, VA, USA (2) Chaska, MN, USA (2) Cheyenne, WY, USA (2) Chicago, IL, USA (4) Dallas, TX, USA (4) Des Moines,IA, USA (3) Detroit, MI, USA Englewood, CO, USA (2) Honolulu, HI, USA Houston, TX, USA (3) Jacksonville, FL, USA (2) Las Vegas, NV, USA (3) Los Angeles, CA, USA (2) Manassas, VA, USA (3) Memphis, TN, USA Miami, FL, USA (4) Minneapolis, MN, USA (2) Needham Heights, MA, USA (2) Nashville, TN, USA Newark, NJ, USA New York, NY, USA (2) Philadelphia, PA, USA Phoenix, AZ, USA Plano, TX, USA Portland, OR, USA Quincy, WA, USA (3) San Antonio, TX, USA (4) San Jose, CA, USA (4) Salt Lake City, UT, USA (3) Seattle, WA, USA (2) Secaucus, NJ, USA (2) Southfield, MI, USA

Source: Microsoft, *Azure Learn: Azure Content Delivery Network Coverage by Metro*, https://learn.microsoft.com/en-us/azure/cdn/cdn-pop-locations (last accessed June 7, 2024)

15. Microsoft, as shown below, also lists the "Dallas metro" area as the location for an Azure public MEC site. Microsoft describes these site as follows: "Azure public multi-access edge compute (MEC) sites are small-footprint extensions of Azure. They're placed in or near mobile operators' data centers in metro areas, and are designed to run workloads that require low latency while being attached to the mobile network. . . . Azure public MEC provides secure, reliable, high-bandwidth connectivity between applications that run close to the user while being served by the Microsoft global network." Microsoft, *What is Azure Public MEC?*, https://learn.microsoft.com/en-us/azure/public-multi-access-edge-compute-mec/overview (last accessed June 8, 2024).



Source: Microsoft, *Microsoft Datacenters*, https://datacenters.microsoft.com/globe/explore?info=edge_dallas (last accessed June 8, 2024)

16. Microsoft further has a 470,000 square foot Azure data center at 5150 Rogers Rd., San Antonio, TX.

Source: DataCenters.com, *Microsoft Azure: South Central US-Texas* https://www.datacenters.com/microsoft-azure-south-central-us-texas (last accessed June 7, 2024)

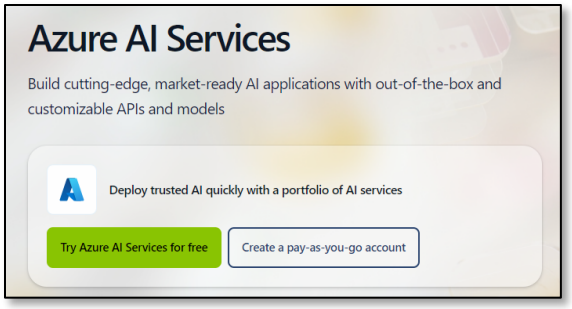
17. In fact, Microsoft's "South Central US" Azure region is centered in Texas and has been since 2008.

Brazil	Canada	Chile	Mexico	**United States**	Azure Government

ast US 3	North Central US	South Central US
oming soon	Start free >	Start free >
Georgia	Illinois	Texas
Coming soon	2009	2008
Coming soon	Coming soon	Available with 3 zones
Azure compliance offerings	Azure compliance offerings	Azure compliance offerings
Stored at rest in the United States Learn more	Stored at rest in the United States Learn more	Stored at rest in the United States Learn more
Coming soon	Cross-region options: Azure Site Recovery Region Pairing	Cross-region options: Azure Site Recovery Region Pairing
Coming soon	See products in this region	See products in this region
Coming soon	All customers and partners	All customers and partners

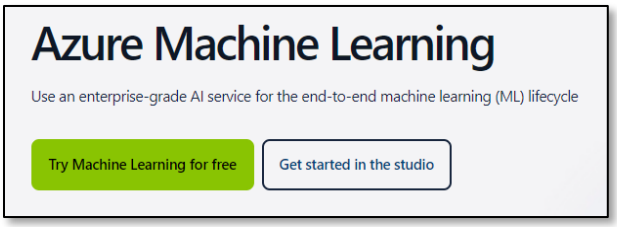
Source: Microsoft, *Azure Geographies*, https://azure.microsoft.com/en-us/explore/global-infrastructure/geographies/#geographies (last accessed June 7, 2024)

18. Beyond purposefully locating infringing hardware in the State of Texas and this District, Microsoft, directly and/or through subsidiaries and agents (including distributors, retailers, and others), makes, imports, ships, distributes, offers for sale, sells, uses, and advertises (including offering products and services through its websites) its Azure AI services and products in the United States, the State of Texas, and this District. For example, Microsoft, through its website, purposefully and knowingly offers and sells its Azure AI services—which run on and rely on its Azure AI system and infrastructure—to customers within this District:



Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 8, 2024, in Longview, Texas)

19. As another example, Microsoft, through its website, purposefully and knowingly sells and offers its Azure Machine Learning services—which also use the Microsoft Azure AI infrastructure—to customers within this District:



Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed on June 8, 2024, in Longview, Texas)

20. At minimum, Microsoft, directly and/or through its subsidiaries and agents (including distributors, retailers, and others), has purposefully and voluntarily put its Azure AI services and products into the stream of commerce with the expectation that they will be purchased and used by consumers in this District in an infringing manner. These infringing products and/or services have been and continue to be purchased and used by consumers in this District.

21. Finally, Microsoft last year announced a multi-billion-dollar deal with specialist cloud provider CoreWeave to use its datacenters for some of its Azure AI workloads. *See* Sebastian Moss, *CoreWeave Plans $1.6bn AI Cloud Data Center in Plano, Texas*, DCD (July 25, 2023) https://www.datacenterdynamics.com/en/news/coreweave-plans-16bn-ai-cloud-data-center-in-

plano-texas/; Sebastian Moss, *Microsoft Signs Multi-Billion Dollar Deal with GP Cloud Provider CoreWeave to Meet AI Needs*, DCD (June 2, 2023) https://www.datacenterdynamics.com/en/news/microsoft-signs-multi-billion-dollar-deal-with-gpu-cloud-provider-coreweave-to-meet-ai-needs/. One of those datacenters is a $1.6 billion datacenter in Plano, within this District. *See* Sebastian Moss, *CoreWeave Plans $1.6bn AI Cloud Data Center in Plano, Texas*, DCD (July 25, 2023) https://www.datacenterdynamics.com/en/news/coreweave-plans-16bn-ai-cloud-data-center-in-plano-texas/. Microsoft also recently announced a $1.5 billion investment in the Condor Galaxy 3 AI supercomputer being built in Dallas, Texas, by the Abu Dhabi, United Arab Emirates-based technology holding group G42.

22. Thus, Microsoft is subject to this Court's general and specific jurisdiction pursuant to due process and/or the Texas Long Arm Statute due at least to Microsoft's substantial business in the State of Texas and this District, including through its past and ongoing infringing activities, because Microsoft regularly does and solicits business herein, and/or Microsoft has engaged in persistent conduct and/or has derived substantial revenues from goods and services provided in the State of Texas and this District.

23. Venue is likewise proper in this District pursuant to 28 U.S.C. § 1400(b) because Microsoft has regular and established physical places of business in this District and has committed acts of patent infringement in the District.

PARTEC AND BFX

24. Plaintiff ParTec is a leading provider of modular supercomputers, quantum computers, and software for modular computing systems. ParTec was founded in 1999 and is based in Munich, Germany.

25. ParTec specializes in the development and manufacture of modular supercomputers and quantum computers, as well as accompanying system software. Its services include the distribution of future-oriented high-performance computers (HPC) and quantum computers (QC), as well as consulting and support services in all areas of development, construction, and operation of these advanced systems. The approach of modular supercomputing represents a unique selling point for ParTec.

26. ParTec has been deeply involved with the European high-performance computing and quantum computer research community from the beginning, and its developments and inventions are being used all over the world. For example, ParTec is the lead partner in the construction of the first Exascale[2] supercomputer in Europe: the JUPITER Exascale Supercomputer. ParTec is responsible for procurement, delivery, installation, hardware, software, and maintenance of JUPITER. The supercomputer will be built with the dynamic Modular System Architecture (dMSA) developed and patented by ParTec.

[2] Exascale computing is a level of supercomputing capable of at least one exaflop floating point calculations per second to support expansive workloads. An exaflop is a measure of performance for a supercomputer that can calculate at least 10^18 or one quintillion floating point operations per second. In exaflop, the exa- prefix means a quintillion—*i.e.*, a billion billion.

27. When completed, Jupiter will have three times the computing capability of Europe's current most powerful supercomputer and will provide the equivalent power of 10 million modern desktop computers. The overall system will require the space of about four tennis courts and will use over 260 km of high-performance cabling, allowing it to move over 2,000 Terabits per second, the equivalent of 11,800 full copies of Wikipedia every second. The total JUPITER order is approximately €300 million.

28. ParTec also developed QBridge in collaboration with Quantum Machines. QBridge is a software solution that enables seamless integration of high-performance and quantum computers.

29. In addition, ParTec is actively working on expanding its Parastation Modulo software, used in modular supercomputers. The latest expansion—Parastation Modulo 2.0—aims to bridge the gap to embed quantum computers into modular supercomputers.

30. In November 2023, industry-publication HPCWire named the MareNostrum 5 pre-exascale computer set up by ParTec AG (together with Eviden) as the best development of the preceding twelve months in the field of high-performance computing. With a performance capacity of up to 314 petaflops—*i.e.*, at least 314 million billion calculations per second—MareNostrum 5 is one of the most powerful supercomputers in the world and is tailored to support medical research in Europe through the development of drugs and vaccines, simulations of virus propagation, artificial intelligence applications, and the processing of large amounts of data.

31. ParTec recently announced the development of its first quantum computer product line, named EIGER. The system is based on superconducting qubit[3] technology and can be scaled

[3] "A qubit, or quantum bit, is the basic unit of information used to encode data in quantum computing and can be best understood as the quantum equivalent of the traditional bit used by classical computers to encode information in binary." IBM, *What is a Qubit?*,

from small qubit counts to advanced quantum processing unit (QPU)[4] technologies. This flexibility allows research organizations, data centers, and industrial customers to carry out work ranging from early quantum exploration to advanced research.

32. Finally, ParTec recently announced plans to construct a production facility for quantum computers in the Greater Munich area. The "ParTec Quantum Factory" is expected to start operations in the second half of 2024. ParTec will initially invest five million euros in the construction of a production facility for assembly and testing of cryogenic and non-cryogenic systems.

33. ParTec's achievements and innovation in supercomputing has not only enabled commercial success, but also resulted in an expansive patent portfolio containing over 150 patents and patent families.

34. Plaintiff BFX is based in Grünwald, Germany, and is ParTec's licensing agent.

35. Under an August 29, 2021 "Exclusive License Agreement" ParTec granted BFX a perpetual, exclusive license to ParTec's patents to use and sublicense the licensed rights in the following field: the development, manufacture and sale of microchips and processors. Excluded were rights related to the application of the system architecture of supercomputers (high performance computing), including cloud computing, and the integration of quantum computers in HPC environments. Under a November 11, 2022 "Agency Agreement," ParTec made BFX its licensing agent to license ParTec's patent rights in the field of computer system architecture, high performance computing, and cloud computing.

---

https://www.ibm.com/topics/qubit (last accessed June 7, 2024).

[4] "A quantum processing unit, or QPU, uses qubits and quantum circuit model architecture to solve problems that are too computationally intensive for classical computing." Yuval Boger, *Why the QPU Is the Next GPU*, BuiltIn (Mar. 8, 2024), https://builtin.com/articles/quantum-processing-unit-qpu.

36. Under a December 21, 2023 "Contribution Agreement," ParTec contributed to its subsidiary, FL Systems AG & Co. KG, an "Exclusive License Agreement." The Exclusive License Agreement—entered the same day between ParTec and FL Systems AG & Co. KG—granted FL Systems AG & Co. KG a perpetual, exclusive license to ParTec's patents, limited to the field of application of the system architecture of supercomputers (High Performance Computing), including cloud computing and the integration of quantum computers in HPC environments. Contemporaneously, FL Systems AG & Co. KG entered an "Agency Agreement" with BFX. FL Systems AG & Co. KG and BFX then entered an exclusive sublicensing agreement for the Asserted Patents. Under those agreements, BFX is responsible for negotiating and entering license agreements, and has the exclusive right to negotiate and enter license agreements, for the Asserted Patents—in the field of computer system architecture, high performance computing, and cloud computing. BFX is also responsible for preparation of patent applications, maintenance of property rights, and enforcement of patent rights, including the right to seek future damages and past damages incurred by FL Systems AG & Co. KG—again, in the field of computer system architecture, high performance computing, and cloud computing.

37. Under a separate December 21, 2023 "Contribution Agreement," BFX transferred, via contribution, the August 29, 2021 "Exclusive License Agreement" between ParTec and BFX to FL Chips AG & Co. KG, a BFX subsidiary. Contemporaneously, FL Chips AG & Co. KG entered an "Agency Agreement" with BFX. FL Chips AG & Co. KG and BFX then entered an exclusive sublicensing agreement for the Asserted Patents. Under those agreements, BFX is responsible for negotiating and entering license agreements, and has the exclusive right to negotiate and enter license agreements, for the Asserted Patents—in the field of the development and manufacture of microchips and processors. BFX is also responsible for preparation of patent

applications, maintenance of property rights, and enforcement of patent rights, including the right to seek future damages and past damages incurred by FL Chips AG & Co. KG—again, in the field of the development and manufacture of microchips and processors.

38. Thus, BFX is currently responsible for managing, licensing, and enforcing the Asserted Patents, with the exclusive rights to do the same.

THE ASSERTED PATENTS

39. This complaint asserts causes of action for infringement of United States Patent No. 10,142,156 (the "'156 Patent"), United States Patent No. 11,934,883 (the "'883 Patent"), and United States Patent No. 11,537,442 (the "'442 Patent") (collectively, the "Asserted Patents").

40. On November 27, 2018, the U.S. Patent and Trademark Office duly and legally issued the '156 Patent, which is entitled "Computer Cluster Arrangement for Processing a Computation Task and Method for Operation Thereof." A true and correct copy of the '156 Patent is attached as **Exhibit A**.

41. On March 19, 2024, the U.S. Patent and Trademark Office duly and legally issued the '883 Patent, which is entitled "Computer Cluster Arrangement for Processing a Computation Task and Method for Operation Thereof." The '883 Patent is designated a continuation of the application that resulted in the '156 Patent. A true and correct copy of the '883 Patent is attached as **Exhibit B**.

42. The '156 Patent and '883 Patents generally claim a computer cluster-booster system for processing a computation task and a method for operating the introduced computer cluster-booster system. The claimed system comprises computation nodes that dynamically outsource specific computation tasks to boosters.

43. Dr. Thomas Lippert is the sole inventor of the '156 and '883 patents. Dr. Lippert

received his diploma in Theoretical Physics in 1987 from the University of Würzburg. He completed his Ph.D. theses in theoretical physics at Wuppertal University on simulations of lattice quantum chromodynamics and at Groningen University in the field of parallel computing with systolic algorithms.

44. Dr. Lippert is a world-renowned leader in modular supercomputing and cluster computing. He is the head of the Jülich Supercomputing Center—one of the premier high-performance computing centers in the world with approximately 250 experts working on all aspects of supercomputing and simulation. Dr. Lippert is also a Director at the Institute of Advanced Simulation in Germany, and he holds the chair for Modular Supercomputing and Quantum computing at Goethe University in Frankfurt, where he explores the development and practical application of modular supercomputers and quantum computers. Dr. Lippert's articles have been cited over 16,000 times according to Google Scholar, and he was named to HPCWire's People to Watch in both 2022 and 2010.

45. ParTec and Dr. Lippert have worked together collaboratively for decades. Dr. Lippert was an early ParTec customer with the launch of the ALICE supercomputer while Dr. Lippert was a member of the Department of Physics of the University of Wuppertal, in Germany. ParTec delivered the Cluster Operating System, ParaStation. After being awarded the position of Director of the Jülich Supercomputing Center, Dr. Lippert invited ParTec to cooperate on the further development of ParaStation and on the co-design and co-development of supercomputers with cluster architecture. This cooperation led to the Supercomputer JUROPA 2 in 2009 at the JSC, which reached the number 10 position on the TOP500 list of the fastest supercomputers worldwide. In 2017, Dr. Lippert invited ParTec to build the Supercomputer JURECA at JSC, proving concept of the modular supercomputing architecture on large scale,

followed by the modular supercomputer JUWELS in 2018. ParTec will build the modular exascale supercomputer JUPITER at Dr. Lippert's hosting site—the Jülich Supercomputing Center—in the coming year.

46. On December 27, 2022, the U.S. Patent and Trademark Office duly and legally issued the '442 Patent, which is entitled "Application Runtime Determined Dynamical Allocation of Heterogenous Compute Resources." The '442 Patent generally claims a heterogeneous computing system and a method for operating the introduced system, comprising computation nodes and booster nodes arranged to compute a computation task with subtasks, where subtasks are initially assigned to and processed by certain computation nodes and booster nodes in the system and then information relating to that processing is used to further distribute computation tasks. A true and correct copy of the '442 Patent is attached as **Exhibit C**.

47. Dr. Lippert is one of the inventors of the '442 Patent. Bernhard Frohwitter—who is also well-known in the supercomputing field—is the other inventor. Mr. Frohwitter holds a degree in mechanical engineering and is an attorney at law, admitted to the German bar. Since 2004, Mr. Frohwitter has been the Majority Shareholder, Managing Director, and Chairman of the Board at ParTec. Mr. Frohwitter also currently serves at its Chief Executive Officer.

48. Each of the Asserted Patents was assigned to ParTec Cluster Competence Center GmbH, which changed its corporate form and name on March 4, 2021, becoming ParTec AG. Since assignment, ParTec has owned and continues to own each of the Asserted Patents. ParTec is the sole owner of the Asserted Patents.

49. BFX has the exclusive right to license the patents to third parties and enforce the Asserted Patents.

50. Collectively, Plaintiffs hold all rights and title to the Asserted Patents, including the

sole and exclusive right to bring claims for infringement.

THE INVENTIONS OF THE ASSERTED PATENTS

51. The Asserted Patents generally teach systems and methods for cluster computing. Specifically, they teach how to use dynamic resource management so that compute tasks are assigned dynamically throughout cluster computers to help increase utilization, reduce required power, and maximize computing speed.

52. Before the inventions described in the patents, general-purpose processors were supplemented by permanently assigned accelerators. But the fixed assignment of accelerators regularly resulted in over or under-allocation of resources. The resulting shortage or oversupply of resources was costly and energy inefficient. Ex. A ('156) at 1:36–46.

53. ParTec's inventions solved those problems by providing a modular computing system that combines different types of computing hardware, such as processing units and accelerators, and makes them freely and dynamically assignable to each other. Ex. A ('156) at 2:21–25. The inventions also teach systems and methods for dynamically assigning tasks between processing units and accelerators using information learned during the computation process so that the assignments can be adapted and improved upon. Ex. C ('442) at 1:66–2:18. Collectively, the inventions allow for higher utilization of resources overall, which has enabled multiple advances, including improved scalability, resilience, and energy efficiency.

MICROSOFT'S USE OF PARTEC'S PATENTED TECHNOLOGY

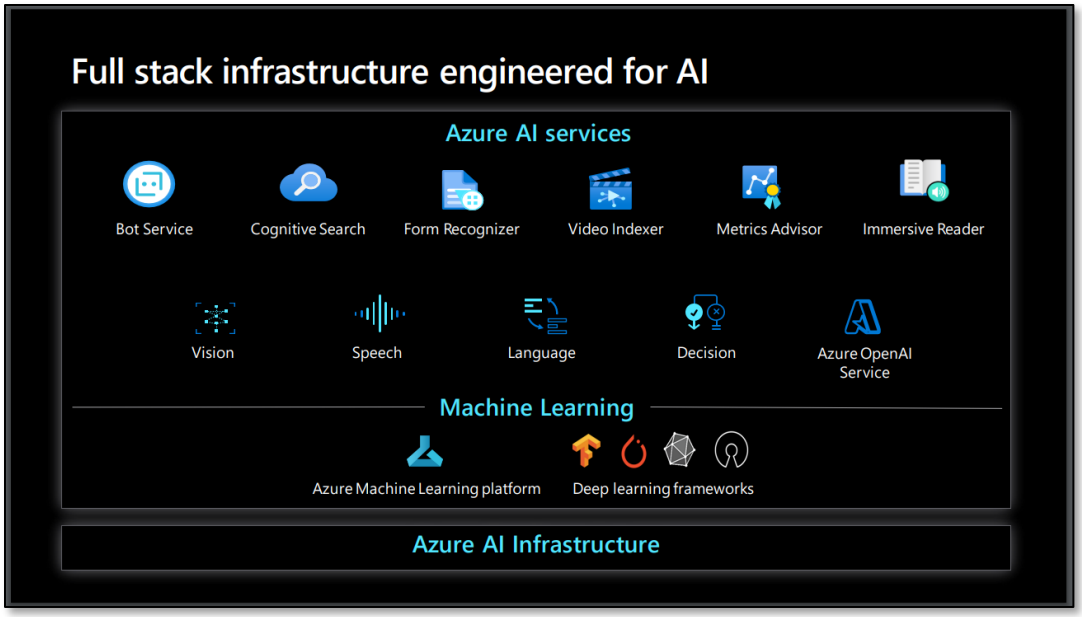
54. Microsoft makes, uses, offers to sell, sells, and/or imports into the United States products and/or systems that infringe the patents-in-suit, including what Microsoft refers to as the Microsoft Azure AI system or infrastructure.[5] *See* Microsoft Mechanics, *What runs GPT-4o?*

[5] This complaint uses the terms "Azure AI system" and "Azure AI infrastructure" interchangeably.

Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024) (Microsoft explaining that the Azure "AI system . . . refers to the specialized hardware and software stack behind [Microsoft's] AI supercomputer"); Microsoft, *Unleashing Innovation: The New Era of Compute Powering Azure AI Solutions* (May 21, 2024), https://azure.microsoft.com/en-us/blog/unleashing-innovation-the-new-era-of-compute-powering-azure-ai-solutions/ (Microsoft referring to "Azure's AI infrastructure"); Microsoft, *Azure AI Infrastructure*, https://azure.microsoft.com/en-us/solutions/high-performance-computing/ai-infrastructure/ (last accessed July 7, 2024) (Microsoft referring to Azure's "cloud AI supercomputing infrastructure").

55. Azure is Microsoft's highly successful cloud platform, comprising "more than 200 products and cloud services." Microsoft, *What is Azure?*, https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-azure/ (last accessed June 7, 2024). Ninety-five percent of Fortune 500 companies use Azure, and Microsoft invests over $1 billion annually in the platform. *Id.* Microsoft charges for Azure on a pay-as-you-go basis, meaning subscribers receive a bill each month that charges them for the specific resources and services they have used.

56. Microsoft Azure comprises a host of products and services. *See* Microsoft, *Azure Products*, https://azure.microsoft.com/en-us/products/ (last accessed June 7, 2024). Among others, is Microsoft Azure AI, which helps enterprises train, deploy, and scale AI, including large AI models. *See* Microsoft, *AI + Machine Learning*, https://azure.microsoft.com/en-us/products/category/ai (last accessed June 7, 2024) (describing AI products and services). Microsoft's AI products and service are shown below:



Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024)

57. As Microsoft notes, the computers necessary to run AI models are "super resource intensive and also pretty expensive." Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024). "Performance requirements for AI . . . are significantly different from other enterprise applications. AI requires infrastructure built specifically for compute-intensive, large-scale AI workloads." Microsoft, *Azure AI Infrastructure*, https://azure.microsoft.com/en-us/solutions/high-performance-computing/ai-infrastructure/ (last accessed July 7, 2024).

58. To provide the necessary computing, Microsoft built what it describes as a "specialized hardware and software stack that can support the efficient running [of] models of . . . massive scale": the Azure AI system. Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024).

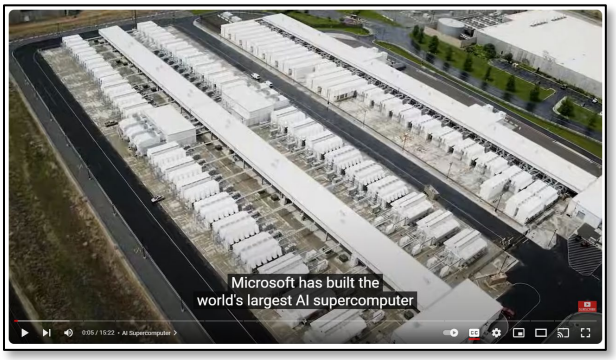
59. The Microsoft Azure AI system includes "a full technology stack with CPUs, GPUs, DPUs, systems, [and] networking." Microsoft, *Azure Blog / AI + Machine Learning*,

https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024). That stack is depicted below:

Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024)

60. Microsoft touts its "Azure AI infrastructure" as "deliver[ing] the scale, efficiency and performance organizations need to innovate." Microsoft Azure, *Accelerate Large-Scale AI Innovation and Empower Decision Making at Unprecedented Cloud Scale*, https://www.youtube.com/watch?v=SAhFOqJ88Ac, (last accessed June 7, 2024).

61. Microsoft also boasts that the system is "the world's largest AI supercomputer":



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

62. TOP500 has ranked the system as the number three supercomputer in the world and the largest cloud-based supercomputer. *See* Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024); TOP500, *June 2024*, https://top500.org/lists/top500/2024/06/ (last accessed June 7, 2024) (June 2024 list).

63. The original infrastructure Microsoft built, which has since been expanded and enlarged, cost Microsoft hundreds of millions of dollars. *See* Emma Roth, *Microsoft Spent Hundreds of Millions of Dollars on a ChatGPT Supercomputer* (Mar. 13, 2023), https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai.

64. Microsoft claims that its "Azure's AI infrastructure" "deliver[s] the most complete compute platform for AI workloads." Microsoft, *Unleashing Innovation: The New Era of Compute Powering Azure AI Solutions* (May 21, 2024), https://azure.microsoft.com/en-us/blog/unleashing-innovation-the-new-era-of-compute-powering-azure-ai-solutions/. According to Microsoft, its AI infrastructure "supports different scenarios for AI supercomputing, such as building large models from scratch, running inference on pre-trained models, using model as a service providers, and fine-tuning models for specific domains." *Id.* Through a "combination of advanced AI accelerators, datacenter designs, and optimized compute and networking topology that drive cost efficiency per workload," Microsoft claims that "the Azure platform ensures [customers] the best AI performance with optimized cost." *Id.*

65. Microsoft further boasts that its Azure AI infrastructure provides "[w]orld-class infrastructure performance for AI workloads" and "AI infrastructure optimized for Azure machine learning"—from a "trusted leader in AI."

World-class infrastructure performance for AI workloads

Deliver high-powered performance to your most compute-intensive AI workloads, including deep learning, with a purpose-built portfolio of AI infrastructure. Enjoy the latest in GPU performance and networking with Azure N-Series virtual machines (VMs) and seamlessly orchestrate your simulations on the cloud with Azure Batch and Azure CycleCloud.

AI Infrastructure optimized for Azure Machine Learning

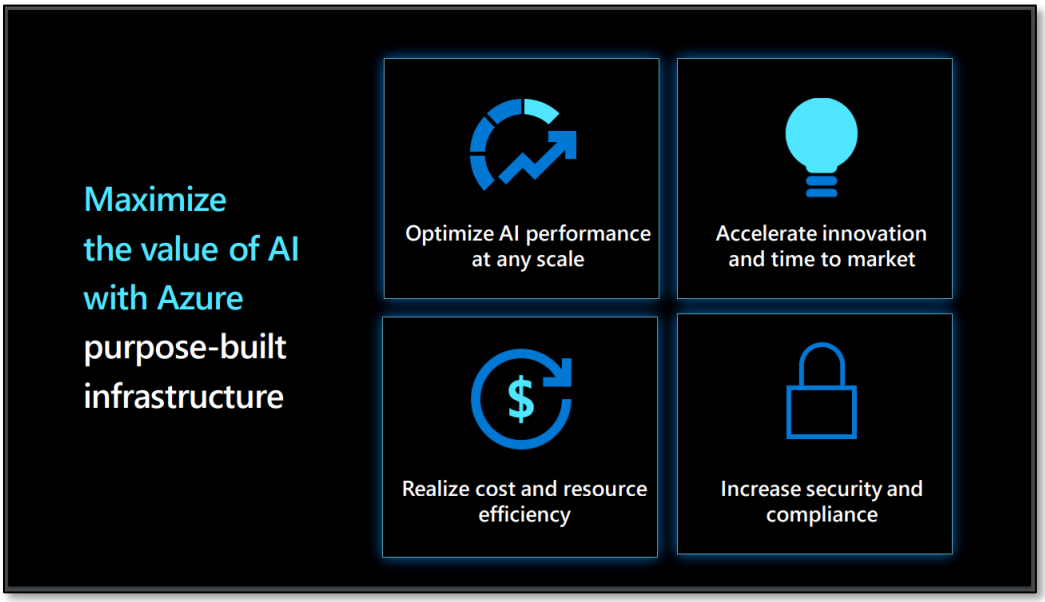
Build, deploy, and manage high-quality models faster using Azure Machine Learning. Accelerate your time to value using industry-leading machine learning operations, open-source interoperability, and integrated tools—all of which are fully supported and optimized by AI Infrastructure.

Trusted leader in AI

Microsoft is a proven leader in AI, delivering the AI-optimized infrastructure that helps build and train some of the industry's most advanced AI solutions, including OpenAI and NVIDIA, as well as Azure Machine Learning and Azure AI Services.

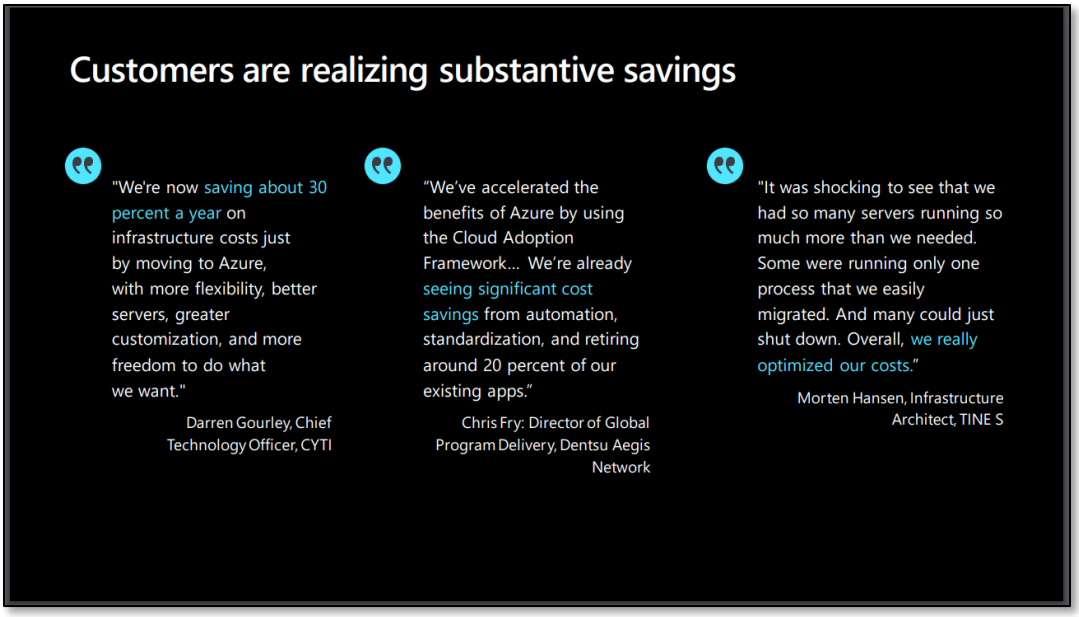
Source: Microsoft, *Azure AI Infrastructure*, https://azure.microsoft.com/en-us/solutions/high-performance-computing/ai-infrastructure/#features (last accessed June 7, 2024)

66. Microsoft similarly touts that customers can "maximize the value of AI with Azure purpose-built infrastructure"—and, in particular, "optimize AI performance at any scale," "accelerate innovation and time to market," "realize cost and resource efficiency," and "increase security and compliance."



Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024)

67. Microsoft further boasts that Azure AI has fueled "substantive" cost savings for customers:

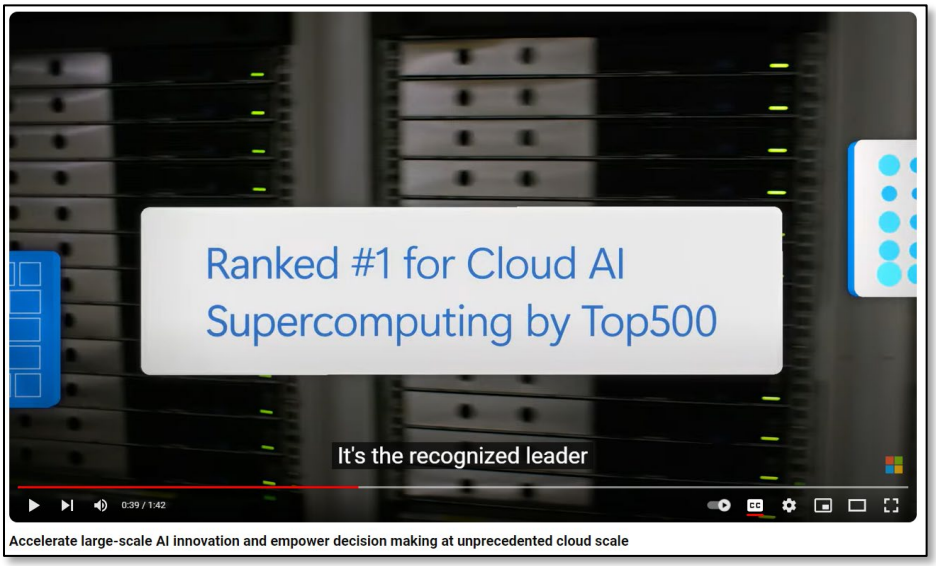


Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024)

68. In one video, Microsoft touts its Azure AI infrastructure as providing the "best performance, scalability, and built-in security needed by demanding AI workloads." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024). Microsoft further claims that the "flexibility to manage AI applications isn't just desirable for a growing number of organizations, it's a necessity. But investing in large fixed hardware and software upgrades can be costly in the long run." The solution, according to Microsoft, is its Azure AI system. *Id.*

69. In another video, Microsoft boasts that its AI infrastructure can "process a single job synchronously across thousands of interconnected GPUs" and "help[] [customers] drive down costs, by speeding up your decision making capabilities, model training times, and ability to go to market, all while delivering the same performance and scalability that's fueling newest AI innovations." Microsoft Azure, *Accelerate Large-Scale AI Innovation and Empower Decision Making at Unprecedented Cloud Scale*, https://www.youtube.com/watch?v=SAhFOqJ88Ac (last

accessed June 7, 2024). Microsoft further claims that it has a "proven track record engineering AI optimized infrastructure and delivering supercomputing performance for some of the most complex deep learning systems," including ChatGPT, and that it is "the recognized leader for performance and scale in the cloud." *Id.*



Source: Microsoft Azure, *Accelerate Large-Scale AI Innovation and Empower Decision Making at Unprecedented Cloud Scale*, https://www.youtube.com/watch?v=SAhFOqJ88Ac (last accessed June 7, 2024)

70. As part of highlighting the benefits and advantages of the Azure AI infrastructure, Microsoft has touted the dynamic assignment technology taught by the Asserted Patents. Microsoft, for example, has highlighted that its Azure AI infrastructure "support[s] a modular approach to deploy whatever GPU demand calls for" "to take advantage of the best cost performance." Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024). Microsoft has also boasted that its Azure AI infrastructure is "elastic, and able to quickly scale resources up or down to optimize operational costs." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*,

https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024). As Nidhi Chappell, Microsoft General Manager of Azure HPC and AI, put it: the components of the system are "pretty expensive resources. You cannot afford to have them sit idle, right? I want to make sure their utilization is pretty, pretty, pretty high." Timothy Prickett Morgan, *Inside the Infrastructure that Microsoft Builds to Run AI*, The Next Platform (Mar. 21, 2023), https://www.nextplatform.com/2023/03/21/inside-the-infrastructure-that-microsoft-builds-to-run-ai/. As Microsoft further claims, "an efficient infrastructure for running" AI models—with their large size—"is really critical." Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024).

71. Azure is big business for Microsoft. Microsoft recently released 2024 Q1 numbers and reported that revenue in Intelligent Cloud was $26.7 billion for the quarter, noting that server products and cloud services revenue increased 24%, driven by Azure. For its 2023 Q4 earnings release, Microsoft reported that "Microsoft's revenue from its Azure cloud platform and related services rose 30% in the quarter," with CEO Satya Nadella stating during the call that "[w]e've moved from talking about AI to applying AI at scale." Nadella also cited the growing adoption of Microsoft's AI tools across different industries and technologies. Despite the impressive historical growth, Microsoft is predicting an even more promising future for AI infrastructure:

Understanding the state of AI and demand for new infrastructure

2024 is shaping up to be an even more promising year for AI than its predecessor. With the rapid pace of technological advancements, AI infrastructure is becoming more diverse and widespread than ever before. From cloud to edge, CPUs to GPUs, and application-specific integrated circuits (ASICs), the AI hardware and software landscape is expanding at an impressive rate.

Source: Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/new-infrastructure-for-the-era-of-ai-emerging-technology-and-trends-in-2024/ (last accessed June 7, 2024)

72. Microsoft has spent over $100 billion in capital expenditures, much of which is attributed to server infrastructure and building out the accused features, including $40 billion in the last 12 months alone. Microsoft's CFO attributed the rise in capex spending to supporting cloud demand, including "the need to scale our AI infrastructure." Microsoft is now on track to spend over $50 billion in capital expenditures this year, a 50% increase from 2023. Microsoft is also rumored to be building a $100 million supercomputer. Reuters, *Microsoft, OpenAI Plan $100 Billion Datacenter Project, Media Report Says* (Mar. 29, 2024), https://www.reuters.com/technology/microsoft-openai-planning-100-billion-data-center-project-information-reports-2024-03-29/.

73. Across industries, 95% of businesses in a recent survey stated that they plan to increase their AI usage over the next two years. They further reported AI as being critical to success. Microsoft Azure, *The State of AI Infrastructure, 2024 Edition*, https://clouddamcdnprodep.azureedge.net/gdc/gdcBkY6mR/original.

74. Given the growth of AI and other complex computing tasks, the second worldwide quantum computer study by the International Data Corporation (IDC) predicts that potential customers' spending on quantum computers will increase from $1.1 billion in 2022 to $7.6 billion in 2027, with a compound annual growth rate (CAGR) of 48.1% (2023–2027). The study further states, "Quantum computing will revolutionize companies' ability to solve some of the most complex challenges."

COUNT ONE
INFRINGEMENT OF U.S. PATENT NO. 10,142,156

75. Plaintiffs repeat and incorporate by reference each preceding paragraph as if fully set forth herein and further state:

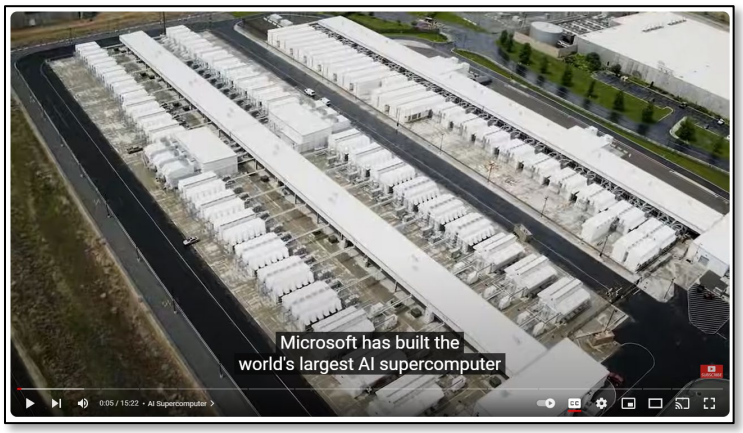
76. Microsoft has infringed and continues to directly infringe the '156 Patent in

violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, or offering for sale in the United States, and/or importing into the United States, without authorization, systems and methods that practice claims of the ’156 Patent, including the Microsoft Azure AI system.

77. For example, Claim 1 is illustrative of the claims of the ’156 Patent. It recites “[a] computer cluster-booster system for processing a computation task, comprising:

> a plurality of hardware computation nodes, each of which interfaces with a communication infrastructure, at least two of the hardware computation nodes being arranged to jointly compute at least a first part of said computation task;
>
> a plurality of hardware boosters, each hardware booster having a compute capacity, at least one hardware booster of the plurality of hardware boosters being arranged to compute at least a second, specific part of said computation task after having been assigned to at least one hardware computation node and under control of that at least one hardware computation node, the at least one hardware booster interfacing with the communication infrastructure; and
>
> a resource manager being arranged to assign the at least one hardware booster to the at least one hardware computation node, including:
>
> > at a start of processing of said computation task, establishing an initial assignment by using a predetermined assignment metric specified as a function of at least one of a group of assignment parameters, and
> >
> > during said processing of said computation task: (i) updating the predetermined assignment metric, and (ii) establishing a dynamic assignment by using the predetermined assignment metric that was updated, and
>
> wherein the plurality of hardware computation nodes and the plurality of hardware boosters are configured such that during processing of said computation task, assignments of hardware computation nodes and hardware boosters can be provided such that at least (i) at least one of the plurality of hardware computation nodes is arranged to communicate with at least one of the plurality of hardware boosters, (ii) at least one of the plurality of hardware boosters is assigned to and shared by more than one of the plurality of hardware computation nodes such that the compute capacity of the at least one of the plurality of hardware boosters is shared between the more than one of the plurality of hardware computation nodes, and (iii) each of the hardware boosters is assignable to each of the hardware computation nodes.”

78. Microsoft's Azure AI system meets every element of this claim.[6] Microsoft touts the Azure AI system as "the world's largest AI supercomputer:"



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

79. The Azure AI system includes "GPUs, networking, and the full stack of AI software." Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024); *see also* Microsoft, *How Microsoft's Bet on Azure Unlocked an AI Revolution* (Mar. 13, 2023), https://news.microsoft.com/source/features/ai/how-microsofts-bet-on-azure-unlocked-an-ai-revolution/ (stating that the Azure AI infrastructure includes "the combination of GPUs, networking hardware and virtualization software required to deliver the compute needed to power the next wave of AI innovation"). Further, it includes "a full technology stack with CPUs, GPUs, DPUs, systems, [and] networking." Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024). That stack is depicted below:

[6] This description of infringement is illustrative and not intended to be an exhaustive or limiting explanation of every manner in which the Microsoft Azure AI system infringes.

Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024)

80. Microsoft further touts its Azure AI system as the solution to AI's need for "infrastructure that is optimized for compute heavy workloads." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024). Hence, the Microsoft Azure AI system is a computer-cluster booster system for processing a computation task.

81. The Microsoft Azure AI system includes a plurality of hardware computation nodes—for example, CPU cores. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]he supercomputer that we built for OpenAI back in 2020 comprises more than 285,000 AMD InfiniBand connected CPU cores"). A CPU or central processing unit "is a hardware component that's the core computational unit in a server. . . . It fetches instructions from memory, performs the required tasks, and sends output back to memory." Amazon, *What is a CPU (Central Processing Unit)?*, https://aws.amazon.com/what-

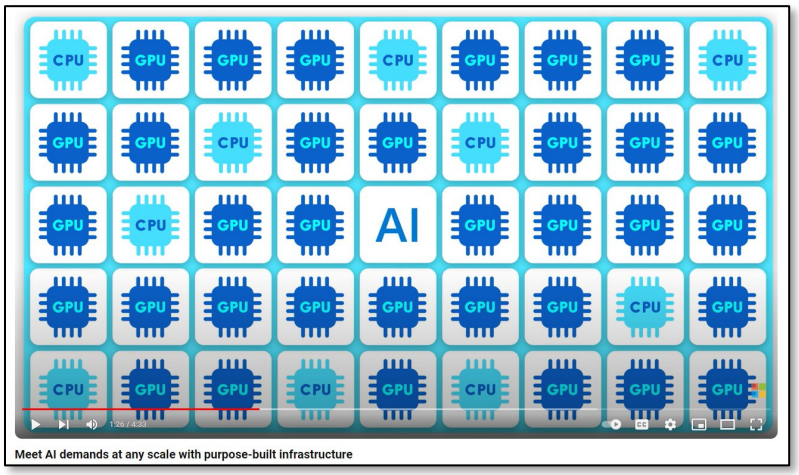
is/cpu/ (last accessed June 7, 2024). "A core is a processing unit of the CPU." Vinicius Fulber-Garcia, *Differences Between Core and CPU*, Baeldung (Mar. 18, 2024), https://www.baeldung.com/cs/core-vs-cpu. In 2020, Azure had 285,000 CPU cores. By November 2023, that number had ballooned to 1.1 million.

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

82. The computation nodes interface with a communication infrastructure—for example, the CPU cores interface using InfiniBand. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]he supercomputer that we built for OpenAI back in 2020 comprises more than 285,000 AMD InfiniBand connected CPU cores"). "InfiniBand is a channel-based fabric that facilitates high-speed communications between interconnected nodes." Robert Sheldon, *What is Infiniband?*, TechTarget, https://www.techtarget.com/searchstorage/definition/InfiniBand (last accessed June 7, 2024).

83. The computation nodes are arranged to jointly compute a first part of a computation

task. In Azure, multiple CPU cores work together to perform computation tasks, as Microsoft depicts in videos describing its AI infrastructure.

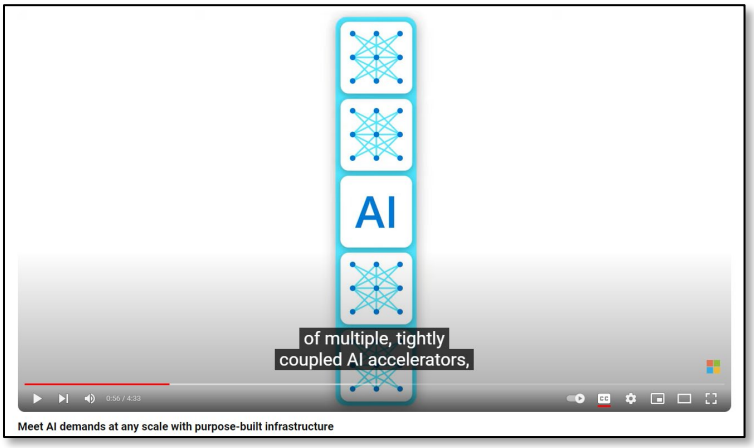


Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

84. The Microsoft Azure AI system includes a plurality of hardware boosters, each with a compute capacity. The Azure AI system includes, for example, GPUs. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]here are 10,000 NVIDIA V100 Tensor Core GPUs that are also InfiniBand connected."). A GPU or graphics processing unit "is an electronic circuit that can perform mathematical calculations at high speed. Computing tasks like graphics rendering, machine learning (ML), and video editing require the application of similar mathematical operations on a large dataset. A GPU's design allows it to perform the same operation on multiple data values in parallel. This increases its processing efficiency for many compute-intensive tasks." Amazon, *What is a GPU?*, https://aws.amazon.com/what-is/gpu/ (last accessed June 7, 2024). In 2020, the Azure AI system had 10,000 GPUs. By November 2023, that number had increased to 14,400 GPUs.

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

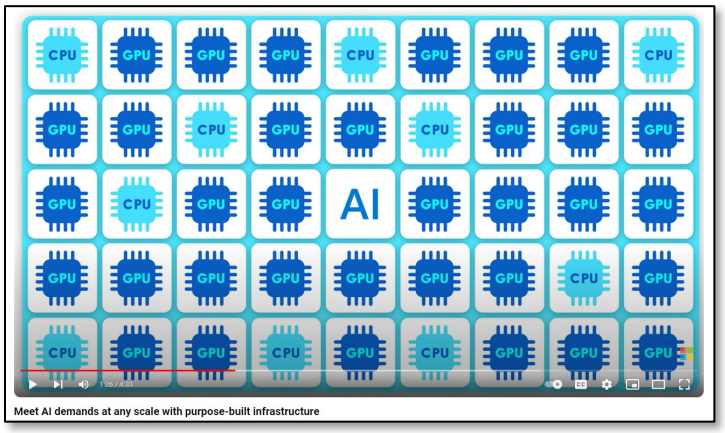
Microsoft touts its AI infrastructure as comprising "multiple, tightly coupled AI accelerators"—referring to GPUs.



Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

85. The hardware boosters of Azure—*e.g.*, the GPUs—can be arranged to compute at least a second, specific part of the computation task after having been assigned to at least one hardware computation node and under control of that at least one hardware computation node. Microsoft in videos describing its AI infrastructure depicts computation nodes (*e.g.*, CPU cores)

and boosters or accelerators (*e.g.*, GPUs) working together in harmony to perform complex computing tasks.



Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

86. Microsoft further touts that the GPUs of the Azure AI system provide users the ability to efficiently accelerate work on powerful systems. *See* Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024) ("H100 and Quantum-2 are part of NVIDIA's high-performance computing (HPC) platform—a full technology stack with CPUs, GPUs, DPUs, systems, networking, and a broad range of AI and HPC software—that provides researchers the ability to efficiently accelerate their work on powerful systems, on-premises or in the cloud."). Code, such as Python code running on CPUs, handle the control.

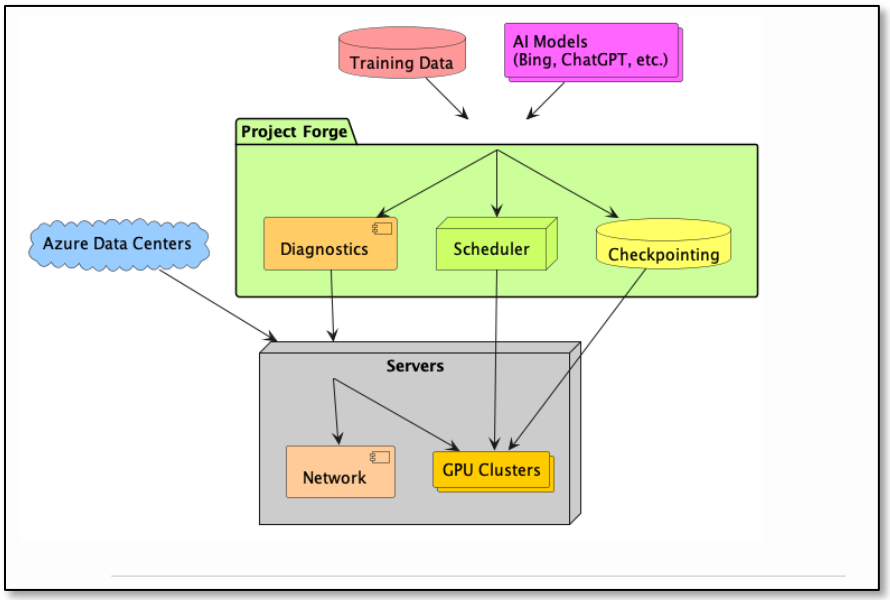
87. The hardware boosters interface with the communication infrastructure. Like the CPU cores, the GPUs are connected using InfiniBand. Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]here are 10,000 NVIDIA V100

Tensor Core GPUs that are also InfiniBand connected.”).

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

88. The Microsoft Azure AI system includes a resource manager arranged to assign hardware boosters (*e.g.*, GPUs) to computation nodes (*e.g.*, CPU cores) that, at the start of processing a computation task, establishes an initial assignment by using a predetermined assignment metric (*e.g.*, utilization) specified as a function of at least one of a group of assignment parameters. “Project Forge”—previously called “Singularity”—is a global scheduler that pools GPU capacity from different regions to help run Microsoft’s global scale AI workloads and maintain high levels of utilization. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft’s AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 (“And the way we address that in Azure is with a containerization and global scheduler service that we’ve been working on called Project Forge, which is designed specifically to help run Microsoft’s global scale AI workloads and maintain really high levels of utilization. Project Forge introduces transparent checkpointing,

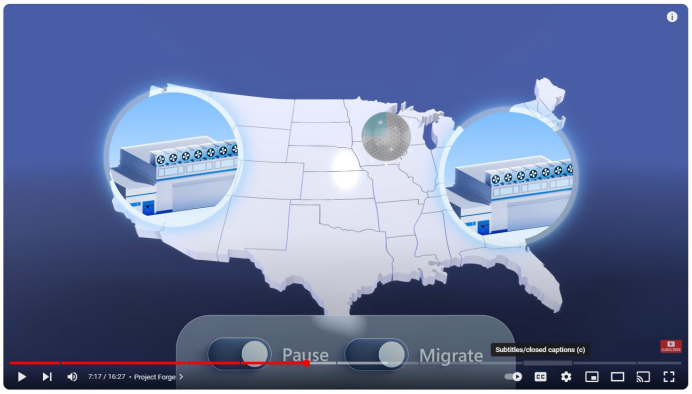
where it periodically saves the state of a model incrementally, without the model's code needing to do anything. That way, if anything fails, it can quickly resume for the most recent checkpoint. We combine this with our integrated global scheduler that pools GPU capacity from regions around the world. So, if you need to pause a job to prioritize another one, that allows us to migrate that pause job to another region if necessary and available, with minimal impact on its progress."); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("At the heart of Singularity is a novel, workload-aware scheduler that can transparently preempt and elastically scale deep learning workloads to drive high utilization without impacting their correctness or performance across a global fleet of AI accelerators (e.g., GPUs, FPGAs).").



Source: Perez Rivas Consulting, *Inside Microsoft's AI Supercomputer Powering ChatGPT and Large Language Models* (July 26, 2023), https://www.dr-perez-rivas-consulting.com/2023/07/26/inside-microsoft-s-ai-supercomputer-powering-chatgpt-and-large-language-models/ (depicting Project Forge's relationship to the system)

89. Further, the resource manager—during the processing of a computation task—can update the predetermined assignment metric and establish a dynamic assignment by using the

predetermined assignment metric that was updated. Project Forge (Singularity), for example, adapts assignments to changes in load. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("So, we're going to train the model which just takes a short period of time and now the GPU is going to go idle. And at that point Project Forge sees that it's gone idle and automatically takes a checkpoint, because it knows that I'm not busy using the GPUs. Normal circumstances, I would've been hogging the GPU at that point and preventing it from being used for somebody else's workload. But now that GPU can be repurposed . . . ."); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("All jobs in Singularity are preemptible, migratable, and dynamically resizable (elastic) by default: a live job can be dynamically and transparently (a) pre-empted and migrated to a different set of nodes, cluster, data centre or a region and resumed exactly from the point where the execution was pre-empted, and (b) resized (i.e., elastically scaled up/down) on a varying set of accelerators of a given type."); *id.* ("No idling of resources: Singularity treats the entire fleet of accelerators as a single logical, shared cluster, and avoids any resource fragmentation or static reservation of capacity."); *id.* ("For example, Singularity adapts to increasing load on an inference job, freeing up capacity by elastically scaling down or pre-empting training jobs."); *id.* ("Resizing/Elasticity: Singularity enables all jobs to be dynamically and elastically scaled up or down in a transparent manner to use a variable number of AI accelerators."); *id.* ("The binding between the job workers in Singularity and the accelerator devices is dynamic and constantly changing during the lifetime of the job.").



Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024) (depicting the migration of a job)

As Microsoft has touted, its Azure AI infrastructure "support[s] a modular approach to deploy whatever GPU demand calls for" "to take advantage of the best cost performance." Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024). As Microsoft further touts, its Azure AI infrastructure is "elastic, and able to quickly scale resources up or down to optimize operational costs." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024).

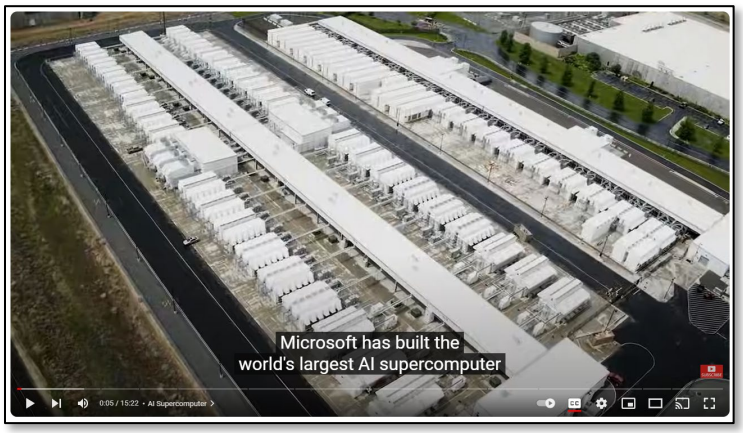
90. Finally, in the Microsoft Azure AI system, the hardware computation nodes and the hardware boosters are configured such that during processing of a computation task, assignments of computation nodes and boosters can be provided such that the hardware computation nodes are arranged to communicate with the hardware boosters and at least one of the hardware boosters is assigned to and shared by more than one of the hardware computation nodes such that the compute capacity of the hardware booster(s) is shared by more than one of the plurality of hardware computation nodes. For example, the GPUs of the Azure AI infrastructure "are designed specifically for multi-tenant cloud compute"—meaning GPUs are shareable. Microsoft, *What*

Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("And the new H100 VM series in Azure, powered by NVIDIA H100 Tensor Core GPUs, lets you choose one or more GPUs, cluster up to eight GPUs per VM using NVLink, and scale out to thousands if your demands grow, using NVIDIA Quantum-2 InfiniBand networking. That's up to 30x higher performance in inferencing and 4x higher for training compared to NVIDIA's previous A100 generation of GPUs. These are designed specifically for multi-tenant cloud compute, giving us the ability to isolate customers sharing the same server from one another, and with that we can achieve the elastic scale that we need."). Hence, a hardware booster can be assigned to and shared by more than one computation node such that the compute capacity of hardware boosters is shared between the hardware computation nodes.

91. Moreover, each of the hardware boosters is assignable to each of the computation nodes. For example, idle GPUs can be repurposed and assigned to a different computation node. Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("So, we're going to train the model which just takes a short period of time and now the GPU is going to go idle. And at that point Project Forge sees that it's gone idle and automatically takes a checkpoint, because it knows that I'm not busy using the GPUs. Normal circumstances, I would've been hogging the GPU at that point and preventing it from being used for somebody else's workload. But now that GPU can be repurposed . . . .").

92. Microsoft directly infringes Claim 1 of the '156 Patent—and the '156 Patent more

generally—through multiple infringing acts. For example,[7] Microsoft has combined all elements of the claimed system and thus "made" the system, as Microsoft admits:



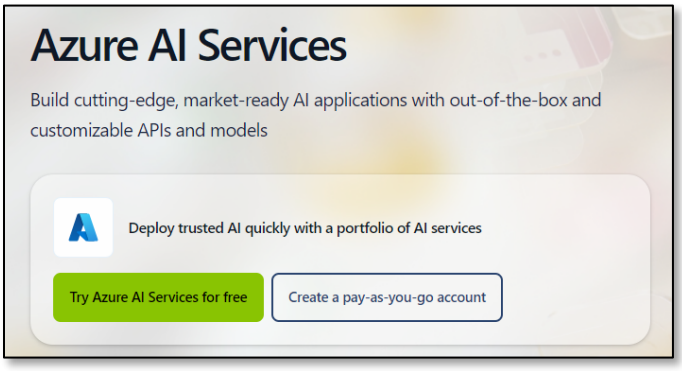
Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

93. Further, Microsoft uses the Azure AI infrastructure to run its own computation tasks and thus uses the system. *See* Timothy Prickett Morgan, *Inside the Infrastructure that Microsoft Builds to Run AI*, The Next Platform (Mar. 21, 2023), https://www.nextplatform.com/2023/03/21/inside-the-infrastructure-that-microsoft-builds-to-run-ai/ (Nidhi Chappell, Microsoft General Manager of Azure HPC and AI: "Whether it is internal teams running Bing, ChatGPT, or whatever – everything is running on Azure public infrastructure. . . . We use the same infrastructure, we make it available internally and externally."); Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*, Yahoo! Finance (Mar. 13, 2023), https://finance.yahoo.com/news/microsoft-strung-together-tens-thousands-130035397.html ("Now Microsoft uses that same set of resources it built for OpenAI to train and run its own large artificial intelligence models, including the new Bing search bot

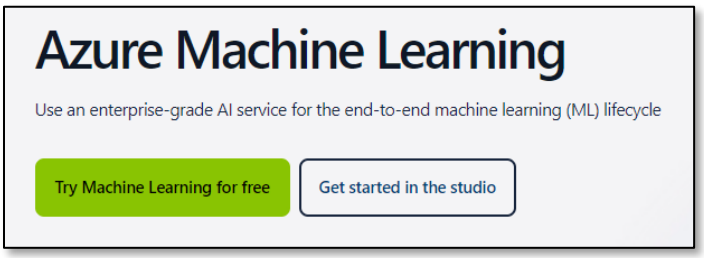
[7] The following infringing acts are illustrative and not intended to be an exhaustive or limiting list of each of Microsoft's infringing acts.

introduced last month."). Microsoft also uses the Azure AI infrastructure, the system, to run the computation tasks of its customers. *Id.*

94. And Microsoft "sells the system to . . . customers." Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*, Yahoo! Finance (Mar. 13, 2023), https://finance.yahoo.com/news/microsoft-strung-together-tens-thousands-130035397.html. It relatedly, and necessarily, offers to sell the system:



Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 8, 2024, in Longview, Texas)



Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed on June 8, 2024, in Longview, Texas)

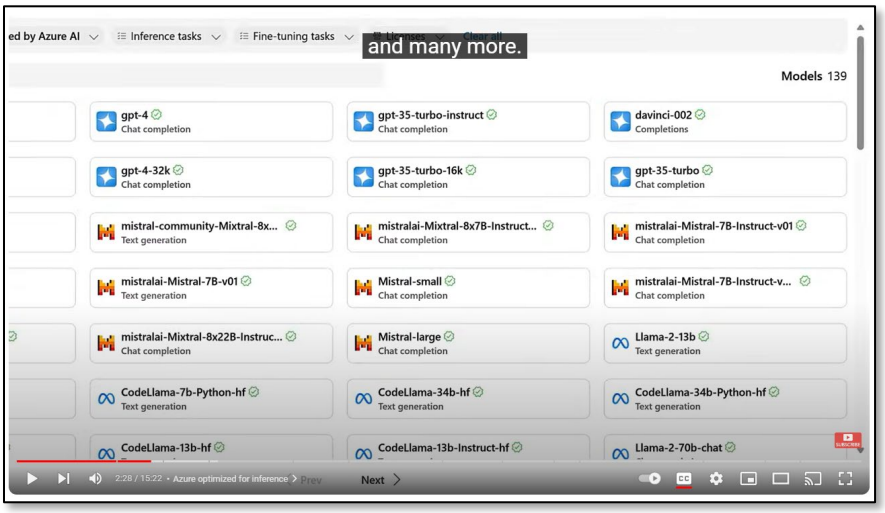
95. In addition to directly infringing the '156 Patent by making, using, selling, offering to sell, and/or importing infringing products into the United States, Microsoft also indirectly infringes one or more claims of the '156 Patent. Where acts constituting direct infringement of the '156 Patent may not be performed by Microsoft, such acts constituting direct infringement of the '156 Patent are performed by Microsoft's customers or end-users who act at the direction and/or

control of Microsoft, with Microsoft's knowledge.

96. Microsoft indirectly infringes one or more claims of the '156 Patent by active inducement in violation of 35 U.S.C. § 271(b), by at least manufacturing, supplying, distributing, selling, and/or offering for sale the Microsoft Azure AI system to its clients with full knowledge and intent that use of the same would constitute direct infringement of the '156 Patent.

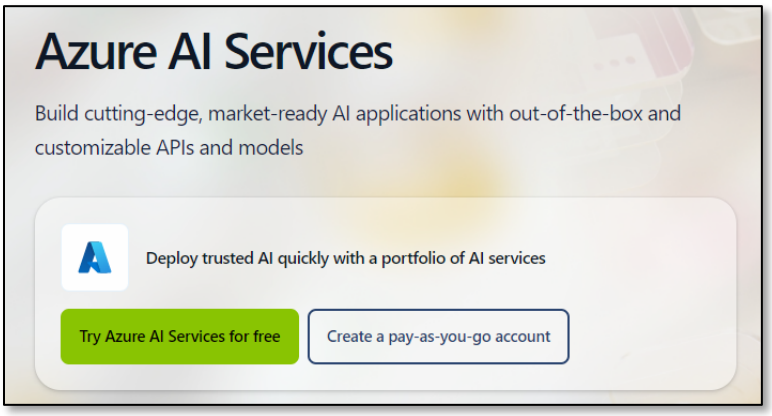
97. Microsoft has been aware of the '156 Patent at least as of the filing of this Complaint.

98. Moreover, Microsoft intends to cause, and has taken affirmative steps to induce, infringement by customers and end-users by at least, *inter alia*, encouraging, promoting, instructing, and/or directing the infringing use of the Microsoft Azure AI system. For example, Microsoft's "model as a service option in Azure" allows users to use Microsoft's AI "infrastructure to access and run the most sophisticated AI models, such as GPT-3.5 Turbo, GPT-4, Meta's Llama, Mistral, and many more," which Microsoft has touted as important for organizations that do not have the resources to build out and train their own AI models.



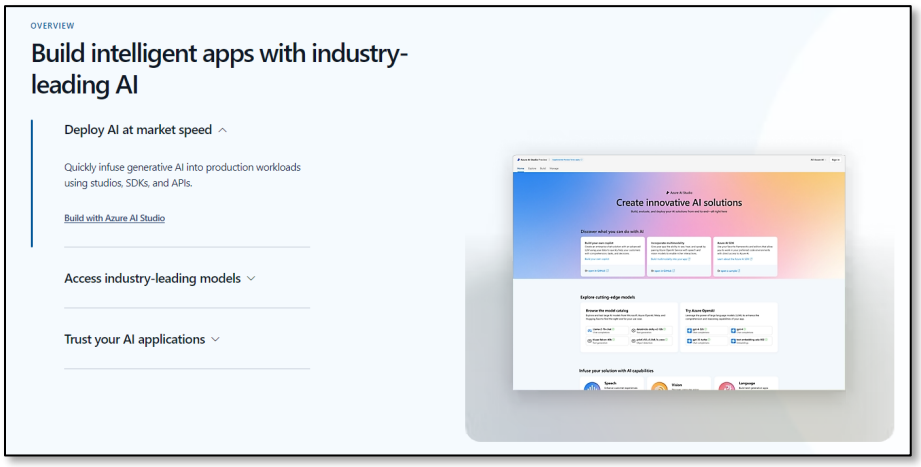
Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

99. Microsoft encourages users to use and take advantage of this option and Microsoft's Azure AI infrastructure more generally. Microsoft promotes its Azure AI services, and encourages users and customers to use these services, to "[b]uild cutting-edge, market-ready AI applications with out-of-the-box and customizable APIs and models."



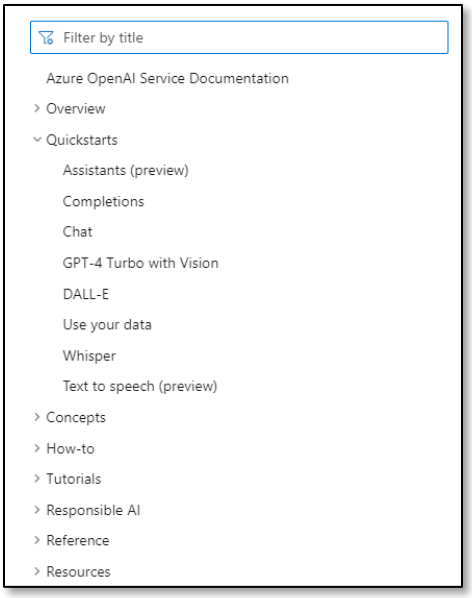
Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

100. Microsoft further promotes its Azure AI services as allowing customers to "build intelligent apps with industry-leading AI," including "deploy[ing] AI at market speed," "access[ing] industry-leading models," and "trust[ing] your AI applications."

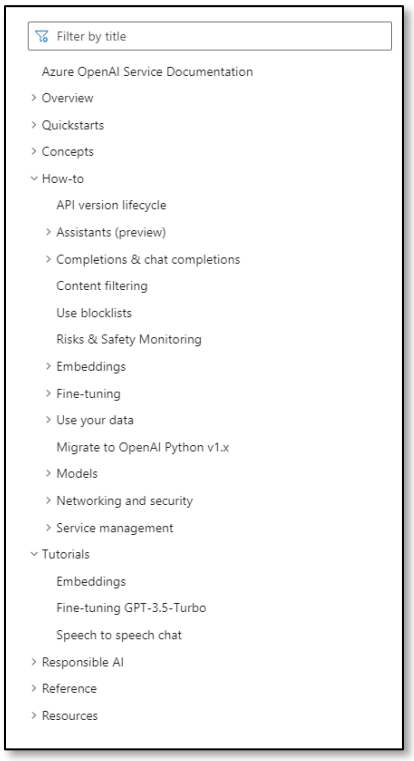


Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

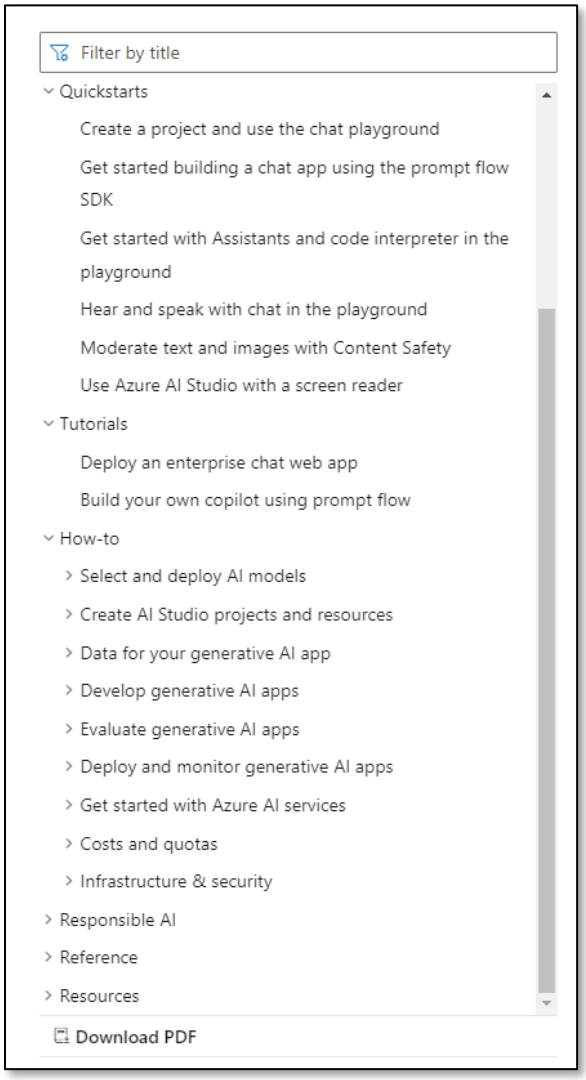
101. Microsoft also offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to use Azure AI:



Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)

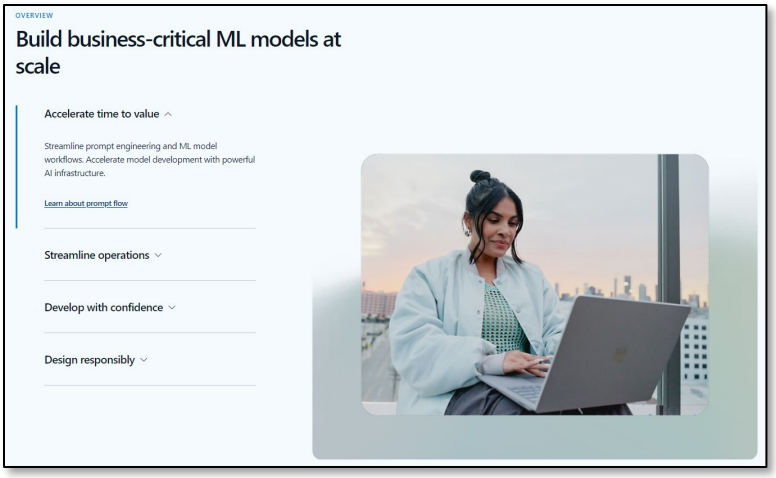


Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)



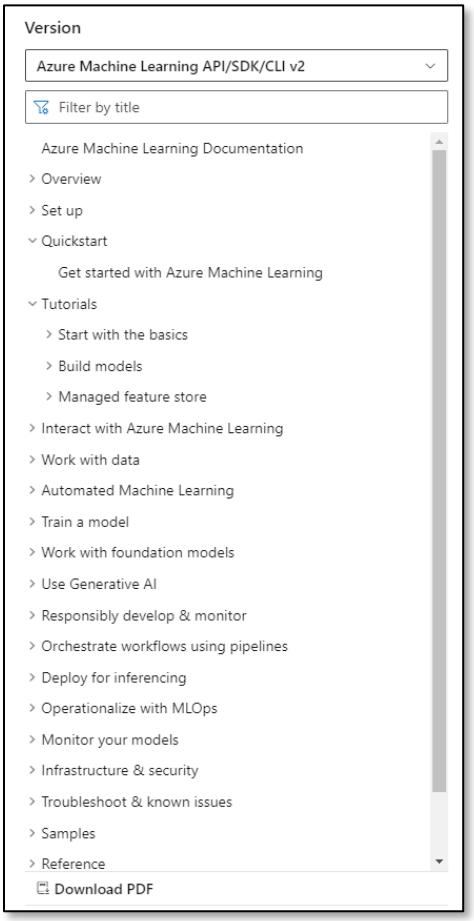
Source: Microsoft, *Azure AI Frequently Asked Questions*, https://learn.microsoft.com/en-us/azure/ai-studio/faq (last accessed June 7, 2024)

102. Similar to its Azure AI services, Microsoft actively promotes and induces use of Azure machine learning, which relies on Microsoft's Azure AI infrastructure. Microsoft promotes Azure machine learning, and encourages users and customers to use the same, to "build business-critical [machine learning] models at scale," allowing customers to "accelerate time to value," "streamline operations," "develop with confidence," and "design responsibly."



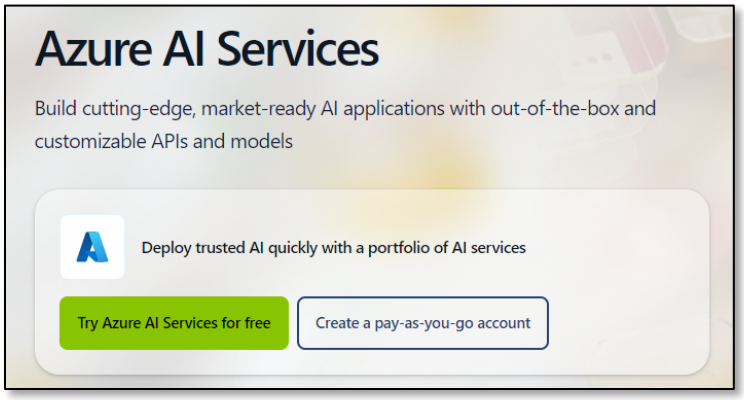
Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

103. As with Azure AI services, Microsoft offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to use its machine learning services.

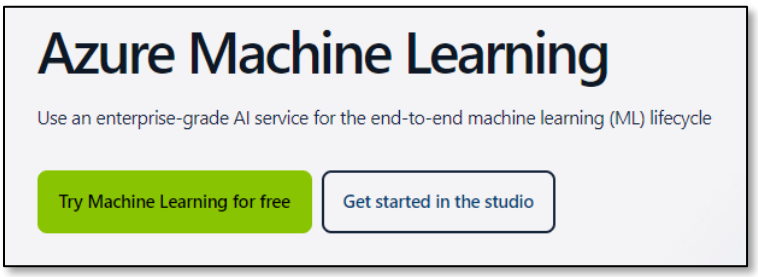


Source: Microsoft, *Azure Machine Learning Documentation*, https://learn.microsoft.com/en-us/azure/machine-learning/?view=azureml-api-2 (last visited June 7, 2024)

104. To encourage use of the foregoing products and services, Microsoft even offers free 30-day trials. After that, users pay as they go.



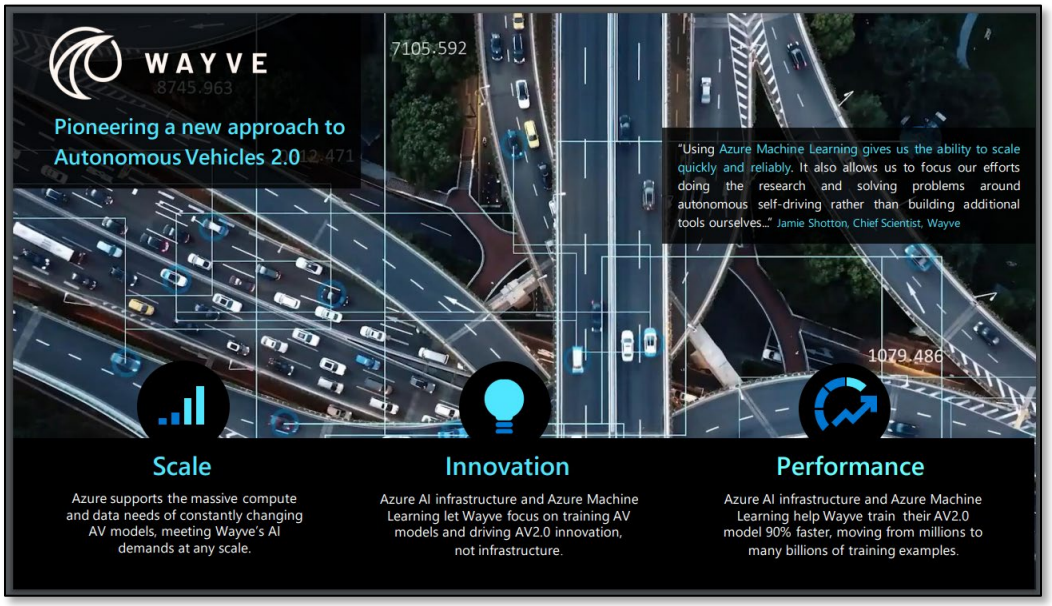
Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)



Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

105. Customers and end-users have heeded Microsoft's encouragement. Microsoft touts the various ways its customers have used Azure AI. *See* Microsoft, *AI Customer Stories*, https://www.microsoft.com/en-us/ai/ai-customer-stories (last accessed June 7, 2024) (collecting stories). Volvo, for example, uses Azure AI to "streamlin[e] invoice processing." *Id.*; *see also* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024) (explaining how

Wayve—a self-driving car company—uses Azure AI).



Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024) (same)

106. As detailed above, Microsoft's Azure AI infrastructure infringes at least Claim 1 of the '156 Patent. Accordingly, by encouraging, promoting, instructing, and/or directing users to use Microsoft Azure AI, Microsoft is actively inducing infringement of the '156 Patent in violation of 35 U.S.C. § 271(b).

107. Microsoft also indirectly infringes one or more claims of the '156 Patent by contributory infringement in violation of 35 U.S.C. § 271(c). Microsoft is aware that components of its Microsoft Azure AI system are a material and substantial part of the invention claimed by the '156 patent, and that they are designed for a use that is both patented and infringing, and that has no substantial non-infringing uses.

108. Microsoft's acts of infringement have caused damage to Plaintiffs, and Plaintiffs are entitled to recover from Microsoft (or any successor entity to Microsoft) the damages sustained by Plaintiffs as a result of Microsoft's wrongful acts in an amount subject to proof at trial.

109. To the extent applicable, Plaintiff has complied with 35 U.S.C. § 287(a) with respect to the ’156 Patent.

COUNT TWO
INFRINGEMENT OF U.S. PATENT NO. 11,934,883

110. Plaintiffs repeat and incorporate by reference each preceding paragraph as if fully set forth herein and further state:

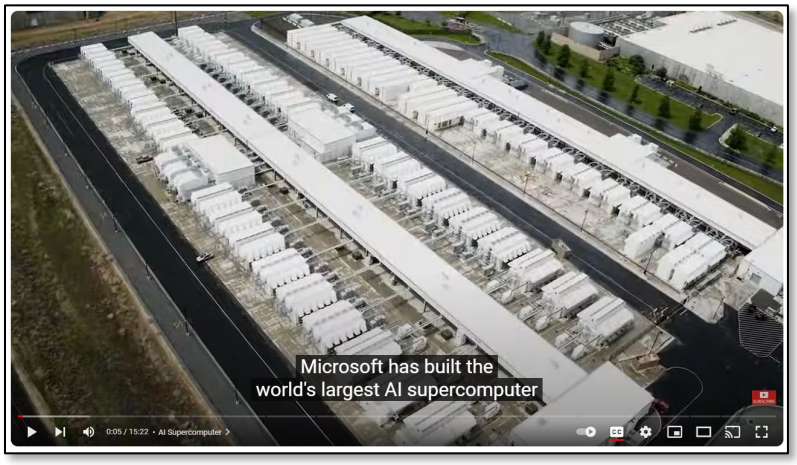
111. Microsoft has infringed and continues to directly infringe the ’883 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, or offering for sale in the United States, and/or importing into the United States, without authorization, systems and methods that practice claims of the ’883 Patent, including the Microsoft Azure AI system.

112. For example, Claim 1 is illustrative of the claims of the ’883 Patent. It recites “[a] computer system for processing a computation task, comprising:

> a plurality of hardware computation nodes, a plurality of hardware boosters, and a resource manager, the plurality of hardware computation nodes and the plurality of hardware boosters each interfacing a communication infrastructure;
>
> the resource manager being arranged to:
>
> > assign a selected hardware booster of the plurality of hardware boosters to a first hardware computation node of the plurality of hardware computation nodes for computation of a part of the computation task,
> >
> > provide assignment information to the first hardware computation node after the assignment of the selected hardware booster so as to enable the first hardware computation node to outsource the part of the computation task to the assigned selected hardware booster under control of the first hardware computation node,
> >
> > initialize a static assignment and further to establish a dynamic assignment during the processing of the computation task,
> >
> > accomplish the assignments as a function of a predetermined assignment metric:
> >
> > > provide the static assignment at the start of the processing of the computation task by using the predetermined assignment metric, and

update the predetermined assignment metric during the processing of the computation task."

113. Microsoft's Azure AI system meets every element of this claim.[8] Microsoft touts the Azure AI system as "the world's largest AI supercomputer:"



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

114. The Azure AI system includes "GPUs, networking, and the full stack of AI software." Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024); *see also* Microsoft, *How Microsoft's Bet on Azure Unlocked an AI Revolution* (Mar. 13, 2023), https://news.microsoft.com/source/features/ai/how-microsofts-bet-on-azure-unlocked-an-ai-revolution/ (stating that the Azure AI infrastructure includes "the combination of GPUs, networking hardware and virtualization software required to deliver the compute needed to power the next wave of AI innovation"). Further, it includes "a full technology stack with CPUs, GPUs, DPUs, systems, [and] networking." Microsoft, *Azure Blog / AI + Machine Learning*,

[8] This description of infringement is illustrative and not intended to be an exhaustive or limiting explanation of every manner in which Microsoft Azure AI system infringes.

https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024). That stack is depicted below:

Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024)

115. Microsoft further touts its Azure AI infrastructure as the solution to AI's need for "infrastructure that is optimized for compute heavy workloads." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024). Hence, the Microsoft Azure AI system is a computer cluster system for processing a computation task.

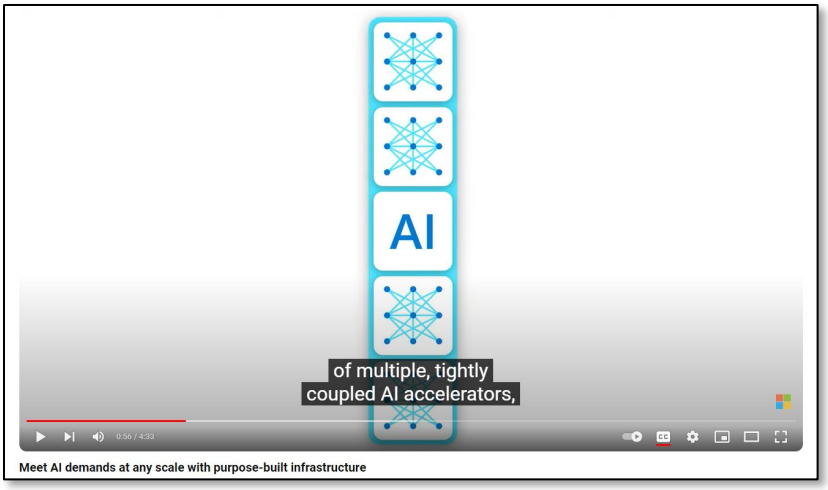
116. The Microsoft Azure AI system includes a plurality of hardware computation nodes—for example, CPU cores. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]he supercomputer that we built for OpenAI back in 2020 comprises more than 285,000 AMD InfiniBand connected CPU cores"). In 2020, Azure had 285,000 CPU cores. By November 2023, it had 1.1 million.

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

117. The Microsoft Azure AI system includes a plurality of hardware boosters. The Azure AI system includes, for example, GPUs. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]here are 10,000 NVIDIA V100 Tensor Core GPUs that are also InfiniBand connected."). In 2020, the Azure AI system had 10,000 GPUs. By November 2023, that number had increased to 14,400 GPUs.

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

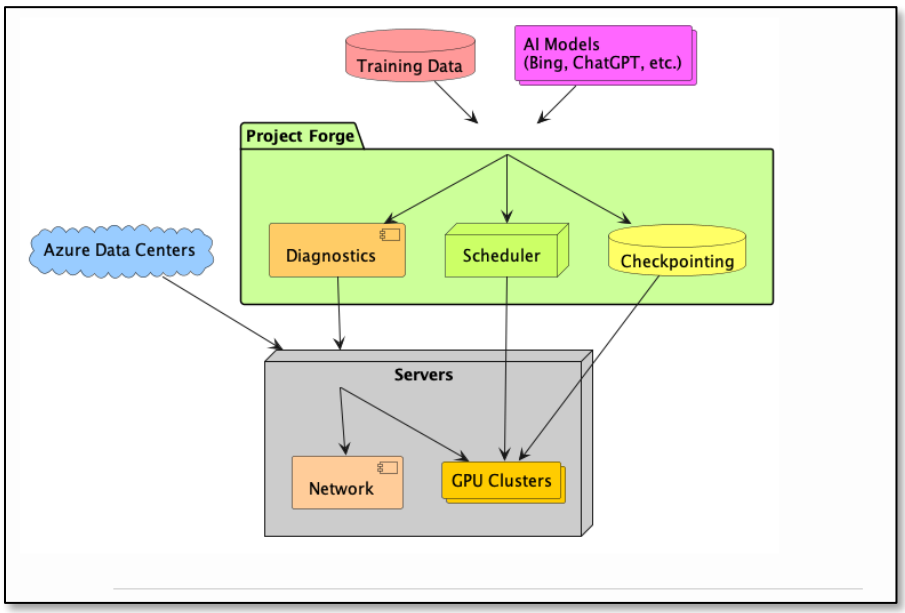
Microsoft touts its AI infrastructure as comprising "multiple, tightly coupled AI accelerators"—referring to GPUs.



Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

118. The Microsoft Azure AI system includes a resource manager. "Project Forge"—previously called "Singularity"—is a global scheduler that pools GPU capacity from different regions to help run Microsoft's global scale AI workloads and maintain high levels of utilization. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("And the way we address that in Azure is with a containerization and global scheduler service that we've been working on called Project Forge, which is designed specifically to help run Microsoft's global scale AI workloads and maintain really high levels of utilization. Project Forge introduces transparent checkpointing, where it periodically saves the state of a model incrementally, without the model's code needing to do anything. That way, if anything fails, it can quickly resume for the most recent checkpoint. We combine this with our integrated global scheduler that pools GPU

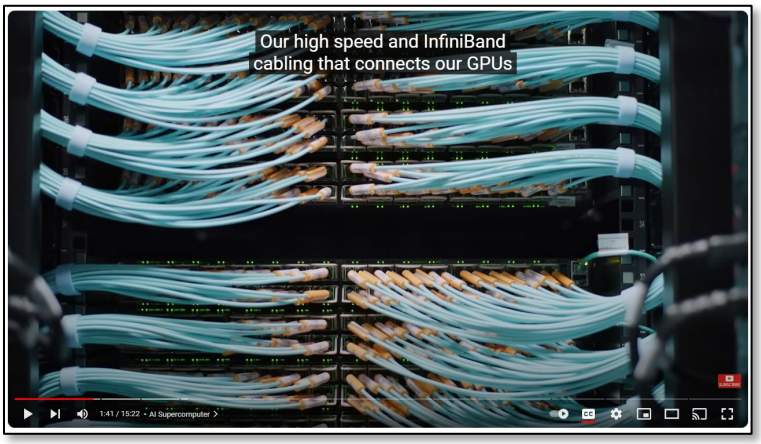
capacity from regions around the world. So, if you need to pause a job to prioritize another one, that allows us to migrate that pause job to another region if necessary and available, with minimal impact on its progress."); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("At the heart of Singularity is a novel, workload-aware scheduler that can transparently preempt and elastically scale deep learning workloads to drive high utilization without impacting their correctness or performance across a global fleet of AI accelerators (e.g., GPUs, FPGAs).").



Source: Perez Rivas Consulting, *Inside Microsoft's AI Supercomputer Powering ChatGPT and Large Language Models* (July 26, 2023), https://www.dr-perez-rivas-consulting.com/2023/07/26/inside-microsoft-s-ai-supercomputer-powering-chatgpt-and-large-language-models/ (depicting Project Forge's relationship to the system)

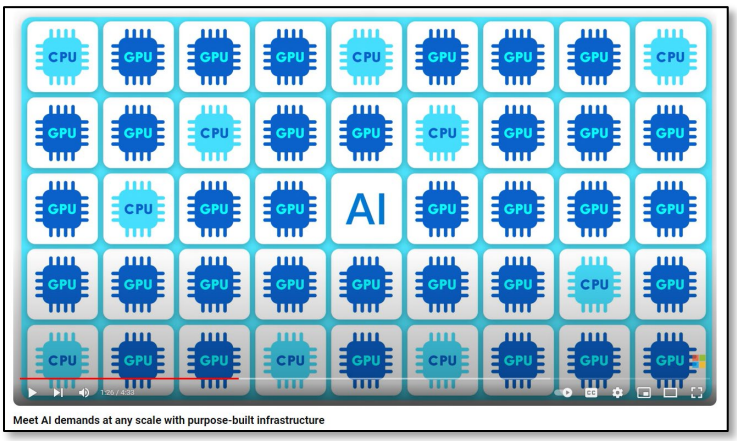
119. The hardware computation nodes and the hardware boosters each interface with a communication infrastructure. For example, both the CPU cores and GPUs of the Azure AI infrastructure interface using InfiniBand. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-

microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]he supercomputer that we built for OpenAI back in 2020 comprises more than 285,000 AMD InfiniBand connected CPU cores"); *id.* ("[T]here are 10,000 NVIDIA V100 Tensor Core GPUs that are also InfiniBand connected.").



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

120. The resource manager is arranged to assign a selected hardware booster to a hardware computation node for computation of a part of the computation task. Microsoft in videos describing its AI infrastructure depicts computation nodes (*e.g.*, CPU cores) and boosters or accelerators (*e.g.*, GPUs) working together in harmony to perform complex computing tasks.



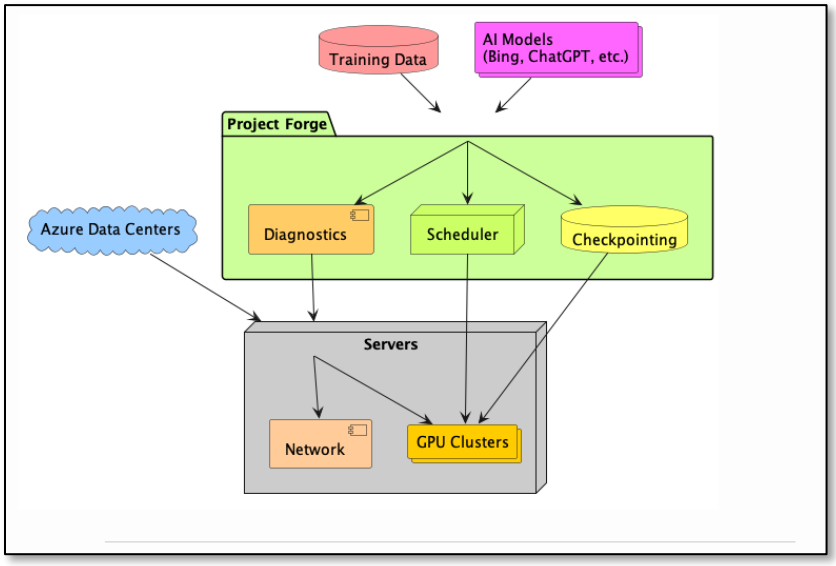
Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

121. Microsoft further touts that the GPUs of the Azure AI system provide the ability to efficiently accelerate their work on powerful systems. *See* Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024) ("H100 and Quantum-2 are part of NVIDIA's high-performance computing (HPC) platform—a full technology stack with CPUs, GPUs, DPUs, systems, networking, and a broad range of AI and HPC software—that provides researchers the ability to efficiently accelerate their work on powerful systems, on-premises or in the cloud."). Code, such as Python code running on CPUs, handle the control.

122. The resource manager is arranged to provide assignment information to the first hardware computation node after the assignment of the selected hardware booster so as to enable the first hardware computation node to outsource the part of the computation task to the assigned selected hardware booster under control of the first hardware computation node. The specific details regarding the GPU (like device IDs) are communicated to the container instance through environmental variables or configuration files generated dynamically. The Python application, now aware of the GPU, can offload specific computational tasks to the GPU.

123. The resource manager is arranged to initialize a static assignment and further to establish a dynamic assignment during the processing of the computation task. "Project Forge"—previously called "Singularity"—is a global scheduler the pools GPU capacity from different regions to help run Microsoft's global scale AI workloads and maintain high levels of utilization. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("And the way we address that in Azure is with a containerization and global scheduler service that we've

been working on called Project Forge, which is designed specifically to help run Microsoft's global scale AI workloads and maintain really high levels of utilization. Project Forge introduces transparent checkpointing, where it periodically saves the state of a model incrementally, without the model's code needing to do anything. That way, if anything fails, it can quickly resume for the most recent checkpoint. We combine this with our integrated global scheduler that pools GPU capacity from regions around the world. So, if you need to pause a job to prioritize another one, that allows us to migrate that pause job to another region if necessary and available, with minimal impact on its progress."); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("At the heart of Singularity is a novel, workload-aware scheduler that can transparently preempt and elastically scale deep learning workloads to drive high utilization without impacting their correctness or performance across a global fleet of AI accelerators (e.g., GPUs, FPGAs).").



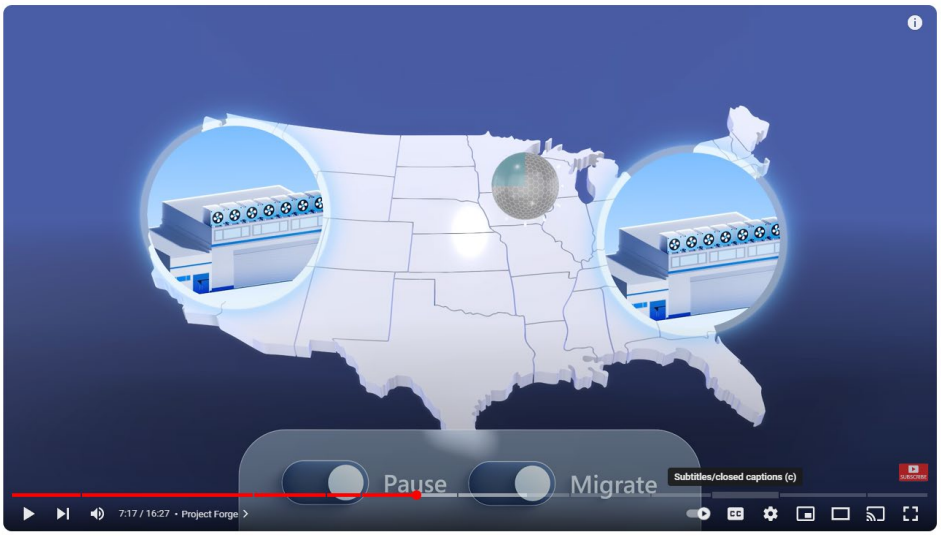
Source: Perez Rivas Consulting, *Inside Microsoft's AI Supercomputer Powering ChatGPT and Large Language Models* (July 26, 2023), https://www.dr-perez-rivas-consulting.com/2023/07/26/inside-microsoft-s-ai-supercomputer-powering-chatgpt-and-large-language-models/ (depicting Project Forge's relationship to the system)

124. The resource manager is arranged to accomplish the assignments as a function of a

predetermined assignment metric by providing the static assignment at the start of the processing of the computation task by using the predetermined assignment metric. For example, "[a] job arriving with a demand for N GPU (based on soft quota) may get more than N or fewer than N GPUs, depending on the competing cluster load." Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024); *see also id.* ("Further, it enables jobs with lower SLA to opportunistically use spare capacity and be quickly pre-empted (without lost work) when jobs with higher SLA arrive.").

125. Finally, the resource manager is arranged to accomplish the assignments as a function of a predetermined assignment metric by updating the predetermined assignment metric during the processing of the computation task. Project Forge (Singularity), for example, adapts to changes in load. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("So, we're going to train the model which just takes a short period of time and now the GPU is going to go idle. And at that point Project Forge sees that it's gone idle and automatically takes a checkpoint, because it knows that I'm not busy using the GPUs. Normal circumstances, I would've been hogging the GPU at that point and preventing it from being used for somebody else's workload. But now that GPU can be repurposed . . . ."); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("All jobs in Singularity are preemptible, migratable, and dynamically resizable (elastic) by default: a live job can be dynamically and transparently (a) pre-empted and migrated to a different set of nodes, cluster, data centre or a region and resumed exactly from the point where the execution was pre-empted, and (b) resized (i.e., elastically scaled up/down) on a

varying set of accelerators of a given type."); *id.* ("No idling of resources: Singularity treats the entire fleet of accelerators as a single logical, shared cluster, and avoids any resource fragmentation or static reservation of capacity."); *id.* ("For example, Singularity adapts to increasing load on an inference job, freeing up capacity by elastically scaling down or pre-empting training jobs."); *id.* ("Resizing/Elasticity: Singularity enables all jobs to be dynamically and elastically scaled up or down in a transparent manner to use a variable number of AI accelerators."); *id.* ("The binding between the job workers in Singularity and the accelerator devices is dynamic and constantly changing during the lifetime of the job.").

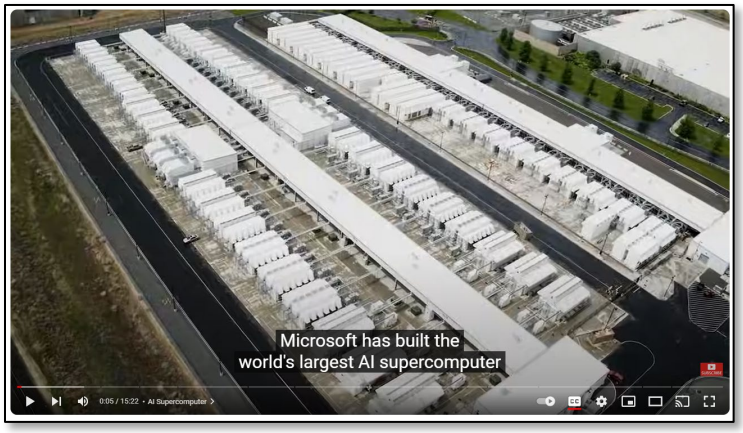


Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024) (depicting the migration of a job).

As Microsoft has touted, its Azure AI infrastructure "support[s] a modular approach to deploy whatever GPU demand calls for" "to take advantage of the best cost performance." Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024). As Microsoft further touts, its Azure AI infrastructure is "elastic, and able to quickly scale resources up or down to optimize operational costs." Microsoft Azure, *Meet AI Demands at any Scale with*

Purpose-Built Infrastructure, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024).

126. Microsoft directly infringes Claim 1 of the '883 Patent—and the '883 Patent more generally—through multiple infringing acts. For example,[9] Microsoft has combined all elements of the claimed system and thus "made" the system, as Microsoft admits:



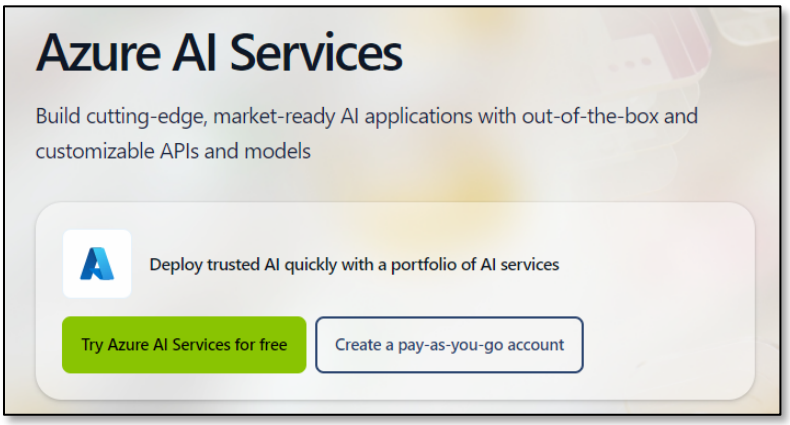
Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

127. Further, Microsoft uses the Azure AI infrastructure to run its own computation tasks and thus uses the system. *See* Timothy Prickett Morgan, *Inside the Infrastructure that Microsoft Builds to Run AI*, The Next Platform (Mar. 21, 2023), https://www.nextplatform.com/2023/03/21/inside-the-infrastructure-that-microsoft-builds-to-run-ai/ (Nidhi Chappell, Microsoft General Manager of Azure HPC and AI: "Whether it is internal teams running Bing, ChatGPT, or whatever – everything is running on Azure public infrastructure. . . . We use the same infrastructure, we make it available internally and externally."); Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*,

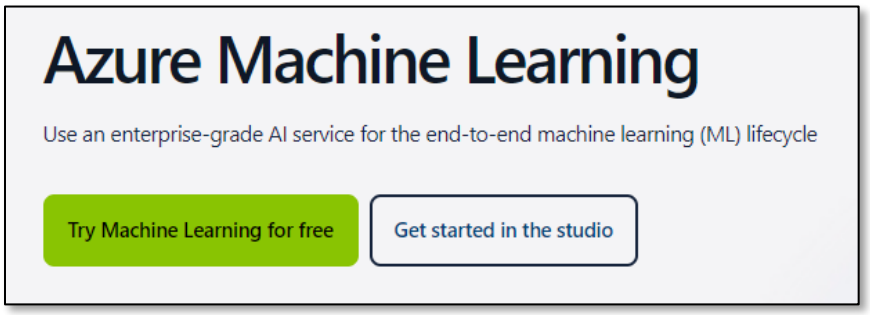
[9] The following infringing acts are illustrative and not intended to be an exhaustive or limiting list of each of Microsoft's infringing acts.

Yahoo! Finance (Mar. 13, 2023), https://finance.yahoo.com/news/microsoft-strung-together-tens-thousands-130035397.html ("Now Microsoft uses that same set of resources it built for OpenAI to train and run its own large artificial intelligence models, including the new Bing search bot introduced last month."). Microsoft also uses the Azure AI infrastructure, the system, to run the computation tasks of its customers. *Id.*

128. And Microsoft "sells the system to . . . customers." Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*, Yahoo! Finance (Mar. 13, 2023), https://finance.yahoo.com/news/microsoft-strung-together-tens-thousands-130035397.html. It relatedly, and necessarily, offers to sell the system:



Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 8, 2024, in Longview, Texas)



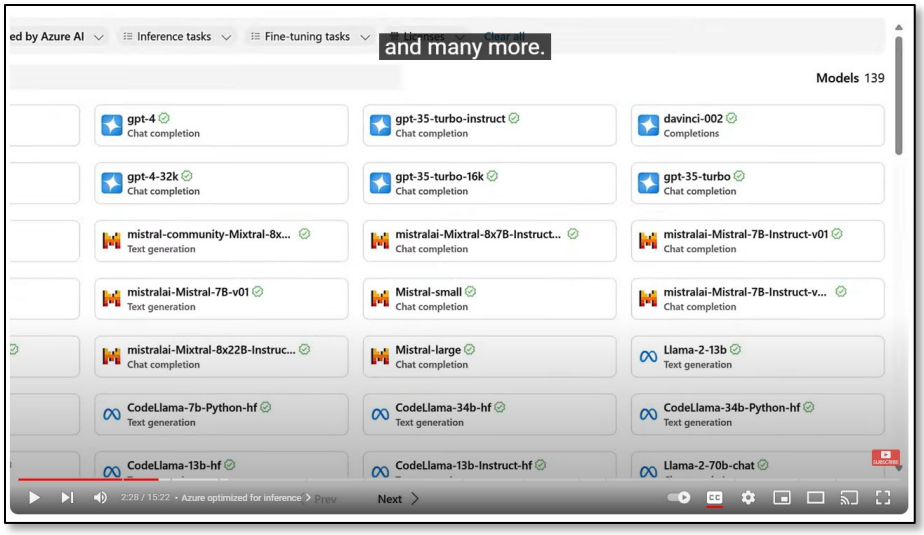
Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed on June 8, 2024, in Longview, Texas)

129. In addition to directly infringing the ’883 Patent by making, using, selling, offering to sell, and/or importing infringing products into the United States, Microsoft also indirectly infringes one or more claims of the ’883 Patent. Where acts constituting direct infringement of the ’883 Patent may not be performed by Microsoft, such acts constituting direct infringement of the ’883 Patent are performed by Microsoft’s customers or end-users who act at the direction and/or control of Microsoft, with Microsoft’s knowledge.

130. Microsoft indirectly infringes one or more claims of the ’883 Patent by active inducement in violation of 35 U.S.C. § 271(b), by at least manufacturing, supplying, distributing, selling, and/or offering for sale the Microsoft Azure AI system to its clients with full knowledge and intent that use of the same would constitute direct infringement of the ’883 Patent.

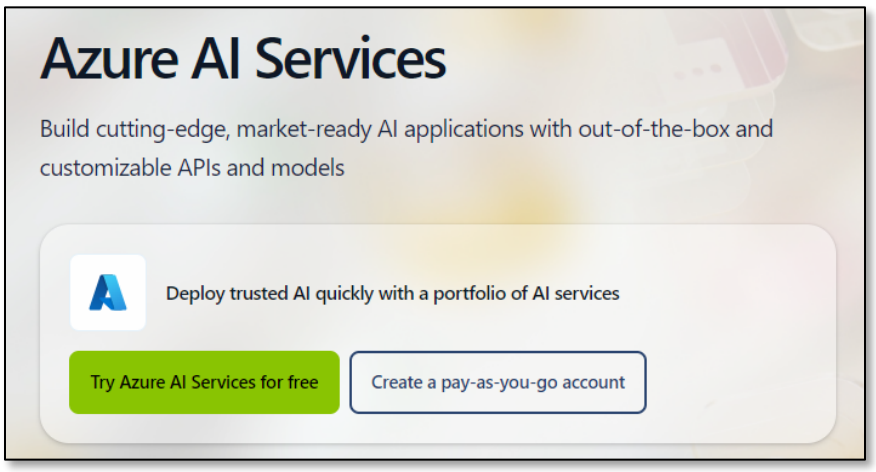
131. Microsoft has been aware of the ’883 Patent at least as of the filing of this Complaint.

132. Moreover, Microsoft intends to cause, and has taken affirmative steps to induce, infringement by customers and end-users by at least, *inter alia*, encouraging, promoting, instructing, and/or directing the infringing use of the Microsoft Azure AI system. For example, Microsoft’s “model as a service option in Azure” allows users to use Microsoft’s AI “infrastructure to access and run the most sophisticated AI models, such as GPT-3.5 Turbo, GPT-4, Meta’s Llama, Mistral, and many more,” which Microsoft has touted as important for organizations that do not have the resources to build out and train their own AI models.



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

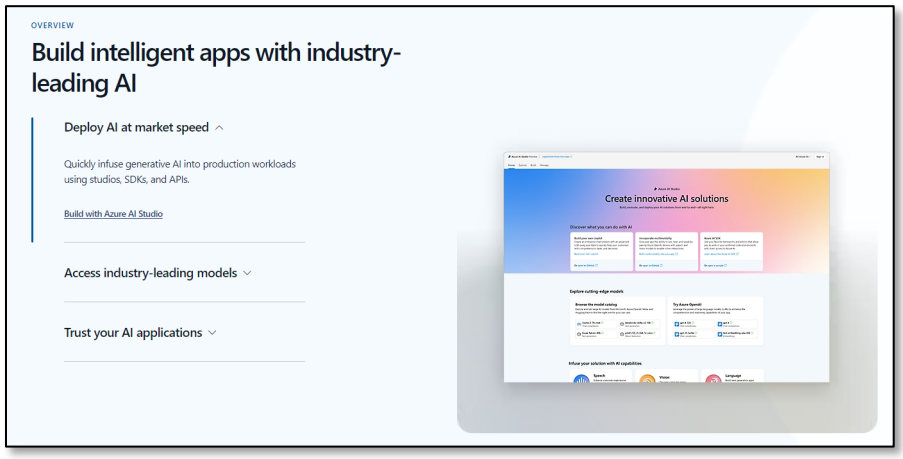
133. Microsoft encourages users to use and take advantage of this option and Microsoft's Azure AI infrastructure more generally. Microsoft promotes its Azure AI services, and encourages users and customers to use these services, to "[b]uild cutting-edge, market-ready AI applications with out-of-the-box and customizable APIs and models."



Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

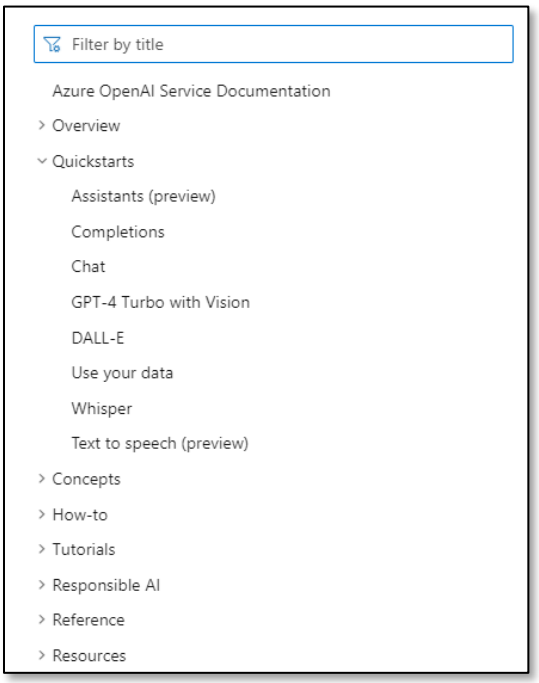
134. Microsoft further promotes its Azure AI services as allowing customers to "build

intelligent apps with industry-leading AI," including "deploy[ing] AI at market speed," "access[ing] industry-leading models," and "trust[ing] your AI applications."

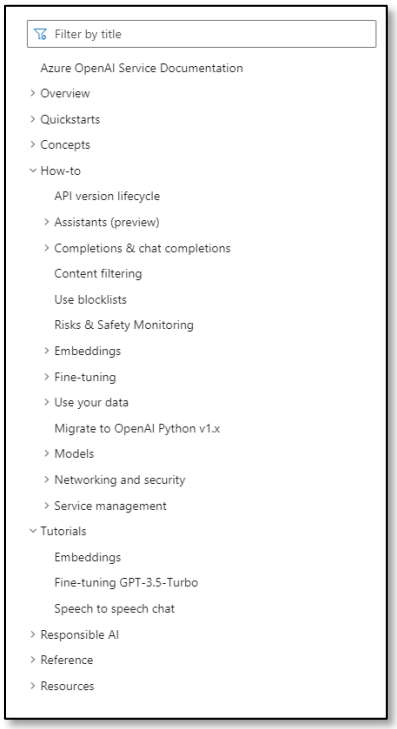


Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

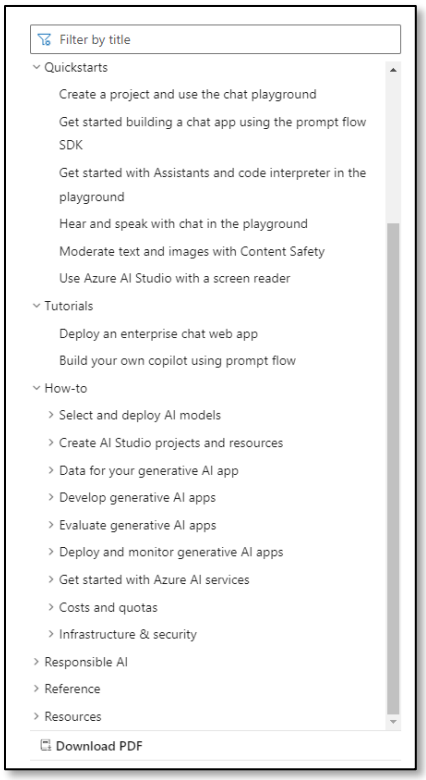
135. Microsoft also offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to use Azure AI:



Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)

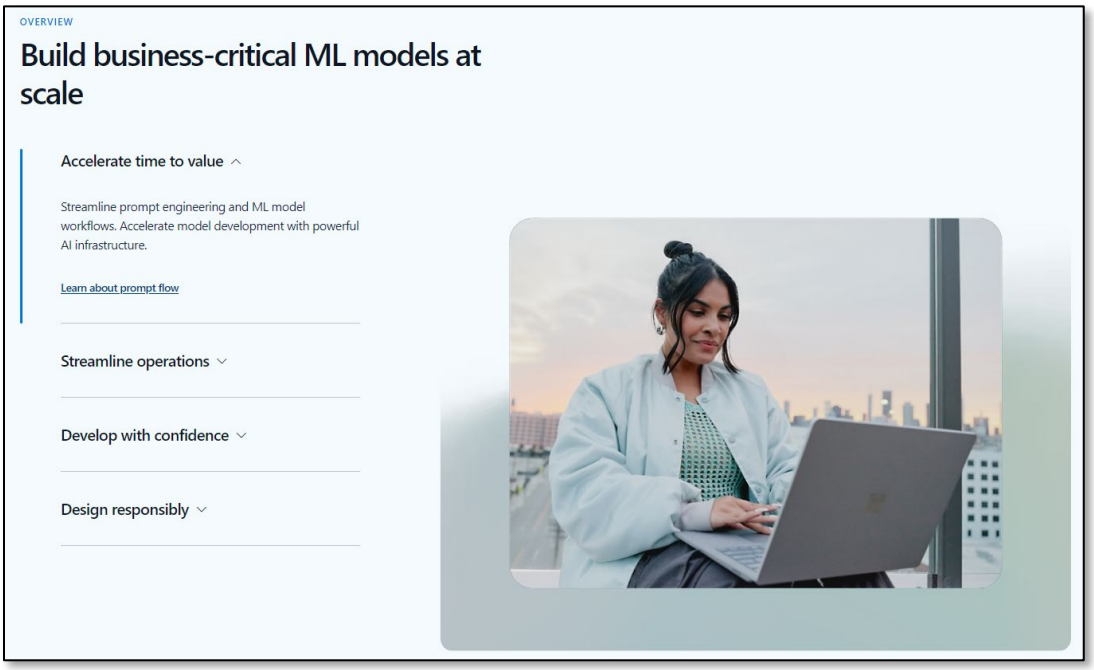


Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)



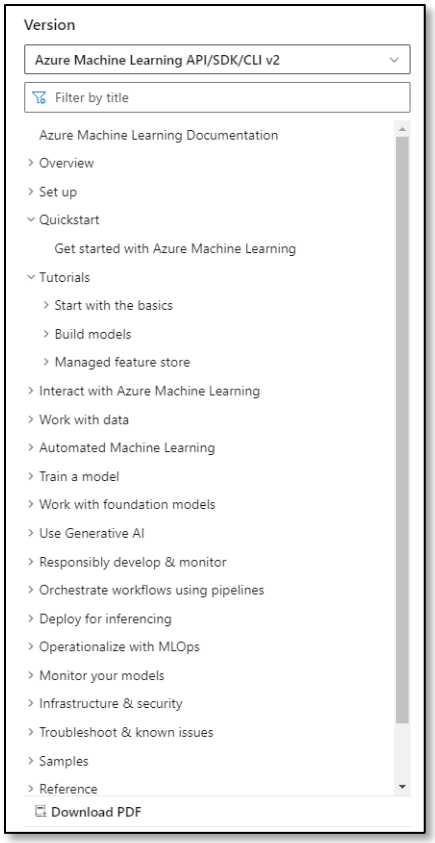
Source: Microsoft, *Azure AI Frequently Asked Questions*, https://learn.microsoft.com/en-us/azure/ai-studio/faq (last accessed June 7, 2024)

136. Similar to its Azure AI services, Microsoft actively promotes and induces use of Azure machine learning, which uses Microsoft's Azure AI infrastructure. Microsoft promotes Azure machine learning, and encourages users and customers to use the same, to "build business-critical [machine learning] models at scale," allowing customers to "accelerate time to value," "streamline operations," "develop with confidence," and "design responsibly."



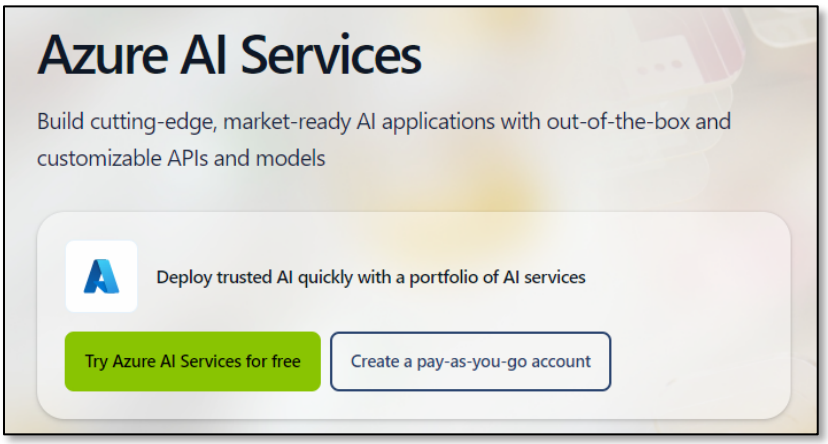
Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

137. Also, as with Azure AI services, Microsoft offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to its machine learning services.

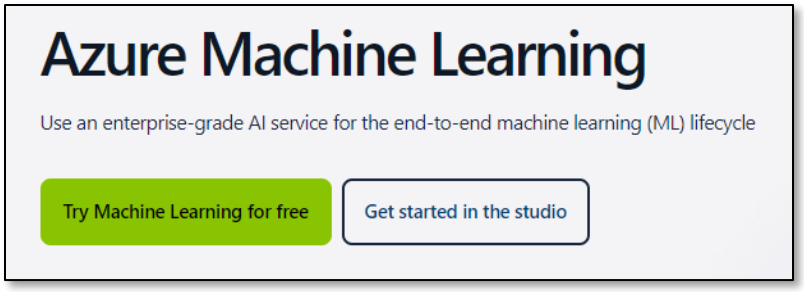


Source: Microsoft, *Azure Machine Learning Documentation*, https://learn.microsoft.com/en-us/azure/machine-learning/?view=azureml-api-2 (last visited June 7, 2024)

138. To encourage use of the foregoing products and services, Microsoft even offers free 30-day trials. After that, users pay as they go.

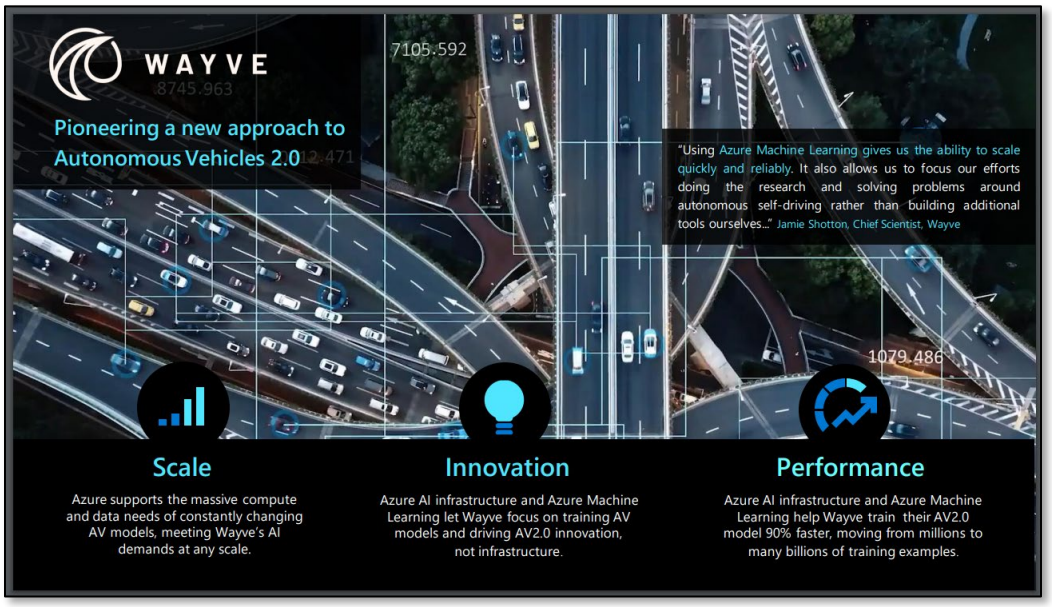


Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)



Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

139. Customers and end-users have heeded Microsoft's encouragement. Microsoft touts the various ways its customers have used Azure AI. *See* Microsoft, *AI Customer Stories*, https://www.microsoft.com/en-us/ai/ai-customer-stories (last accessed June 7, 2024) (collecting stories). Volvo, for example, uses Azure AI to "streamlin[e] invoice processing." *Id.*; *see also* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024) (explaining how Wayve—a self-driving car company—uses Azure AI).



Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024) (same)

140. As detailed above, Microsoft's Azure AI infrastructure infringes at least Claim 1 of the '883 Patent. Accordingly, by encouraging, promoting, instructing, and/or directing users to use Microsoft Azure AI, Microsoft is actively inducing infringement of the '883 Patent in violation of 35 U.S.C. § 271(b).

141. Microsoft also indirectly infringes one or more claims of the '883 Patent by contributory infringement in violation of 35 U.S.C. § 271(c). Microsoft is aware that components of Microsoft Azure are a material and substantial part of the invention claimed by the '883 patent, and that they are designed for a use that is both patented and infringing, and that has no substantial non-infringing uses.

142. Microsoft's acts of infringement have caused damage to Plaintiffs, and Plaintiffs are entitled to recover from Microsoft (or any successor entity to Microsoft) the damages sustained by Plaintiffs as a result of Microsoft's wrongful acts in an amount subject to proof at trial.

143. To the extent applicable, Plaintiff has complied with 35 U.S.C. § 287(a) with respect to the '883 Patent.

COUNT THREE
INFRINGEMENT OF U.S. PATENT NO. 11,537,442

144. Plaintiffs repeat and incorporate by reference each preceding paragraph as if fully set forth herein and further state:

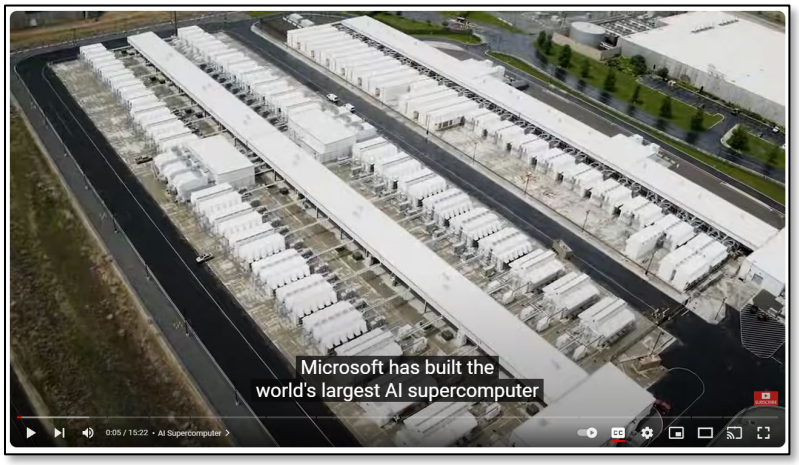
145. Microsoft has infringed and continues to directly infringe the '442 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, or offering for sale in the United States, and/or importing into the United States, without authorization, systems and methods that practice claims of the '442 Patent, including the Microsoft Azure AI system.

146. For example, Claim 1 is illustrative of the claims of the '442 Patent. It recites "[a]

method of operating a heterogeneous computing system comprising a plurality of computation nodes and a plurality of booster nodes, at least one of the plurality of computation nodes and a plurality of booster nodes being arranged to compute a computation task, the computation task comprising a plurality of sub-tasks, the method comprising:

> in a first computing iteration, assigning and processing the plurality of sub-tasks by at least a portion of the plurality of computation nodes and at least a portion of the plurality of booster nodes in a first distribution; and
>
> generating, using information relating to the processing of the plurality of sub-tasks by at least the portion of the plurality of computation nodes and at least the portion of the plurality of booster nodes, a further distribution of the plurality of sub-tasks between the plurality of computation nodes and the plurality of booster nodes for processing thereby in a further computing iteration."

147. Microsoft's Azure AI system meets every element of this claim.[10] Microsoft touts Azure as "the world's largest AI supercomputer:"



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

148. The Azure AI system includes "GPUs, networking, and the full stack of AI software." Microsoft, *Azure Blog / AI + Machine Learning*, https://azure.microsoft.com/en-

[10] This description of infringement is illustrative and not intended to be an exhaustive or limiting explanation of every manner in which Microsoft Azure AI system infringes.

us/blog/roundup-of-ai-breakthroughs-by-microsoft-and-nvidia/ (last accessed June 7, 2024). Further, it includes "a full technology stack with CPUs, GPUs, DPUs, systems, [and] networking." *Id.* Hence, the Azure AI infrastructure is a heterogeneous computing system.

149. Microsoft Azure includes a plurality of computation nodes—for example, CPU cores. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]he supercomputer that we built for OpenAI back in 2020 comprises more than 285,000 AMD InfiniBand connected CPU cores"). In 2020, Azure had 285,000 CPU cores. By November 2023, that number had ballooned to 1.1 million.

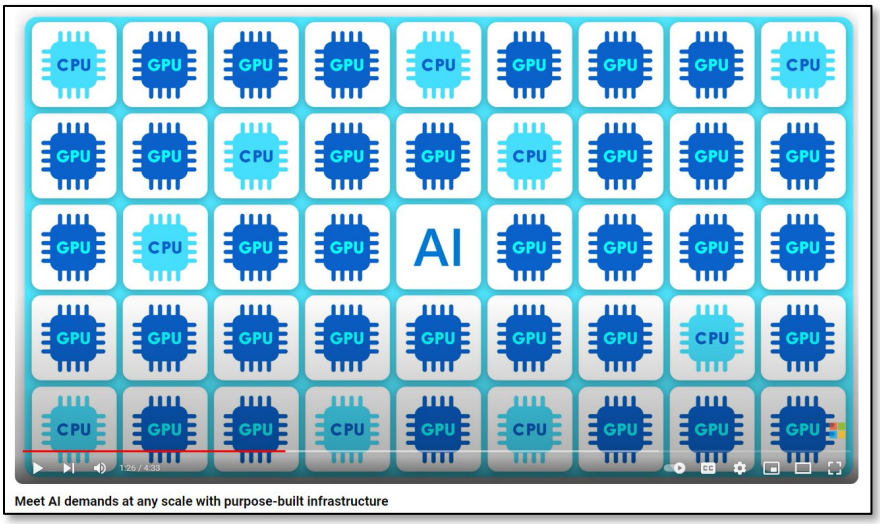
Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

150. Microsoft Azure includes a plurality of booster nodes—for example, GPUs. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("[T]here

are 10,000 NVIDIA V100 Tensor Core GPUs that are also InfiniBand connected."). In 2020, the Azure AI system had 10,000 GPUs. By November 2023, it had 14,400 GPUs.

Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

151. The computation nodes and booster nodes can be arranged to compute a computation task. In Azure, multiple CPUs or CPU cores, for example can work in parallel, each leveraging GPU acceleration, as Microsoft demonstrates in videos describing its AI infrastructure.



Source: Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024)

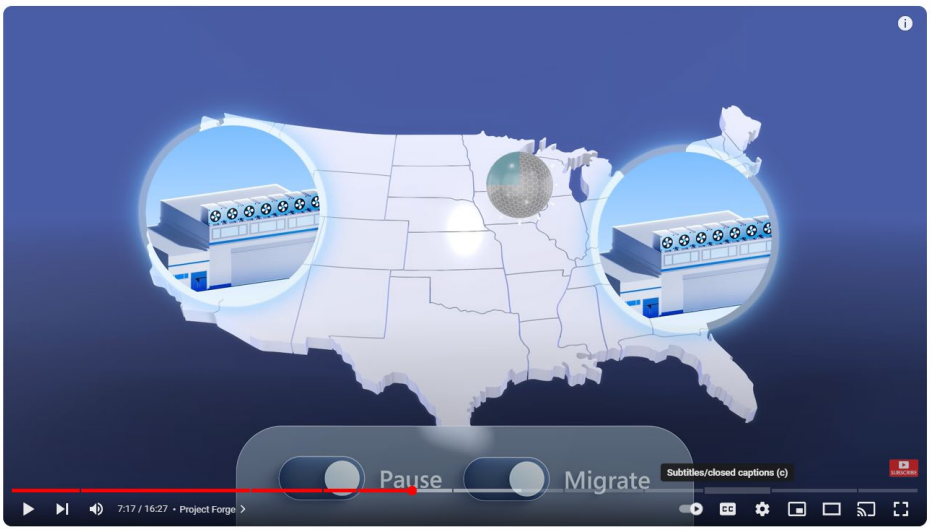
152. The Microsoft Azure AI infrastructure excels by allowing users to write code specific to a pre-defined number of workers. *See* Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) ("[T]he user writes code for a constant world-size (i.e., number of workers), and is oblivious to how many physical devices Singularity places the job on."). Hence, the computation task comprises sub-tasks.

153. The Microsoft Azure AI system, in a first computing iteration, assigns and processes the plurality of sub-tasks by at least a portion of the plurality of computation nodes and at least a portion of the plurality of booster nodes in a first distribution. For example, "[a] job arriving with a demand for N GPU (based on soft quota) may get more than N or fewer than N GPUs, depending on the competing cluster load." Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024); *see also* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 ("And the way we address that in Azure is with a containerization and global scheduler service that we've been working on called Project Forge, which is designed specifically to help run Microsoft's global scale AI workloads and maintain really high levels of utilization. Project Forge introduces transparent checkpointing, where it periodically saves the state of a model incrementally, without the model's code needing to do anything. That way, if anything fails, it can quickly resume for the most recent checkpoint. We combine this with our integrated global scheduler that pools GPU capacity from regions around the world. So, if you need to pause a job to prioritize another one, that allows us to migrate that pause job to another region if necessary and

available, with minimal impact on its progress.”).

154. Finally, the Azure AI infrastructure generates, using information relating to the processing of the plurality of subtasks by at least the portion of the plurality of computation nodes and at least the portion of the plurality of booster nodes, a further distribution of the plurality of sub-tasks between the plurality of computation nodes and the plurality of booster nodes for processing thereby in a further computing iteration. Project Forge (“Singularity”) adapts to increasing load, freeing up capacity by elastically scaling down or pre-empting training jobs. It enables all jobs to be dynamically and elastically scaled up or down in a transparent manner to use a variable number of AI accelerators. *See* Microsoft, *What Runs ChatGPT? Inside Microsoft’s AI Supercomputer | Featuring Mark Russinovich* (May 24, 2023), https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/what-runs-chatgpt-inside-microsoft-s-ai-supercomputer-featuring/ba-p/3830281 (“We combine this with our integrated global scheduler that pools GPU capacity from regions around the world. So, if you need to pause a job to prioritize another one, that allows us to migrate that pause job to another region if necessary and available, with minimal impact on its progress.”); Microsoft, *Singularity: Plant-Scale, Preemptive and Elastic Scheduling of AI Workloads*, https://arxiv.org/pdf/2202.07848 (last accessed June 7, 2024) (“At the heart of Singularity is a novel, workload-aware scheduler that can transparently preempt and elastically scale deep learning workloads to drive high utilization without impacting their correctness or performance across a global fleet of AI accelerators (e.g., GPUs, FPGAs).”); *id.* (“All jobs in Singularity are preemptible, migratable, and dynamically resizable (elastic) by default: a live job can be dynamically and transparently (a) pre-empted and migrated to a different set of nodes, cluster, data centre or a region and resumed exactly from the point where the execution was pre-empted, and (b) resized (i.e., elastically scaled up/down) on a

varying set of accelerators of a given type."); *id.* ("For example, Singularity adapts to increasing load on an inference job, freeing up capacity by elastically scaling down or pre-empting training jobs."); *id.* ("Resizing/Elasticity: Singularity enables all jobs to be dynamically and elastically scaled up or down in a transparent manner to use a variable number of AI accelerators."); *id.* ("The binding between the job workers in Singularity and the accelerator devices is dynamic and constantly changing during the lifetime of the job.")



Source: Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (depicting the migration of a job)

As Microsoft has touted, its Azure AI infrastructure "support[s] a modular approach to deploy whatever GPU demand calls for" "to take advantage of the best cost performance." Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024). As Microsoft further touts, its Azure AI infrastructure is "elastic, and able to quickly scale resources up or down to optimize operational costs." Microsoft Azure, *Meet AI Demands at any Scale with Purpose-Built Infrastructure*, https://www.youtube.com/watch?v=eq2zQq2GtFQ (last accessed June 7, 2024).

155. Microsoft directly infringes Claim 1 of the '442 Patent—and the '442 Patent more generally—through multiple infringing acts. For example,[11] Microsoft uses the Azure AI infrastructure to run its own computation tasks, performing each step of the claimed method in the process. *See* Timothy Prickett Morgan, *Inside the Infrastructure that Microsoft Builds to Run AI*, The Next Platform (Mar. 21, 2023), https://www.nextplatform.com/2023/03/21/inside-the-infrastructure-that-microsoft-builds-to-run-ai/ (Nidhi Chappell, Microsoft General Manager of Azure HPC and AI: "Whether it is internal teams running Bing, ChatGPT, or whatever – everything is running on Azure public infrastructure. . . . We use the same infrastructure, we make it available internally and externally."); Dina Bass, *Microsoft Strung Together Tens of Thousands of Chips in a Pricey Supercomputer for OpenAI*, Yahoo! Finance (Mar. 13, 2023), https://finance.yahoo.com/news/microsoft-strung-together-tens-thousands-130035397.html ("Now Microsoft uses that same set of resources it built for OpenAI to train and run its own large artificial intelligence models, including the new Bing search bot introduced last month."). Microsoft also uses the Azure AI infrastructure to run the computation tasks of its customers—again, performing each step of the claimed method. *Id.*

156. In addition to directly infringing the '442 Patent by making, using, selling, offering to sell, and/or importing infringing products into the United States, Microsoft also indirectly infringes one or more claims of the '442 Patent. Where acts constituting direct infringement of the '442 Patent may not be performed by Microsoft, such acts constituting direct infringement of the '442 Patent are performed by Microsoft's customers or end-users who act at the direction and/or control of Microsoft, with Microsoft's knowledge.

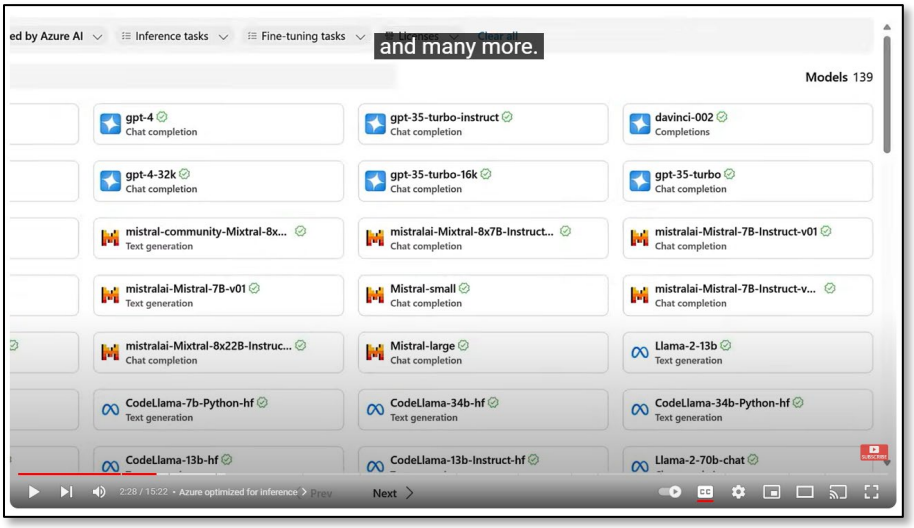
157. Microsoft indirectly infringes one or more claims of the '442 Patent by active

[11] The following infringing acts are illustrative and not intended to be an exhaustive or limiting list of each of Microsoft's infringing acts.

inducement in violation of 35 U.S.C. § 271(b), by at least manufacturing, supplying, distributing, selling, and/or offering for sale the Microsoft Azure AI system to its clients with full knowledge and intent that use of the same would constitute direct infringement of the ’442 Patent.

158. Microsoft has been aware of the ’442 Patent at least as of the filing of this Complaint.

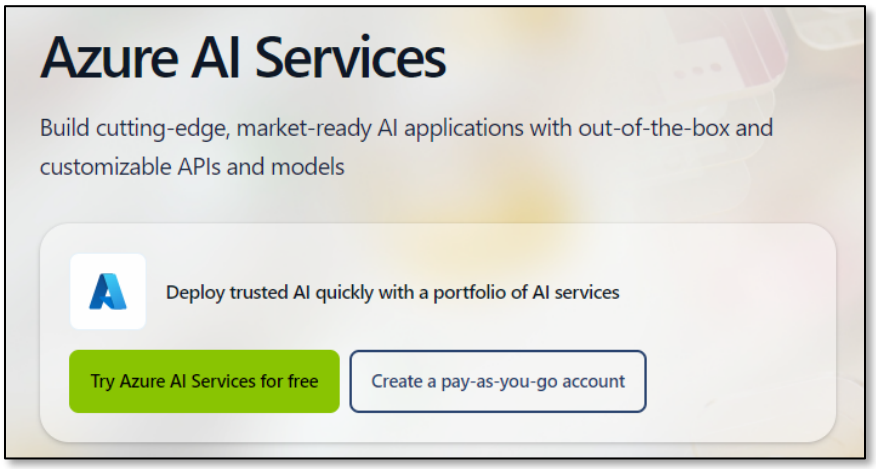
159. Moreover, Microsoft intends to cause, and has taken affirmative steps to induce, infringement by customers and end-users by at least, *inter alia*, encouraging, promoting, instructing, and/or directing the infringing use of the Microsoft Azure AI system. For example, Microsoft’s “model as a service option in Azure” allows users to use Microsoft’s AI “infrastructure to access and run the most sophisticated AI models, such as GPT-3.5 Turbo, GPT-4, Meta’s Llama, Mistral, and many more,” which Microsoft has touted as important for organizations that do not have the resources to build out and train their own AI models.



Source: Microsoft Mechanics, *What runs GPT-4o? Inside the Biggest AI Supercomputer in the Cloud with Mark Russinovich*, https://www.youtube.com/watch?v=DlX3QVFUtQI (last accessed June 7, 2024)

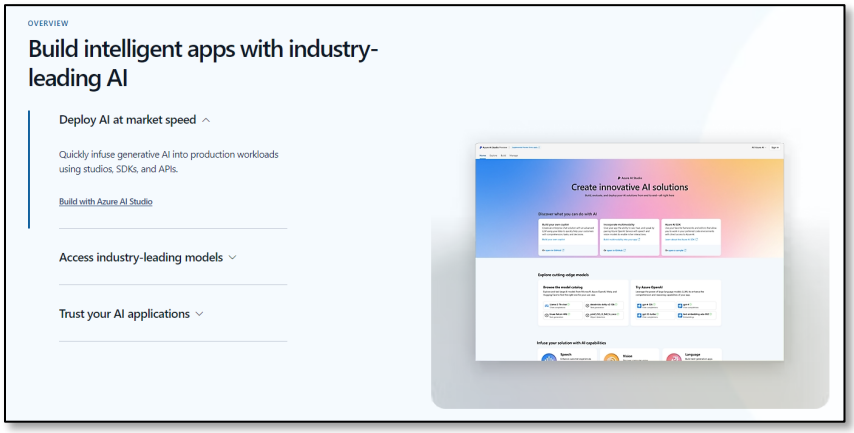
160. Microsoft encourages users to use and take advantage of this option and Microsoft’s Azure AI infrastructure more generally. Microsoft promotes its Azure AI services, and encourages

users and customers to use these services, to "[b]uild cutting-edge, market-ready AI applications with out-of-the-box and customizable APIs and models."



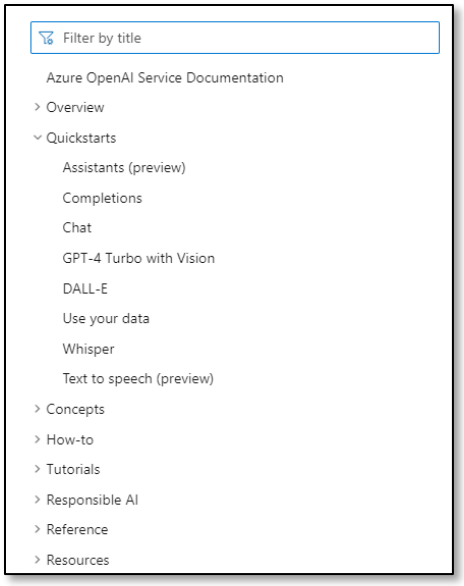
Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

161. Microsoft further promotes its Azure AI services as allowing customers to "build intelligent apps with industry-leading AI," including "deploy[ing] AI at market speed," "access[ing] industry-leading models," and "trust[ing] your AI applications."

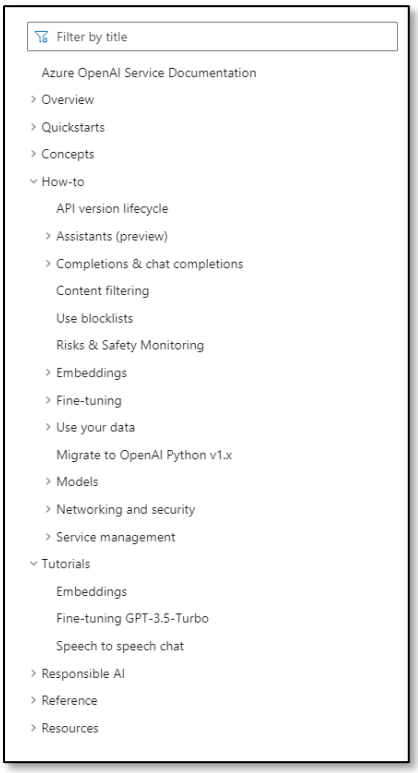


Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)

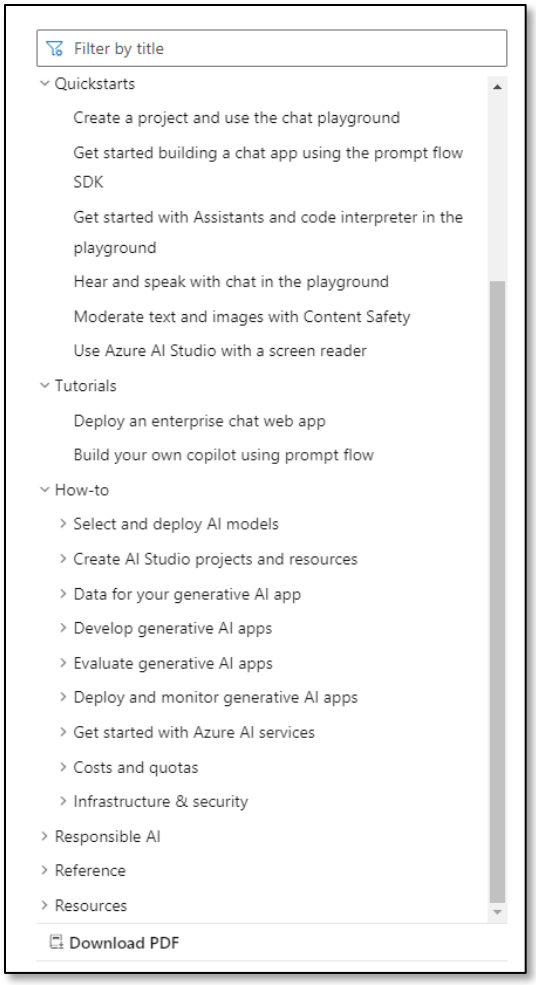
162. Microsoft also offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to use Azure AI:



Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)

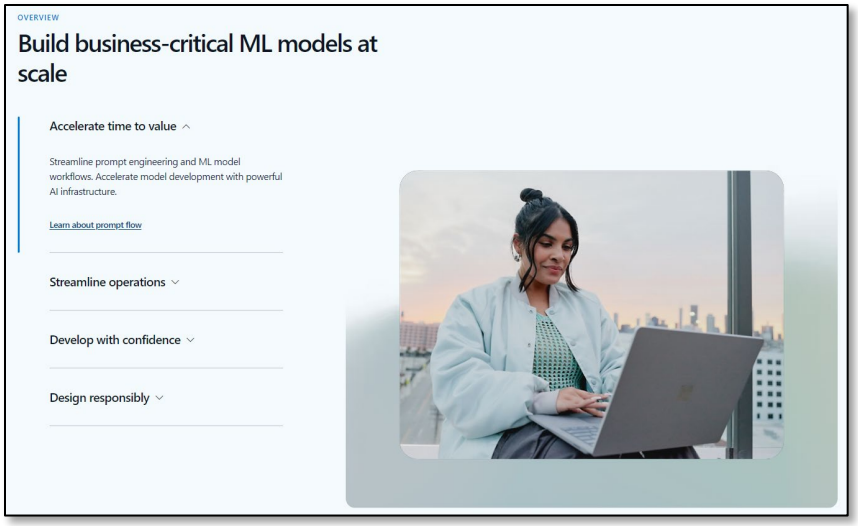


Source: Microsoft, *What is Azure OpenAI Service?*, https://learn.microsoft.com/en-us/azure/ai-services/openai/overview (last accessed June 7, 2024)



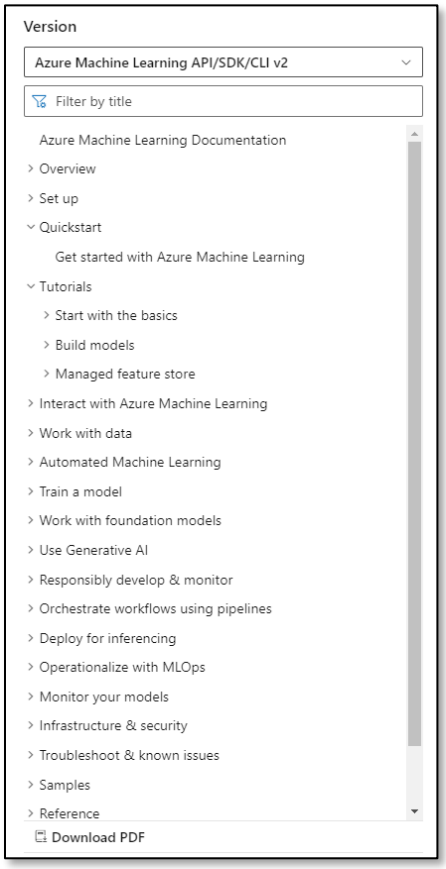
Source: Microsoft, *Azure AI Frequently Asked Questions*, https://learn.microsoft.com/en-us/azure/ai-studio/faq (last accessed June 7, 2024)

163. Similar to its Azure AI services, Microsoft actively promotes and induces use of Azure machine learning, which uses the Azure AI infrastructure. Microsoft promotes Azure machine learning, and encourages users and customers to use the same, to "build business-critical [machine learning] models at scale," allowing customers to "accelerate time to value," "streamline operations," "develop with confidence," and "design responsibly."



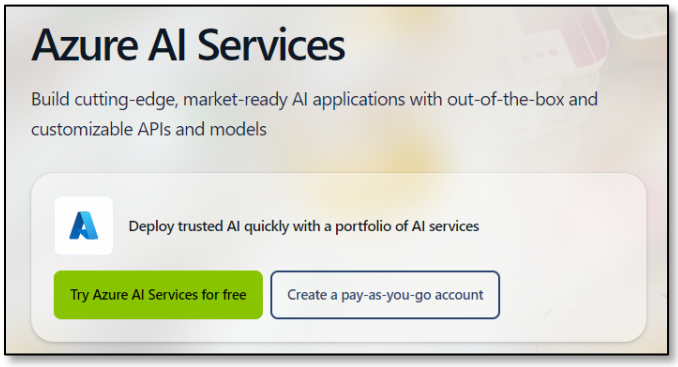
Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

164. Also, as with Azure AI services, Microsoft offers a host of quickstart guides, tutorials, and how-tos explaining to customers and users how to use its machine learning services.

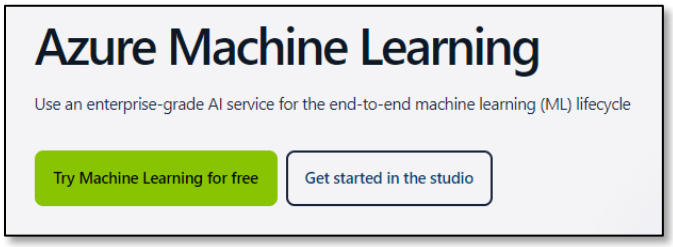


Source: Microsoft, *Azure Machine Learning Documentation*, https://learn.microsoft.com/en-us/azure/machine-learning/?view=azureml-api-2 (last visited June 7, 2024)

165. To encourage use of the foregoing products and services, Microsoft even offers free 30-day trials. After that, users pay as they go.

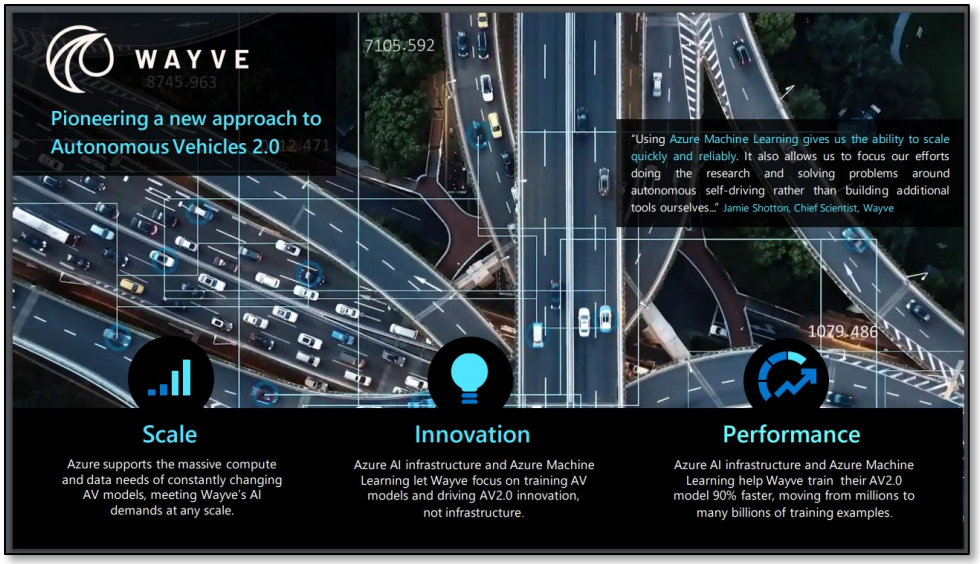


Source: Microsoft, *Azure AI Services*, https://azure.microsoft.com/en-us/products/ai-services/ (last accessed on June 7, 2024)



Source: Microsoft, *Azure Machine Learning*, https://azure.microsoft.com/en-us/products/machine-learning/ (last accessed June 7, 2024)

166. Customers and end-users have heeded Microsoft's encouragement. Microsoft touts the various ways its customers have used Azure AI. *See* Microsoft, *AI Customer Stories*, https://www.microsoft.com/en-us/ai/ai-customer-stories (last accessed June 7, 2024) (collecting stories). Volvo, for example, uses Azure AI to "streamlin[e] invoice processing." *Id.*; *see also* Microsoft, *What Runs ChatGPT? Inside Microsoft's AI Supercomputer*, https://www.youtube.com/watch?v=Rk3nTUfRZmo (last accessed June 7, 2024) (explaining how Wayve—a self-driving car company—uses Azure AI).



Source: Microsoft, *Intro to AI Infrastructure*, https://microsoft.github.io/PartnerResources/assets/msa/AI%20Video.pdf (last accessed June 7, 2024) (same)

167. As detailed above, Microsoft's Azure AI infrastructure infringes at least Claim 1 of the '442 Patent. Accordingly, by encouraging, promoting, instructing, and/or directing users to use Microsoft Azure AI, Microsoft is actively inducing infringement of the '442 Patent in violation of 35 U.S.C. § 271(b).

168. Microsoft also indirectly infringes one or more claims of the '442 Patent by contributory infringement in violation of 35 U.S.C. § 271(c). Microsoft is aware that components of Microsoft Azure are a material and substantial part of the invention claimed by the '442 patent, and that they are designed for a use that is both patented and infringing, and that has no substantial non-infringing uses.

169. Microsoft's acts of infringement have caused damage to Plaintiffs, and Plaintiffs are entitled to recover from Microsoft (or any successor entity to Microsoft) the damages sustained by Plaintiffs as a result of Microsoft's wrongful acts in an amount subject to proof at trial.

170. To the extent applicable, Plaintiff has complied with 35 U.S.C. § 287(a) with respect to the '442 Patent.

DEMAND FOR JURY TRIAL

171. Plaintiffs hereby demand a jury trial for all issues so triable.

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs request entry of judgment in their favor and against Defendant Microsoft as follows:

A. Declaring that Microsoft has infringed United States Patent Nos. 10,142,156, 11,934,883, and 11,537,442;

B. Awarding lost profits and/or reasonable royalty damages to Plaintiffs in an amount no less than a reasonable royalty for Microsoft's infringement of the Asserted Patents, together with prejudgment and post-judgment interest and costs as permitted under 35 U.S.C. § 284;

C. Awarding attorneys' fees pursuant to 35 U.S.C. § 285 or as otherwise permitted by law;

D. Ordering Microsoft to pay supplemental damages to Plaintiffs, including any ongoing royalties and interest, with an accounting, as needed;

E. Enjoining Microsoft from practicing the Asserted Patents; and

F. Awarding such other costs and further relief as the Court may deem just and proper.

Dated: June 10, 2024

Respectfully submitted,

/s/ Justin Nelson w/permission Claire Abernathy Henry

Justin A. Nelson – Lead Counsel
Texas State Bar No. 24034766
SUSMAN GODFREY, L.L.P.
1000 Louisiana Street, Suite 5100
Houston, Texas 77002
Telephone: (713) 651-9366
Facsimile: (713) 654-6666
jnelson@susmangodfrey.com

Matthew R. Berry
Washington State Bar No. 37364
Alexander W. Aiken
Washington State Bar No. 55988
SUSMAN GODFREY, L.L.P.
401 Union St., Suite 3000
Seattle, Washington 98101
Telephone: (206) 516-3880
Facsimile: (206) 516-3883
mberry@susmangodfrey.com
aaiken@susmangodfrey.com

S. Calvin Capshaw
Texas State Bar No. 03783900
CAPSHAW DERIEUX LLP
114 E. Commerce Ave.
Gladewater, TX 75647
Telephone: (903) 845-5770
ccapshaw@capshawlaw.com

Of Counsel:

Claire Abernathy Henry
Texas State Bar No. 24053063
WARD, SMITH & HILL, PLLC
1507 Bill Owens Parkway
Longview, TX 75604
Telephone: (903) 757-6400
Fax: (903) 757-2323
claire@wsfirm.com

Attorneys for Plaintiffs ParTec AG and BF exaQC AG