IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF TEXAS
AUSTIN DIVISION

XOCKETS, INC.,

Plaintiff,

v.

AMAZON.COM, INC. and AMAZON WEB SERVICES, INC.
Defendants.

Civil Action No. 1:25-cv-1021

JURY TRIAL DEMANDED

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff Xockets, Inc. ("Xockets") submits this Complaint for patent infringement against Defendants Amazon.com, Inc. and Amazon Web Services, Inc. (collectively, "Amazon").

BACKGROUND AND OVERVIEW

1. This Complaint asserts infringement of the following United States Patents owned by Xockets, disclosing a new cloud fabric and switching plane architecture: U.S. Patent No. 10,223,297 ("'297 Patent") (DPU Cloud Network Fabric), U.S. Patent No. 9,378,161 (the "'161 Patent") (DPU Cloud Network Fabric), U.S. Patent No. 10,212,092 (the "'092 Patent") (DPU In-Network Computing), and U.S. Patent No. 9,436,640 (the "'640 Patent") (DPU In-Network Computing) (collectively, the "Asserted Patents").

2. This case involves the rampant and widescale exploitation by Amazon of the exclusive intellectual property rights granted to Xockets in these patents for its new cloud fabric and switching plane architecture. Amazon has harnessed that which it did not invent, so it could rise to the forefront of modern cloud infrastructure and services. Amazon claims to be the leader in a competitive race to support the exponentially growing footprint and need for distributed computing infrastructure and services, including operating machine learning and artificial

intelligence applications for customers. To maintain its market dominance along the way, Amazon willfully trampled on the exclusive property rights of Xockets, a visionary startup that Amazon has known about for approximately a decade.

3. **Dr. Parin Dalal and Xockets, Inc.** In 2011, Dr. Parin Dalal invented a new, groundbreaking cloud processor and computing architecture. Anticipating the need for novel processing solutions for the rapidly growing data demands in cloud-based computing, Dr. Dalal invented what is now known as a DPU, or Data Processing Unit. He invented this advanced DPU by designing a new virtual switch computing architecture to offload, accelerate, and isolate data-intensive workloads, enabling accelerated computing in modern cloud data centers.

4. With the development of this new cloud processor, Xockets has enabled the transition now underway of every data center to an AI data center, ushering in a new age of innovation unparalleled in history.

5. Using and building upon that new DPU computing architecture, Dr. Dalal also designed an entirely new cloud fabric, deploying an innovative switching architecture in each DPU to manage, facilitate, and dramatically speed up collective communication among server processors (CPUs, GPUs, and hybrids of these server processors) and enable in-network computing operations on data-intensive workloads independent of the existing cloud network fabric. The efficiency gains in distributed computing from Xockets' new cloud fabric have forever changed the cloud industry. Training machine learning models, such as Large-Language Models, which would have taken many years without Xockets' new cloud fabric technology, is now performed in weeks or months.

6. Today's revolutions in machine learning and artificial intelligence ("ML/AI") depend fundamentally on the massive offloading and acceleration power of DPUs and the new

cloud fabric invented by Dr. Dalal. It is no exaggeration to say that Amazon's modern cloud infrastructure and services—and the ongoing ML/AI boom—would not be possible without Dr. Dalal's inventions.

7. In 2012, Dr. Dalal founded Xockets, where he and his team went on to design, develop, and build the world's first-ever advanced DPUs. In May 2012, Xockets filed the first of many patent applications disclosing Xockets' DPU and its virtual switch computing architecture for a new cloud processor, and a switching plane architecture for forming a new cloud fabric that operates independent of the existing cloud network fabric.

8. The cloud industry, meanwhile, headed in the wrong direction for years—trying to process data-intensive workloads using software running on server processors—until Dr. Dalal's DPU inventions were revealed in public demonstrations and patent filings. Dr. Dalal knew that exponential growth in the data processing required in clouds would eventually overwhelm the conventional computing architectures of server processors in performing their core computing functions in running applications for customers, resulting in systemic inefficiencies, increased costs, reduced performance for customers, and lower revenue for cloud providers.

9. Dr. Dalal and the Xockets team went against conventional wisdom in developing the DPU and its new virtual switch computing architecture, which moves data-intensive computing from server processors to programmable hardware acceleration pipelines in the DPU, for accelerating cloud infrastructure services, including offloading security, networking, and storage operations, and thereby frees up server processors for their core computing functions.

10. But Dr. Dalal and the Xockets team did not stop with offloading of compute functions alone. They also developed a new cloud *fabric* that reimagines how the computing components of a cloud network interact. Under the old paradigm, data competed for limited

bandwidth on the cloud network, leading to bottlenecking, reduced speed, and packet loss. Even as compute power steadily increased, the capacity of the standard cloud network infrastructure remained limiting. Dr. Dalal tackled this intractable networking problem by inventing a specific type of *second* switching plane using DPUs that allows data to bypass the traditional, constraining cloud infrastructure and unlock the power of distributed computing. That power has fueled today's ML/AI explosion.

11. **2015: Xockets' Strata demonstration.** In the fall of 2015, Xockets debuted DPUs to the industry at the Strata + Hadoop World Conference in New York City. Amazon Web Services ("AWS") attended Strata, hosting its own space just a few booths down from Xockets.

12. **2017: Amazon's "Deep Dive" into Xockets' technologies.** Xockets' DPU technologies—and their significant financial and strategic benefits— quickly grabbed the attention of major industry players, including Amazon. In May 2017, Dr. Dalal and the Xockets team met with AWS for what Amazon called a "Deep Dive" meeting. AWS was focused on investing in acceleration of its cloud infrastructure and services using DPUs and indicated that it was interested in acquiring Xockets. Under this premise, Xockets gave Amazon access to its design and development materials.

13. Amazon then conducted extensive diligence into Xockets' DPU technologies, knowing that mistakes in designing and developing DPUs could cost Amazon years of delay and future loss of its cloud business.[1] Amazon's AWS cloud engineers separately interviewed every member of the Xockets team to learn about its DPU technologies, its challenges, and its solutions. But, after learning everything it could about Xockets' groundbreaking DPU innovations, Amazon

[1] https://www.wsj.com/tech/amazon-ai-chips-supercomputer-aws-annapurna-trainium-a943be71?st=Tm4NKj&reflink=desktopwebshare_permalink ("If you make a mistake in software, it could cost you a week or two to fix it. . . . With hardware, you can lose nine months to a year.").

did not acquire Xockets. Nor did it seek to take a license to Xockets' patents. Instead, Amazon engineers built DPUs using Xockets' computing and switching architectures.

14. At the time of the 2017 Deep Dive meeting with Xockets, Amazon was hoping to develop a DPU computing architecture for Amazon's own cloud computing platform and services. And, following the Deep Dive meeting with Xockets, Amazon did exactly that by introducing its infringing AWS Nitro System ("Nitro DPU"). Within a year after the Xockets meeting, Amazon debuted "Nitro v3," which deploys in Amazon's own data centers the DPU virtual switch computing architecture invented by Dr. Dalal. Amazon shamelessly left Xockets behind after the Deep Dive meeting and turned its eye towards monetizing Xockets' patented inventions internally, without Xockets' knowledge or consent.

15. In the years since, Amazon has continued to develop and roll out additional infringing DPU technologies developed by Xockets. Amazon and has achieved staggering profits using Xockets' DPU technologies. Today's ML/AI revolution, enabled by Xockets' DPU technologies, has only expanded Amazon's gains.

16. On information and belief, Amazon has installed more than 20 million DPUs[2] across the servers in its data centers, with at least one DPU in each of its servers. Amazon has received and will continue to receive extraordinary financial and business benefits from the use of Xockets' DPU technologies.

17. On information and belief, there is more than $15,000 per server in cost savings from the use of DPUs in Amazon's data centers. In addition, increased revenues from DPU-enabled accelerated performance could be more than double the cost savings to cloud providers like Amazon, exceeding $30,000 per server over a three-year lifespan.

[2] https://www.geekwire.com/2023/inside-the-ai-chip-race-how-a-pivotal-happy-hour-changed-amazons-strategy-in-the-cloud.

18. **<u>2024: Xockets' Sales Process</u>.** In March of 2024, Xockets began a sales process with an investment bank to sell its IP and/or an exclusive license to Amazon or another major cloud provider in the hope that it could avoid litigation. Xockets contacted Amazon in connection with this process, given the ubiquitous deployment of Xockets' DPU technologies by cloud giants, including Amazon. Amazon declined to participate in the sales process and has never sought to take a license.

19. Amazon instead chose to employ "an 'infringe now, pay later' strategy,"[3] irreparably harming Xockets and the value of its exclusive intellectual property rights.

20. **<u>Amazon's Annapurna Labs in Austin, Texas</u>.** Amazon infringes Xockets' Asserted Patents, including through its research, design, development, testing, and deployment of infringing server systems at its Annapurna Labs in Austin, Texas. Annapurna is "one of the world's most influential research labs powering modern artificial intelligence," and is "central to Amazon's high-stakes AI strategy."[4] Annapurna Labs, which Amazon acquired for approximately $350 million in 2015,[5] sits at the center of Amazon's ongoing infringement of the Asserted Patents.

21. As explained by the *Wall Street Journal*, "Amazon has long depended on Amazon Web Services—and Amazon Web Services depends on Annapurna."[6] Indeed, "[t]he company's entire AI strategy is now built on a foundation of chips designed by Annapurna, which is so crucial

[3] Kristen J. Osenga, *The Loss of Injunctions Under* eBay*: Evidence of the Negative Impact on the Innovation Economy*, Hudson Institute (Feb. 28, 2024), https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impact-innovation-economy ("In patent law, this is now known as 'predatory infringement', in which defendants choose a commercial strategy of 'infringe now, pay later,' at worst, or, at best, they get away with infringement through a lengthy legal battle of attrition in which the patent owner ultimately just gives up.").

[4] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

[5] https://www.wsj.com/articles/amazon-announces-supercomputer-new-server-powered-by-homegrown-ai-chips-18c196fc.

[6] https://www.wsj.com/tech/amazon-ai-chips-supercomputer-aws-annapurna-trainium-a943be71.

that analysts have described this custom silicon as the secret sauce of AWS."[7] Amazon has advertised that "[a]ll the chips are designed and built by the Annapurna Labs team."[8]

22. According to Bloomberg, Amazon "does all of its custom AI chip design in the U.S., and most of it happens" at Annapurna Labs in Austin. "Today, the team designs and tests custom hardware and software that power AWS data centers worldwide."[9] Annapurna "doesn't just design the chip."[10] It "oversee[s] computer, electrical, mechanical, thermal, and software engineering for the entire server."[11] Annapurna Labs "positions itself to test and trial entire data center-ready server systems before rolling them out for real."[12]

23. **Amazon's Willful Infringement, Damages, and Injunction.** Amazon has brazenly and willfully infringed, and continues to infringe, Xockets' patents. Xockets thus files this patent-infringement complaint, asking the Court to protect Xockets' exclusive patent rights, including with monetary damages for Amazon's years of willful infringement and injunctive relief to stop any further trespass by Amazon on those exclusive property rights.

24. **The Asserted Patents and the Co-Filed New Cloud Processor Case.** This Complaint asserts infringement of four patents: the '297, '092, '161, and '640 Patents. These four patents are generally directed to a new cloud computing fabric with a switching plane architecture that connects together new cloud processors to form a network computing fabric that can operate

[7] https://www.wsj.com/tech/amazon-ai-chips-supercomputer-aws-annapurna-trainium-a943be71.

[8] https://www.aboutamazon.com/news/aws/take-a-look-inside-the-lab-where-aws-makes-custom-chips.

[9] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[10] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[11] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[12] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

independent of the existing cloud network fabric to accelerate distributed computing and enable ML/AI.[13]

25. The ’297 Patent is generally directed toward server systems in cloud data centers with computation modules, or DPUs, that are structured with a novel switching plane architecture that forms a new cloud fabric that can process data-intensive workloads of server processors independent of the existing cloud network fabric to accelerate distributed computing in data centers, including in machine learning and artificial intelligence applications.

26. The ’092 Patent is generally directed toward server systems with the new cloud fabric having a novel distributed computing architecture that enables in-network computing in the fabric for processing data-intensive workloads, including reduction/combining, multicast acceleration, and other MPI operations used in machine learning and artificial intelligence.

27. The ’161 Patent is generally directed towards server systems with a new cloud fabric having offload processor modules, or DPUs, that are structured with a novel switching plane architecture that enables active traffic management and stream processing in the fabric to accelerate distributed computing, including for machine learning and artificial intelligence.

28. The ’640 Patent is generally directed to server systems with the new cloud fabric having offload processor modules, or DPUs, that are structured with a novel switching plane architecture that enables reduction/combining operations in the fabric for data-intensive workloads to accelerate distributed computing, including for machine learning and artificial intelligence.

[13] Co-filed with this suit in a separate complaint, Xockets alleges that Amazon infringes three other patents: U.S. Patent Nos. 11,080,209, 10,649,924, and 11,082,350, which are generally directed to a new cloud processor (i.e., a computation module or offload processor module, which can be called a DPU) and its computing architecture that operates independent of server processors (e.g., CPUs, GPUs, and hybrids of these server processors), to accelerate data-intensive workloads such as security, networking, and storage operations by offloading these workloads from server processors to programmable pipelines of hardware accelerators in the new cloud processors. The technologies of those patents are described at further length in that co-pending complaint.

THE PARTIES

29. Xockets is a Texas corporation, with its principal place of business located in the Temple Office Park at 2027 South 61st Street, Suite 107, Temple, Texas 76504.

30. Xockets holds all substantial rights in and to the Asserted Patents. These include the "exclusive right" to use the patented inventions during the limited term of each patent to be rewarded for the innovations, including the right to exclude others from using its inventions, and to assert all causes of action for infringement under the Asserted Patents, and the exclusive right to damages and injunctive relief to stop continued infringement, and any other remedies for infringement of the Asserted Patents.

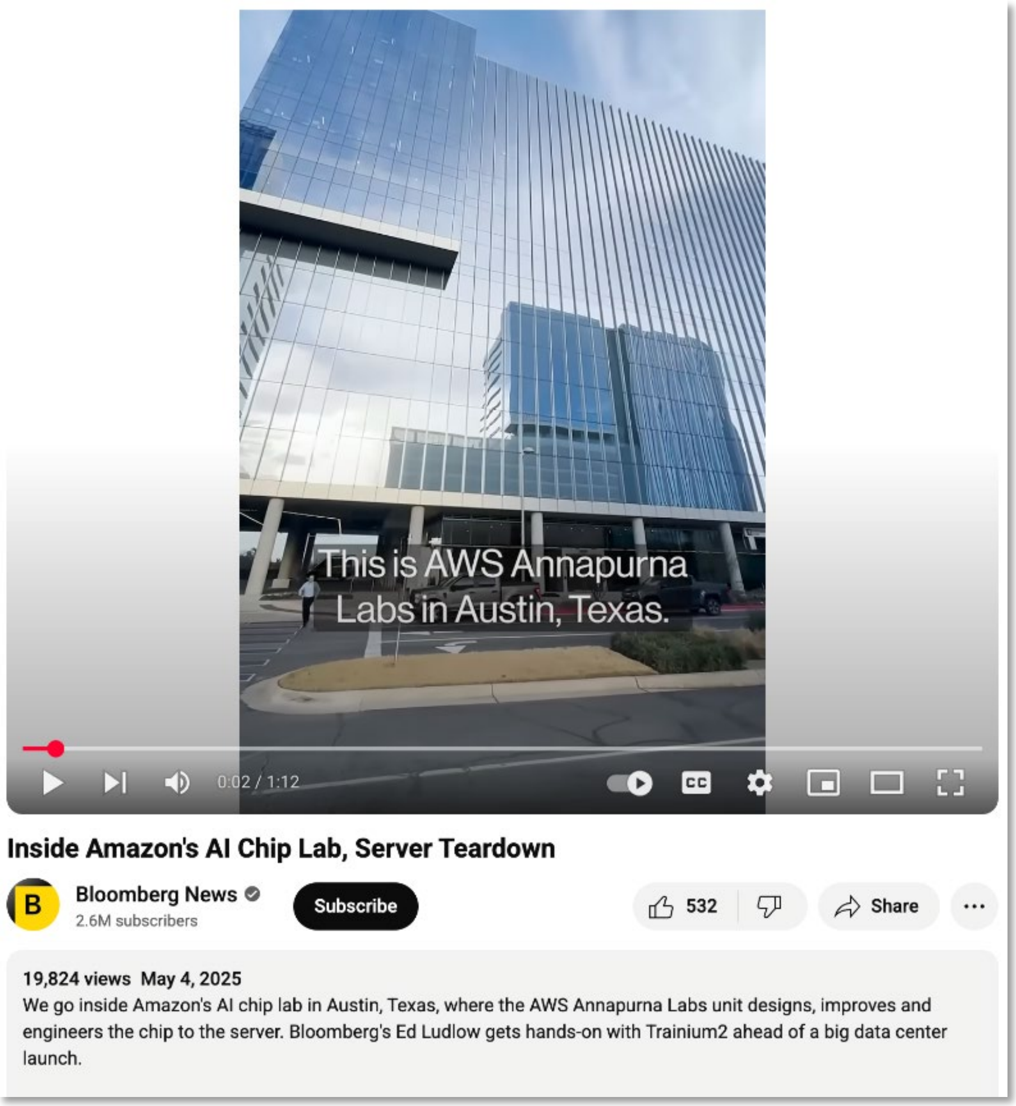
31. On information and belief, Defendant Amazon.com, Inc. is a Delaware corporation with a principal place of business at 410 N. Terry Avenue North, Seattle, Washington 98109. Amazon.com, Inc. can be served through its registered agent for service of process at Corporation Service Company d/b/a CSC – Lawyers Incorporating Service Company, at 211 E. 7th Street, Suite 620, Austin, Texas 78701.

32. On information and belief, Defendant Amazon Web Services, Inc. ("AWS") is a Delaware corporation with a principal place of business at 410 N. Terry Avenue North, Seattle, Washington 98109. Amazon Web Services can be served through its registered agent for service of process at Corporation Service Company d/b/a CSC – Lawyers Incorporating Service Company, at 211 E. 7th Street, Suite 620, Austin, Texas 78701.

33. On information and belief, AWS is a wholly owned subsidiary of Amazon.com, Inc.

34. On information and belief, AWS wholly owns Annapurna Labs (U.S.) Inc.,[14] which, as described in further detail below, has a large research, development, and testing facility in Austin, Texas, which develops, tests, and deploys infringing systems.

[14] *See* https://aws-experience.com/emea/tel-aviv/event/4557f3a5-4a26-4766-be57-35e168bb5165 ("Four years after its inception, Annapurna Labs was acquired by Amazon Web Services (AWS).").



Inside Amazon's AI Chip Lab, Server Teardown

Bloomberg News
2.6M subscribers

19,824 views May 4, 2025
We go inside Amazon's AI chip lab in Austin, Texas, where the AWS Annapurna Labs unit designs, improves and engineers the chip to the server. Bloomberg's Ed Ludlow gets hands-on with Trainium2 ahead of a big data center launch.

Source: https://www.youtube.com/watch?v=cb0y548wcMI

JURISDICTION AND VENUE: DESIGN, DEVELOPMENT, AND TESTING OF INFRINGING SYSTEMS IN AUSTIN, TEXAS

35. This Court has subject matter jurisdiction under 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.*

36. This Court has personal jurisdiction over Amazon because Amazon has committed acts within Texas and this judicial district giving rise to this action and/or has established minimum

contacts with this forum such that the exercise of jurisdiction would not offend traditional notions of fair play and substantial justice.

37. On information and belief, Amazon conducts substantial business in this forum, including (a) engaging in the infringing conduct alleged herein in Texas and in this judicial district; (b) regularly and consistently doing and soliciting business; (c) engaging in other persistent courses of conduct such as providing customer service and warranty repairs in connection with its business operations in Texas and in this judicial district; (d) deriving substantial revenue by its offering of infringing products and services and providing infringing goods to consumers in Texas and in this judicial district; and (e) purposefully establishing substantial, systematic, and continuous contacts with the state of Texas and with this District such that it should reasonably expect to be subject to suit here in this judicial district.

38. In the alternative, Federal Rule of Civil Procedure 4(k)(1)(A) confers personal jurisdiction over Amazon because, upon information and belief, Amazon regularly conducts, transacts, and/or solicits business in Texas and in this judicial district; derives substantial revenue from its business transactions in Texas and in this judicial district; and otherwise avails itself of the privileges and protection of the laws of the State of Texas such that this Court's assertion of jurisdiction over Amazon does not offend traditional notions of fair play and due process. On information and belief, Amazon's unlawful infringing actions have caused and will continue to cause injury to Xockets in Texas and in this judicial district such that Amazon should reasonably expect such actions to have consequences in Texas and in this judicial district.

39. This Court also has personal jurisdiction over Amazon because, directly or through intermediaries, Amazon has committed acts and continues to commit acts of patent infringement in the state of Texas and within this judicial district, including making, using, offering to sell and/or

selling the Accused Instrumentalities in Texas, and/or importing the Accused Instrumentalities into Texas, and/or inducing others to commit acts of patent infringement in Texas. On information and belief, Amazon data centers in this District employ the Accused Instrumentalities, and Amazon has offered to sell to customers in this District services that utilize the Accused Instrumentalities. Furthermore, on information and belief, Amazon has made, used, or tested the Accused Instrumentalities in this District, as further explained below.

40. On information and belief, Amazon has derived substantial revenues from infringing acts in the Western District of Texas, including from the sale and use of infringing products and the Accused Instrumentalities.

41. Venue is proper because, *inter alia*, pursuant to 28 U.S.C. §§ 1391(b) and 1400(b), Amazon has committed acts of infringement in this judicial district and has maintained regular and established places of business in this judicial district, including at least at 11501 Altera Parkway, Austin, Texas 78758.

42. On information and belief, and as set out in further detail below, Amazon has infringed through the conduct of one or more of its subsidiaries, including AWS and Annapurna Labs (U.S.) Inc. AWS is a wholly owned subsidiary of Amazon.com, Inc. On information and belief, Annapurna Labs (U.S.) Inc. is a wholly owned subsidiary of AWS.[15]

43. Ground zero, or home base, for Amazon's infringement of Xockets' patents is Amazon's Annapurna Labs, in Austin, Texas. It is here that Amazon researches, designs, develops, and tests the Accused Instrumentalities, including their features accused in this Complaint.[16]

[15] *See* https://aws-experience.com/emea/tel-aviv/event/4557f3a5-4a26-4766-be57-35e168bb5165 ("Four years after its inception, Annapurna Labs was acquired by Amazon Web Services (AWS).").

[16] *See* https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

Source: https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor

44. Annapurna Labs, which Amazon acquired in 2015 for approximately $350 million,[17] is "one of the world's most influential research labs powering modern artificial intelligence," and is "central to Amazon's high-stakes AI strategy."[18]

45. As a startup, Annapurna specifically sought out Austin as its business location over a decade ago, as the "chip giants" in the industry already had offices there.[19]

46. As explained by the *Wall Street Journal*, "Amazon has long depended on Amazon Web Services—and Amazon Web Services depends on Annapurna."[20] Indeed, "[t]he company's entire AI strategy is now built on a foundation of chips designed by Annapurna, which is so crucial that analysts have described this custom silicon as the secret sauce of AWS."[21] Amazon has advertised that "[a]ll the chips are designed and built by the Annapurna Labs team."[22]

47. According to Bloomberg, Amazon "does all of its custom AI chip design in the U.S., and most of it happens" at Annapurna Labs in Austin. "Today, the team designs and tests custom hardware and software that power AWS data centers worldwide."[23]

[17] https://www.wsj.com/articles/amazon-announces-supercomputer-new-server-powered-by-homegrown-ai-chips-18c196fc.

[18] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

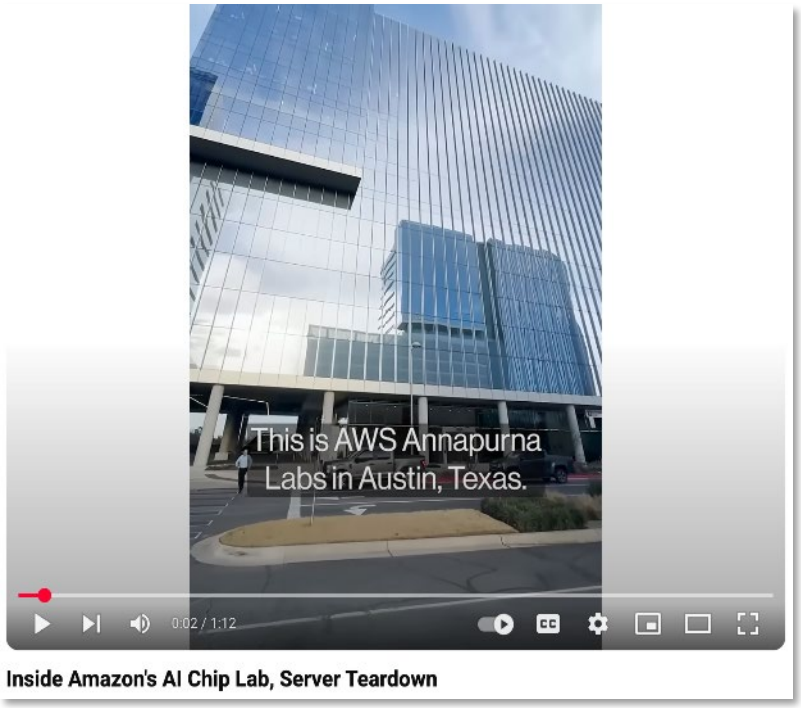
[19] https://www.wsj.com/articles/amazon-announces-supercomputer-new-server-powered-by-homegrown-ai-chips-18c196fc

[20] https://www.wsj.com/tech/amazon-ai-chips-supercomputer-aws-annapurna-trainium-a943be71.

[21] https://www.wsj.com/tech/amazon-ai-chips-supercomputer-aws-annapurna-trainium-a943be71.

[22] https://web.archive.org/web/20250510185341/https:/www.aboutamazon.com/news/aws/take-a-look-inside-the-lab-where-aws-makes-custom-chips.

[23] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.



This is AWS Annapurna Labs in Austin, Texas.

Inside Amazon's AI Chip Lab, Server Teardown

Source: https://www.youtube.com/watch?v=cb0y548wcMI

48. Amazon's Austin lab is where Amazon "integrates, tests and prototypes the hardware that the chips are integrated with, as well as the motherboards and racks that the custom silicon interacts with."[24]

49. Annapurna Labs is where Amazon has developed and continues developing its flagship Trainium chip,[25] which "helps reduce Amazon's reliance on AI chip leader NVIDIA," and "give[s] AWS more control of its own destiny in the market that NVIDIA dominates."[26] Amazon began working on Trainium at Annapurna as early as 2020, and is now developing improved versions, including Trainium3 and Trainium3-based servers.[27] Annapurna "doesn't just design the [Trainium] chip."[28] It "oversee[s] computer, electrical, mechanical, thermal, and software engineering for the entire server."[29] As set forth below, Trainium is a key part of the Accused Instrumentalities.

[24] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

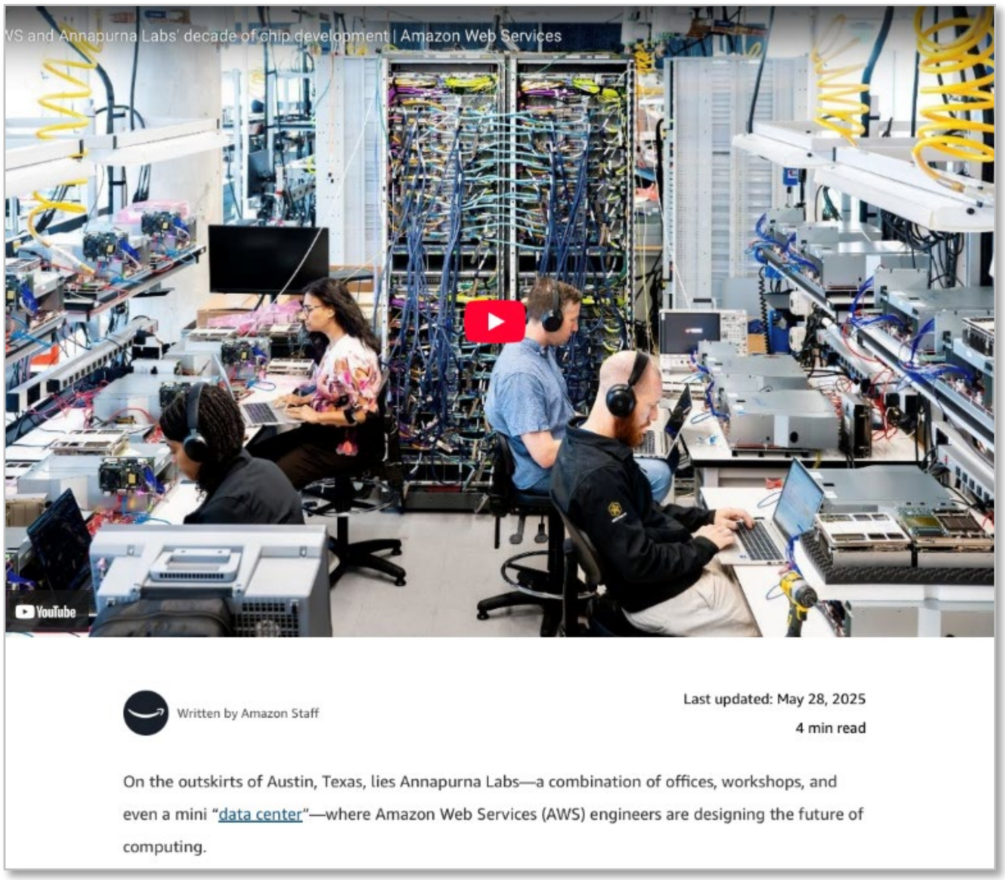
[25] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation; https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[26] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[27] https://www.wsj.com/articles/amazon-announces-supercomputer-new-server-powered-by-homegrown-ai-chips-18c196fc.

[28] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.

[29] https://www.instagram.com/bloombergtv/reel/DJFo0xaJ51h.


VS and Annapurna Labs' decade of chip development | Amazon Web Services
YouTube

Written by Amazon Staff

Last updated: May 28, 2025

4 min read

On the outskirts of Austin, Texas, lies Annapurna Labs—a combination of offices, workshops, and even a mini "data center"—where Amazon Web Services (AWS) engineers are designing the future of computing.

Source: https://www.aboutamazon.com/news/aws/take-a-look-inside-the-lab-where-aws-makes-custom-chips

50. At Annapurna Labs in Austin, Amazon is also developing an "Ultracluster," which is a "massive AI supercomputer made up of hundreds of thousands of its homegrown Trainium chips."[30] As set forth below, the Ultracluster is also a key part of the Accused Instrumentalities.

[30] https://www.wsj.com/articles/amazon-announces-supercomputer-new-server-powered-by-homegrown-ai-chips-18c196fc.

Source: https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation

51. Annapurna Labs "positions itself to test and trial entire data center-ready server systems before rolling them out for real."[31] This allows Amazon personnel at the Austin lab to "understand how chips will work in the field alongside the actual equipment it will run on and provide diagnostics, testing and opportunities for further refinement."[32] Annapurna Labs even includes its own "mini data center, where the team can test and trial new equipment or processes before rolling them out for real."[33] On information and belief, this means that the Accused Instrumentalities have been used and tested by Amazon at Annapurna Labs in this District.

52. Accordingly, Annapurna Labs is a central player in Amazon's infringement of the Xockets' patents. Annapurna employees working in this District will be key witnesses on important facts in this case.

[31] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

[32] https://siliconangle.com/2024/11/27/amazons-secretive-ai-weapon-exclusive-look-inside-aws-annapurna-labs-chip-operation.

[33] https://web.archive.org/web/20250510185341/https:/www.aboutamazon.com/news/aws/take-a-look-inside-the-lab-where-aws-makes-custom-chips.

53. AWS employee profiles on LinkedIn show that employees working on the AWS Nitro DPU—including technical program managers, staff engineers, and cloud solutions personnel—are located in and around Austin, Texas.[34] As explained below, Amazon's server systems with AWS Nitro DPU infringe the Asserted Patents.

Source: https://www.aboutamazon.com/news/aws/take-a-look-inside-the-lab-where-aws-makes-custom-chips

54. Current and recent AWS job postings confirm that infringing conduct takes place in Austin. For example, AWS has recently advertised Austin-based job openings for a System Development Engineer in Infrastructure Security,[35] a Hardware Engineer in ML Acceleration,[36] a Hardware Development Engineer in Storage,[37] and a Senior Hardware Development Engineer, working on AWS EC2,[38] which infringes the Asserted Patents.

[34] *E.g.*, https://www.linkedin.com/in/karimallah; https://www.linkedin.com/in/vinayakvashishtha; https://www.linkedin.com/in/tisha-arora-2a90b115a.

[35] https://web.archive.org/web/20250603045501/https:/www.amazon.jobs/en/jobs/2965322/sys-dev-engineer-infrastructure-annapurna-labs; https://www.amazon.jobs/en/jobs/2975562/system-development-engineer-ii-infrastructure-security-annapurna-labs.

[36] https://www.amazon.jobs/en/jobs/2919834/hardware-engineer-ml-acceleration-annapurna-labs.

[37] https://www.linkedin.com/jobs/view/hardware-development-engineer-storage-aws-hardware-services-at-amazon-web-services-aws-4082074937.

[38] https://www.amazon.jobs/en/jobs/3012000/senior-hardware-development-engineer-hardware-engineering-services.

55. The Accused Instrumentalities also implement NVIDIA's NVLink switch system, which includes NVLink Switch DPUs, which, on information and belief, were developed in Austin by Mellanox Technologies, Ltd.

56. Mellanox was engaged in the business of designing and developing DPUs, switches, and server systems, and it did so in Austin. Those products and activities are central to Amazon's infringement of Xockets' patents, as explained below.

57. On information and belief, NVIDIA acquired Mellanox and its DPU developments in April 2020. This included the acquisition of a major Mellanox facility in Austin.

58. On information and belief, many Mellanox employees present before the acquisition have continued to work for NVIDIA in Austin. Many of those employees continue to work on designing and developing DPUs, including, on information and belief, DPU features and functionality that are part of the Accused Instrumentalities in this case.

59. Accordingly, NVIDIA and Mellanox are central players in Amazon's infringement of the Xockets' patents. NVIDIA and Mellanox employees working in this District will be key witnesses to important facts in this case.

60. Furthermore, on information and belief, Amazon has also deployed the Accused Instrumentalities at AWS data centers within this judicial district, including in the Austin and San Antonio areas. For example, Amazon recently broke ground on a new site for a data center in Round Rock, Texas, just north of Austin.[39]

61. Amazon also maintains regular and established offices in the Western District of Texas, including but not limited to 11501 Altera Parkway, Austin, Texas 78758. On information

[39] https://communityimpact.com/austin/round-rock/development/2025/04/01/mass-grading-begins-on-amazon-site-in-round-rock.

and belief, Amazon employs more than 10,000 people in Austin alone, and presumably many more throughout this District.

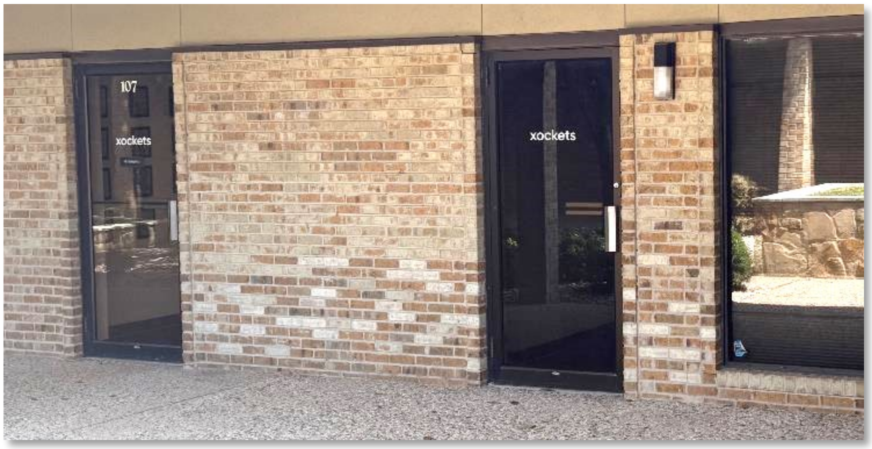
62. Amazon has not disputed this Court's personal jurisdiction over it in other recent patent infringement actions against it. *See, e.g.*, Amazon's Amended Answer ¶ 7, *AlmondNet, Inc. v. Amazon.com, Inc.*, Case No. 21-cv-898 (W.D. Tex. Mar. 10, 2023), ECF No. 93; Amazon's Amended Answer ¶ 16, *Broadband iTV, Inc. v. Amazon.com, Inc.*, Case No. 20-cv-921 (W.D. Tex. Sept. 1, 2021), ECF No. 56.

63. Similarly, Amazon has not contested that venue properly lies in the Western District of Texas in other recent patent infringement actions against it. *See, e.g.*, Amazon's Amended Answer ¶ 8, *AlmondNet, Inc. v. Amazon.com, Inc.*, Case No. 21-cv-898 (W.D. Tex. Mar. 10, 2023), ECF No. 93; Amazon's Amended Answer ¶¶ 12–14, *Broadband iTV, Inc. v. Amazon.com, Inc.*, Case No. 20-cv-921 (W.D. Tex. Sept. 1, 2021), ECF No. 56. In multiple recent patent infringement cases, Amazon has moved to transfer the case to the Austin Division of this District. *See, e.g.*, Sealed Motion for Intra-District Venue Transfer to the Austin Division of the Western District of Texas, *AlmondNet, Inc. v. Amazon.com, Inc.*, Case No. 21-cv-898 (W.D. Tex. June 23, 2022), ECF No. 36; Joint Motion to Transfer Venue to the Austin Division at 1–2, *LS Cloud Storage Techs., LLC v. Amazon.com, Inc.*, Case No. 22-cv-316 (W.D. Tex. Oct. 24, 2022), ECF No. 29.

64. Venue is also proper in this District because this District is the home and principal place of business of Xockets.

65. Xockets' General Counsel Ms. Erin Hudson and Chief IP Counsel Mr. Pierre Hubert reside and work in Austin and the surrounding area.

66. Xockets maintains an office in Temple, Texas, where its patents are displayed on its wall of innovation,[40] and has plans to expand to Austin. From its Temple office, Xockets is researching and building partnerships with researchers in ML/AI, who are working on healthcare solutions to improve peoples' lives.

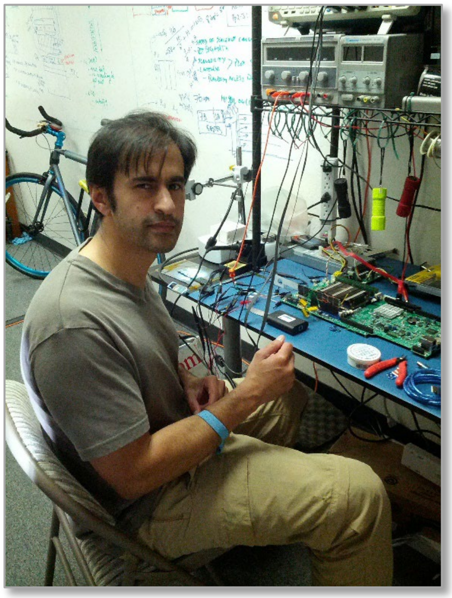


67. Accordingly, this District is the proper forum in which to resolve this dispute. There is a strong local interest in resolving this matter here, insofar as it will significantly impact the intellectual property rights of one of this District's residents.

[40] *See* https://www.xockets.com/our-technology.

XOCKETS AND THE ASSERTED PATENTS

68. Xockets was founded in 2012 by Dr. Parin Dalal and a team of network infrastructure engineers, turned early cloud engineers. Dr. Dalal earned his Ph.D. in theoretical physics and began his career as an engineer designing CPUs and GPUs for computing platforms. Today, Dr. Dalal is a Board Member of Xockets and is a technical advisor to innovative AI startups. Dr. Dalal worked for Google as a Principal Engineer of Machine Learning and Artificial Intelligence. Before joining Google, Dr. Dalal led company-wide strategic AI decision-making at Varian Medical Systems, now a Siemens company, as its Vice President of Advanced AI, developing AI-based formulations of cancer treatments to save lives.



69. In the early 2010s, Dr. Dalal foresaw that the cloud industry's conventional reliance on increased transistor density for ever-faster computing performance—known as Moore's Law—would eventually fail to meet the challenges that data-intensive workloads pose to distributed computing in data centers. Dr. Dalal knew that Moore's Law would reach its limit, as the data workloads driving distributed computing in clouds would grow by orders of magnitude over the coming years. Cloud data centers, Dr. Dalal realized, needed an entirely new computing paradigm.

70. Dr. Dalal and the Xockets team met that challenge, developing a novel computing architecture that would extend into the network of cloud data centers and process data-intensive workloads. Xockets designed and developed a brand new cloud processor that is known today as a Data Processing Unit ("DPU"). His advanced DPU is designed to provide flexible, hardware handling of computing operations at the speed of the network—or line rate—with software-like programmability that relies on the data in packet flows to define programmable pipelines of

hardware accelerators for the computing operations required for processing different packet flows.

71. This programmable hardware acceleration, embedded in the network, can run new, varied, and evolving cloud infrastructure services. It provides the versatility that cloud computing requires to offload infrastructure services and accelerate a wide array of data-intensive workloads and processes. This versatility is essential to widespread adoption of DPUs in the cloud industry. DPUs permit the processing of data-intensive workloads independent of server processors and conventional computing architectures. This offloading frees up server processors so that they can run their primary workloads and applications for customers at higher speeds and lower costs.

72. Dr. Dalal went yet further. He knew that the conventional cloud fabric would continue to constrain what distributed computing could achieve. In a traditional cloud network, all traffic travels through a CPU-managed spine to reach its destination processors for various compute tasks. But that spine is prone to blocking and bottlenecking. Even as bandwidth between CPUs has increased dramatically, the challenge remains. No matter how densely a node is packed, or how much processing power is otherwise built into a cloud object, the amount of data that can be transported between those objects continues to limit computational power.

73. So, Dr. Dalal devised a new cloud fabric. This fabric connects DPUs in a novel way to form a new cloud network fabric for brokering collective communication independent of the limitations of existing cloud networks. This new cloud fabric is designed for even faster, lower-cost collective communication among server processors, and for in-network computing operations such as sorting, organizing, and reducing/combining data-intensive workloads in distributed computing.

74. For example, his inventions disclose a second switching plane that ensures that network packets always have a non-blocking path. No longer must data travel to the top of a rack

along the DCN spine. Instead, packets arrives efficiently and reliably at their destinations through use of the second switching plane.

75. This new cloud fabric enables the training of large AI models across GPUs in a matter of weeks or months rather than many years as would otherwise be required. Accordingly, this new fabric has enabled massive gains in ML/AI, high-performance computing, and general analytics—all of which now take place in the cloud.

76. Xockets disclosed Dr. Dalal's inventions in a series of patent applications filed with the United States Patent and Trademark Office beginning in May 2012. Xockets currently owns nineteen patents, including the Asserted Patents, with numerous applications pending or prepared for filing—all directed to numerous DPU inventions.

77. Xockets' patented inventions include a groundbreaking new cloud computing architecture that dramatically increases the speed and lowers the power and other operating costs in delivering cloud computing servers and services. Xockets' patented DPU architecture offloads from CPUs, GPUs, and/or other host processors in servers and accelerates the data plane and/or control plane of cloud infrastructure services independent of server processors.

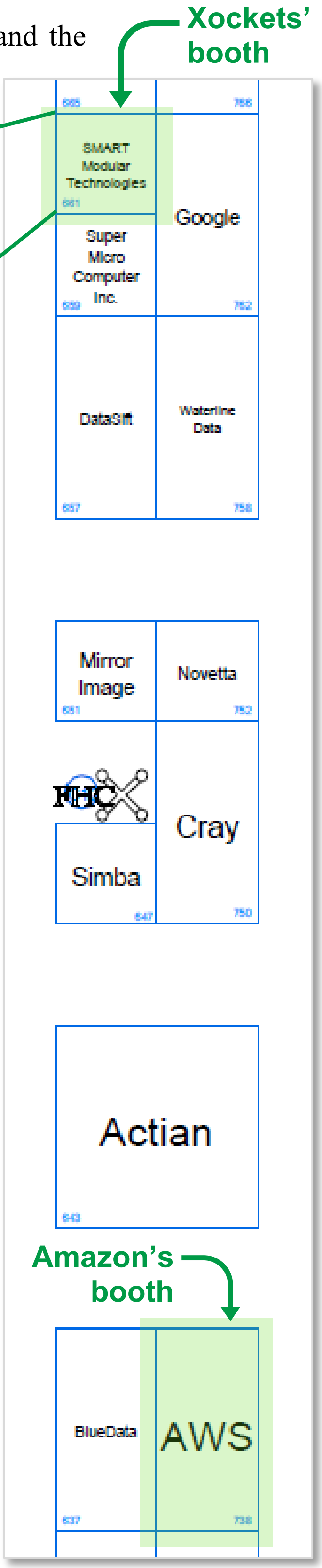
78. Xockets' inventions, for example, disclose a virtual switch computing architecture for offloading from server processors to DPUs, or offload processor modules, accelerating and isolating data-intensive workloads in clouds. Those workloads include security, networking, and storage computing operations that move data between server processors (e.g., CPUs, GPUs, and hybrids of these server processors).

79. Xockets' inventions also disclose a new cloud fabric for distributed computing. A second switching plane allows network packets to travel through non-blocking paths, independent of the main processor, and enables improved distributed computing and ML/AI functionality.

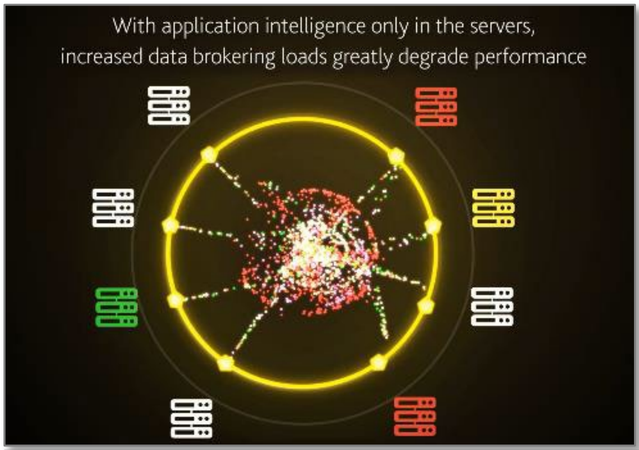
AMAZON LEARNS ABOUT THE PATENTED TECHNOLOGY FROM XOCKETS

80. **2015: Strata Conference**. In the fall of 2015, Dr. Dalal and the Xockets team presented their DPU products at the Javits Convention Center in New York City, at the Strata + Hadoop World Conference, the computing industry's premier big data and network technology conference. At Strata Dr. Dalal and the Xockets team presented the world's very first DPUs, implemented in a product it called the "StreamSwitch."

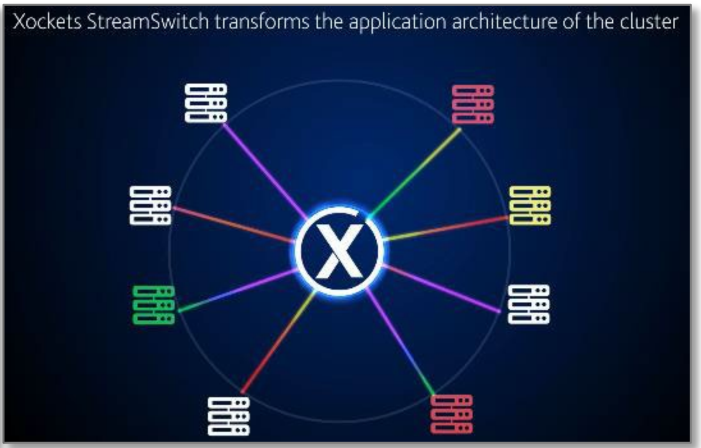
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture
StreamSwitch
Active Application Network
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture



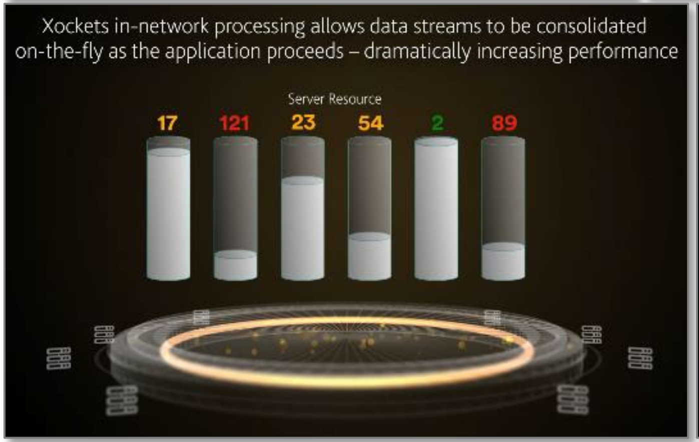
81. It was at Strata that Xockets introduced this new frontier of cloud distributed computing to the industry, demonstrating its value proposition with two demonstration units, one shown at the front of the booth, and another to the left side. Xockets also displayed a video, technical specifications, and benefits across its background display boards. Amazon's AWS booth was situated very close to, only several booths away from, the booth of Xockets and its business partner SMART Modular Technologies, which implemented Xockets' DPU technologies in the manufacture of StreamSwitches.



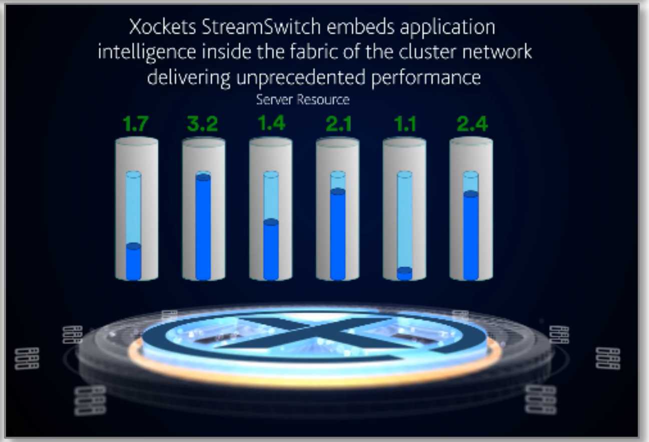
With application intelligence only in the servers, increased data brokering loads greatly degrade performance



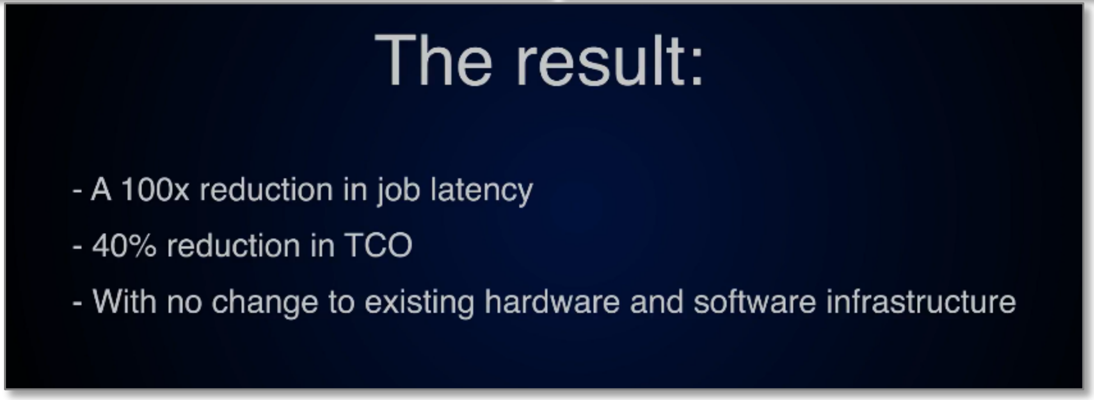
Xockets StreamSwitch transforms the application architecture of the cluster



Xockets in-network processing allows data streams to be consolidated on-the-fly as the application proceeds – dramatically increasing performance



Xockets StreamSwitch embeds application intelligence inside the fabric of the cluster network delivering unprecedented performance



The result:

- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure

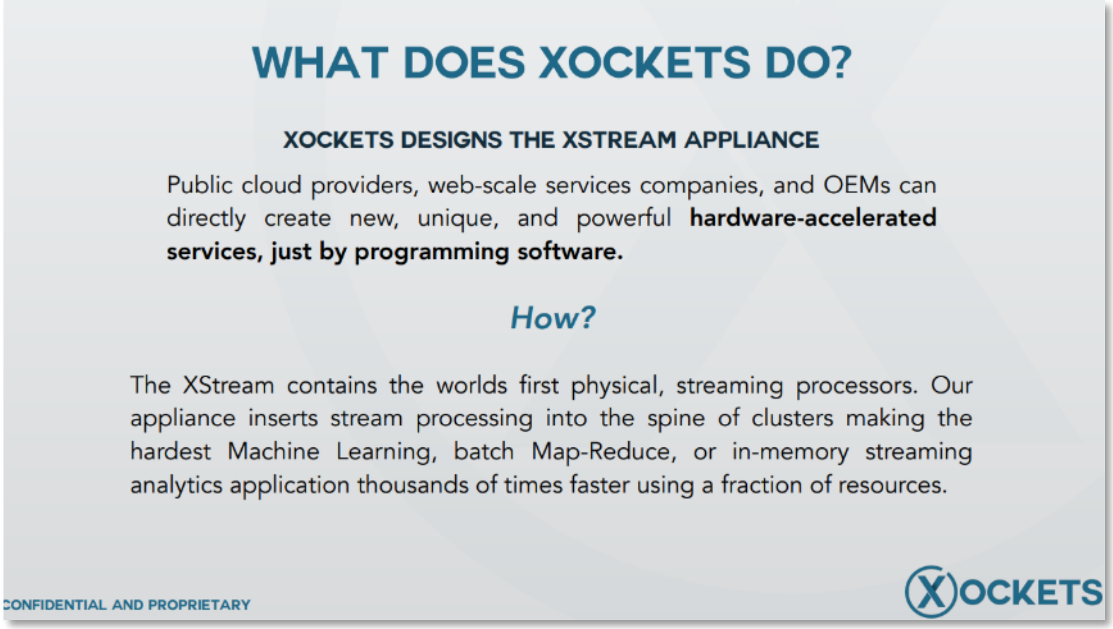
82. **<u>2017: "Deep Dive" Meeting with Amazon</u>**. Big industry players—including Microsoft, NVIDIA, and Amazon—quickly took notice. In May 2017, as part of a discussion around a potential acquisition by Amazon, Dr. Dalal and the entire Xockets engineering team met with Amazon. At the time, Amazon AWS was focused on investing in DPUs to accelerate their cloud servers and services for customers.

83. Proceeding under the premise that Amazon might acquire Xockets, and eager—so Xockets thought—to explore a collaborative relationship, Xockets shared its design and development records with Amazon.

84. As part of this meeting—which Amazon called a “Deep Dive”—Amazon conducted extensive diligence into Xockets’ revolutionary DPU technologies. Amazon interviewed each member of the Xockets team to understand the ins and outs of what Dr. Dalal and his colleagues had developed, and how Xockets’ DPU technologies might solve the world’s growing computing challenges.

85. Nafea Bshara, the head of Amazon’s Austin-based Annapurna Labs, participated in the Deep Dive meeting, where lead engineers from AWS, including Mr. Bshara, separately interviewed each of Xockets’ engineers in separate meeting rooms.

86. In its meeting with Amazon, Xockets presented its DPU technologies, answered questions from AWS leaders, and demonstrated that its DPU technologies had massive acceleration power and cost savings for servers in data centers, including the ability to accelerate Hadoop Big Data analytics workloads by offloading them to Xockets’ StreamSwitch product. This offload showed up to 1000x acceleration.



WHAT DOES XOCKETS DO?

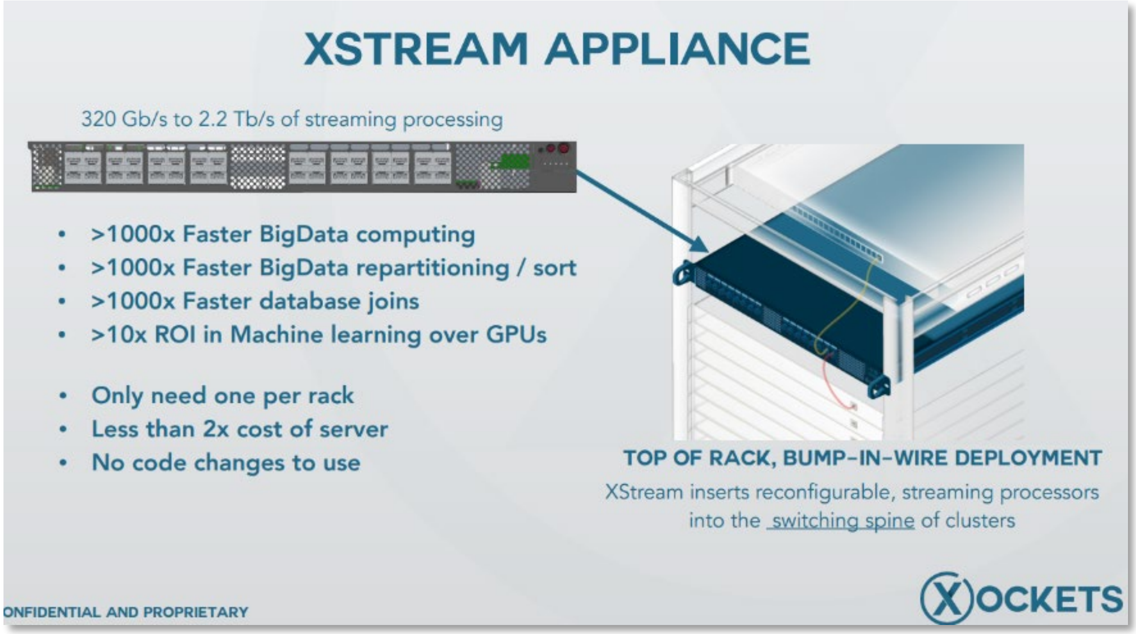
XOCKETS DESIGNS THE XSTREAM APPLIANCE

Public cloud providers, web-scale services companies, and OEMs can directly create new, unique, and powerful **hardware-accelerated services, just by programming software.**

How?

The XStream contains the worlds first physical, streaming processors. Our appliance inserts stream processing into the spine of clusters making the hardest Machine Learning, batch Map-Reduce, or in-memory streaming analytics application thousands of times faster using a fraction of resources.

CONFIDENTIAL AND PROPRIETARY



XSTREAM APPLIANCE

320 Gb/s to 2.2 Tb/s of streaming processing

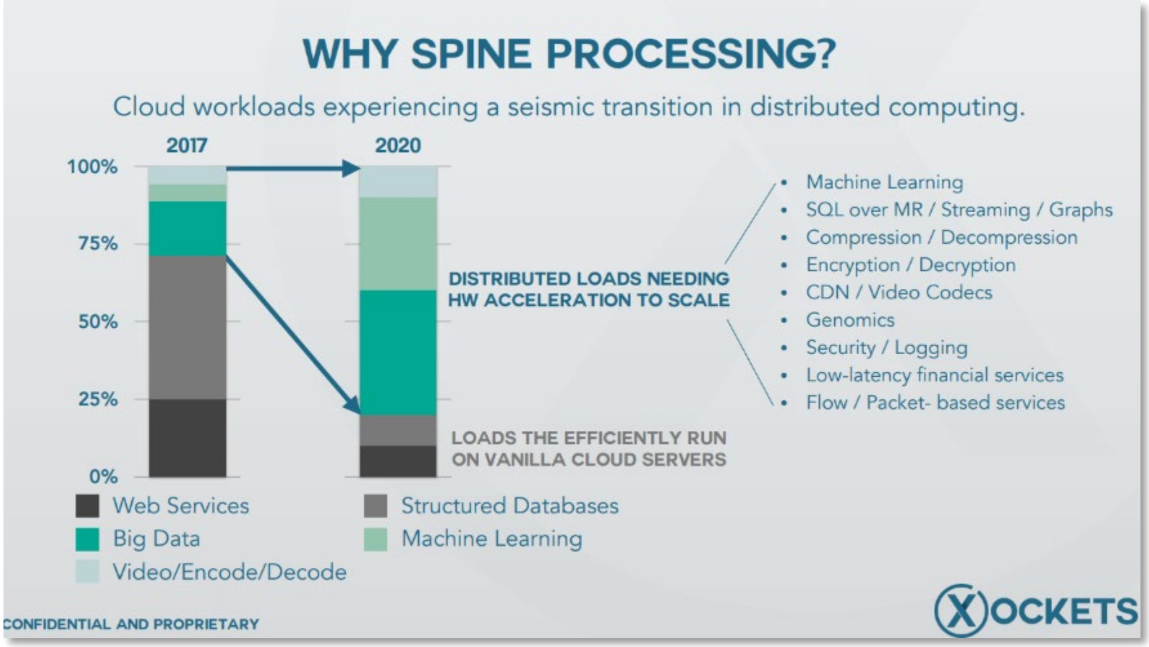
- >1000x Faster BigData computing
- >1000x Faster BigData repartitioning / sort
- >1000x Faster database joins
- >10x ROI in Machine learning over GPUs

- Only need one per rack
- Less than 2x cost of server
- No code changes to use

TOP OF RACK, BUMP-IN-WIRE DEPLOYMENT

XStream inserts reconfigurable, streaming processors into the switching spine of clusters

ONFIDENTIAL AND PROPRIETARY



WHY SPINE PROCESSING?

Cloud workloads experiencing a seismic transition in distributed computing.

- Machine Learning
- SQL over MR / Streaming / Graphs
- Compression / Decompression
- Encryption / Decryption
- CDN / Video Codecs
- Genomics
- Security / Logging
- Low-latency financial services
- Flow / Packet- based services

CONFIDENTIAL AND PROPRIETARY

XOCKETS

87. Amazon also asked Xockets whether and how its DPU technologies could have a genomics application in healthcare, and Xockets obliged, demonstrating in detail that "[t]he major revenue markets forecasted for genomics are: clinical tests for cancer patients and newborns, and research applications for pathogen identification and drug discovery."

88. In connection with the Deep Dive, Xockets also shared extensive financial information, including an NOL statement, fundraising data, and tax credit data.

89. At no time during Xockets' meetings with Amazon did Amazon ever suggest that Xockets' technologies were not novel, were obvious, or were otherwise unworthy of intellectual property protections.

90. Xockets left the Deep Dive meeting hopeful that its decision to share its groundbreaking DPU technologies with Amazon would lead to a fruitful and mutually beneficial partnership with one of the titans of the industry. Xockets and Amazon could together answer the world's growing need for improved cloud infrastructure and services using Xockets' inventions.

91. But Amazon had other plans. It neither acquired Xockets nor sought a license to any of Xockets' patented DPU technologies. Instead, Amazon took Xockets' DPU technologies for itself. Within a year of the Deep Dive meeting, Amazon rolled out a new Nitro product—Nitro v3—which is a Nitro DPU that brazenly deploys Xockets' DPU virtual switch computing architecture on Amazon's own data centers.[41]

92. In the years since Amazon first learned of Xockets' DPU technologies, its flagrant infringement of Xockets' patents has continued—and indeed, *expanded*, especially in today's age of machine learning and AI. The Nitro DPU has now moved past v3, past v4, and into v5. Amazon's cloud-computing infrastructure grows by the day, and it remains built upon Dr. Dalal's inventions.

93. **2024: Amazon's Refusal to Engage in Xockets' Sales Process**. As discussed in further detail below, in 2024, before Xockets enforced its patent rights in litigation, Xockets invited Amazon to participate in a sales process, in view of Amazon's use of Xockets' patented technologies. Amazon declined to participate and continued to infringe Xockets' patents.

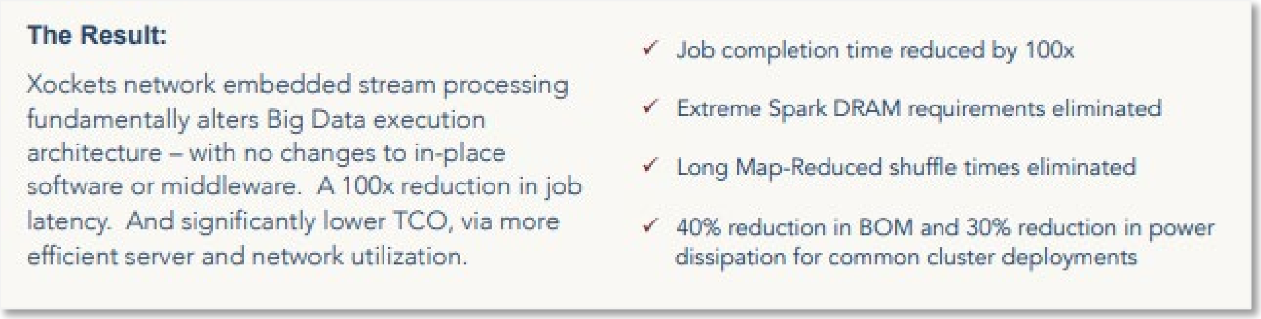
AMAZON HAS PROFITED SUBSTANTIALLY FROM ITS INFRINGEMENT OF XOCKETS' PATENTS.

94. Amazon's infringement has paid dividends, yielding major benefits to its industry standing, competitive position, and bottom line—just as Amazon knew it would.

95. At the Strata industry trade show, Xockets advertised to the industry, including Amazon, the substantial economic advantages promised by its new DPU technologies. For example, Xockets explained that its new cloud architecture resulted in "significantly lower TCO [total cost of ownership], via more efficient server and network utilization," including because its inventions (i) reduced job completion time by 100x, (ii) eliminated extreme Spark DRAM

[41] In this complaint, all references to "Nitro" refer to AWS Nitro v3 and later generations.

requirements, (iii) eliminated long Map-Reduced shuffle times, and (iv) reduced BOM by 40% and power dissipation by 30% for common cluster server deployments in data centers.



The Result:

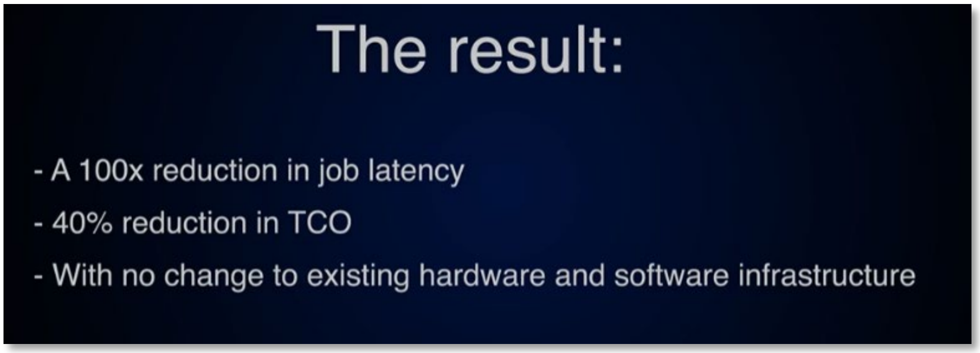
Xockets network embedded stream processing fundamentally alters Big Data execution architecture – with no changes to in-place software or middleware. A 100x reduction in job latency. And significantly lower TCO, via more efficient server and network utilization.

✓ Job completion time reduced by 100x

✓ Extreme Spark DRAM requirements eliminated

✓ Long Map-Reduced shuffle times eliminated

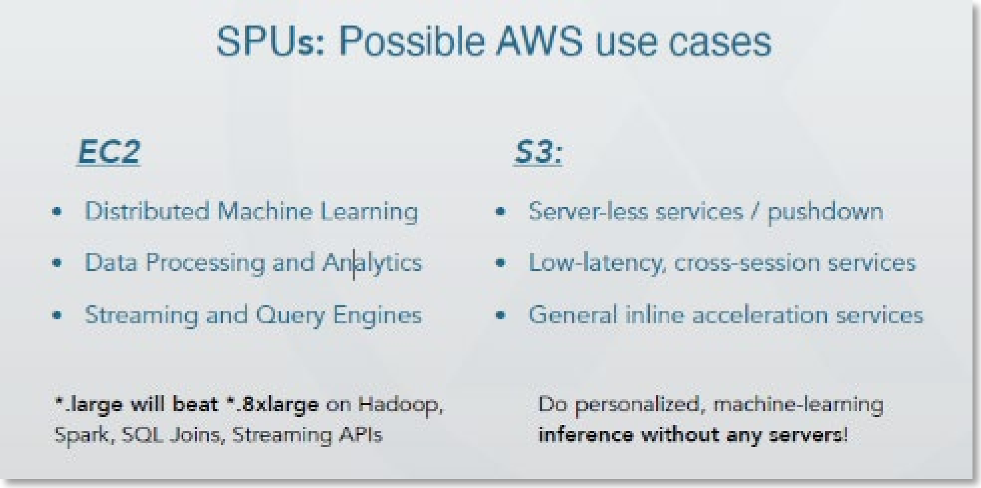
✓ 40% reduction in BOM and 30% reduction in power dissipation for common cluster deployments



The result:

- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure

96. At the 2017 Deep Dive, Dr. Dalal and the Xockets team demonstrated to Amazon that its DPU technologies would be extremely valuable for Amazon, including, by this time, 1000x acceleration. Xockets showed that its DPU technologies offered "AWS use cases" for both Amazon's Elastic Compute Cloud ("EC2") and its cloud object storage system ("S3").



SPUs: Possible AWS use cases

EC2

- Distributed Machine Learning
- Data Processing and Analytics
- Streaming and Query Engines

*.large will beat *.8xlarge on Hadoop, Spark, SQL Joins, Streaming APIs

S3:

- Server-less services / pushdown
- Low-latency, cross-session services
- General inline acceleration services

Do personalized, machine-learning inference without any servers!

97. Xockets showed Amazon that adoption of its technology would result in major cost savings. It explained: "[o]nce bolstered with Xockets technology and patent portfolio, Clouds will be able to use dramatically cheaper, less power-hungry resources to accomplish the same big data and machine learning jobs."

98. Xockets went further, "[e]xplicitly quantifying these savings" through a spreadsheet "informed by public information and conservative assumptions on the workloads used by AWS customers (as representative of the industry) and the bottom line costs to run these jobs."

99. Xockets showed Amazon that its DPU technology would save Amazon approximately 92% in operating expenses and 52% in capital expenditures per year.

100. Despite this massive value proposition, Amazon eventually said no, citing an ostensible "concern as to whether . . . [Xockets'] approach was a technical direction [it] could pursue with AWS."

101. But Amazon quickly went in that exact "technical direction," deploying Xockets' DPU technologies on Nitro v3 by the end of 2017 with a Nitro DPU rollout starting in 2018.

102. Amazon then began touting the massive acceleration and cost-saving benefits offered by Xockets' DPU architecture. In 2020, Amazon's CTO boasted that, "[w]ith the Nitro system, [Amazon's] EC2 performs better across CPU, networking, and storage because [it] moved those functions into dedicated Nitro cards."[42] "Not having to hold back resources for management software means more savings that can be passed on to the consumer."[43]

103. Amazon's infringement, starting shortly after the Deep Dive with Xockets in 2017, has also increased the speed at which Amazon can roll out and scale its cloud technology. Its CTO explained in 2020: "[W]e have launched nearly 4x the number of instances since launching the

[42] https://www.allthingsdistributed.com/2020/09/reinventing-virtualization-with-nitro.html.

[43] https://www.allthingsdistributed.com/2020/09/reinventing-virtualization-with-nitro.html.

Nitro System in 2017. As a result, our customers have a broader tool set to choose from as they optimize price and performance. The faster we innovate, the faster our customers can innovate."[44]

104. Amazon depicted this massive spike in launches as shown below:

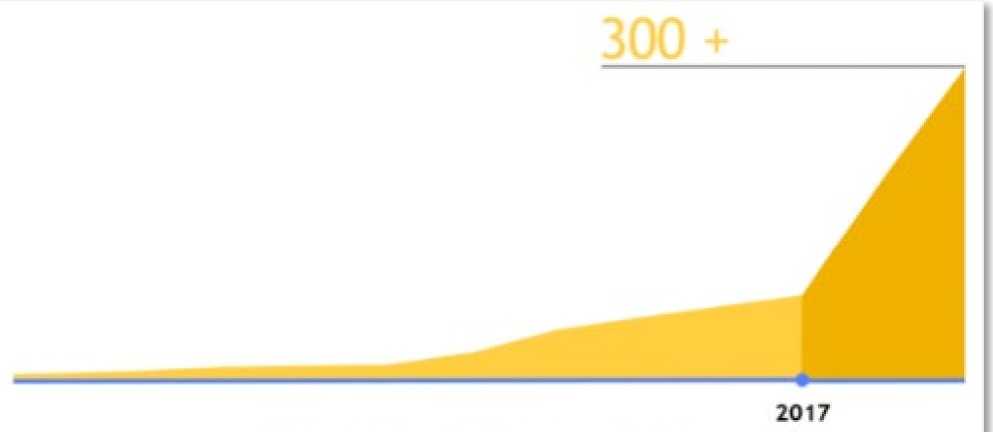


Figure 5: Nitro System enabled innovation

105. In 2018, Amazon SVP James Hamilton explained that "AWS uses network specialized ASICs for all routers and every server includes at least one, and often more."[45] Hamilton explained that "AWS custom ASICs power the Nitro System, handling network virtualization, packet processing, some storage operations, as well as supporting specialized security features."[46] The following year, Hamilton said: "We continue to consume millions of the Nitro ASICs every year so, even though it's only used by AWS, it's actually a fairly high volume server component."[47]

106. In 2021, an industry source echoed that Amazon was enjoying major benefits from these innovations, remarking that "AWS has architected an approach that offloaded the work currently done by the central processor."[48] Amazon had "set the stage for the future allowing

[44] https://www.allthingsdistributed.com/2020/09/reinventing-virtualization-with-nitro.html.
[45] https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor.
[46] https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor.
[47] https://perspectives.mvdirona.com/2019/02/aws-nitro-system.
[48] https://siliconangle.com/2021/06/18/aws-secret-weapon-revolutionizing-computing.

shared memory, memory disaggregation and independent resources that can be configured to support workloads from the cloud to the edge—at much lower cost than can be achieved with general-purpose approaches."[49]

107. According to the same source, AWS Nitro, which is "a set of custom hardware and software that runs on ARM-based chips spawned from Annapurna," "is key to this architecture."[50] "AWS has moved the hypervisor, network and storage virtualization to dedicated hardware that frees up the CPU to run more efficiently."[51] "The reason this is so compelling . . . is that AWS now has the architecture in place to compete at every level of the massive total addressable market, comprising public cloud, on-premises data centers and both the near and far edge."[52] Amazon did so using Xockets' vision and its DPU technologies.

108. Amazon's CTO even quantified the cost-saving benefits, stating in 2020 that, "with this architecture, as much as 30% of the resources in an instance were allocated to the hypervisor and operational management for network, storage, and monitoring."[53]

109. A 2022 study by NVIDIA confirmed that these resource savings have massive financial benefits,[54] including, on information and belief, more than $15,000 per server in power and other operating cost savings over the three-year lifetime of a server.

110. Even accounting for Amazon's lower server cost, power cost, power usage effectiveness, and its higher number of servers, Amazon is enjoying substantial financial savings on a per-server basis given its deployment of the new cloud architecture.

[49] https://siliconangle.com/2021/06/18/aws-secret-weapon-revolutionizing-computing.

[50] https://siliconangle.com/2021/06/18/aws-secret-weapon-revolutionizing-computing.

[51] https://siliconangle.com/2021/06/18/aws-secret-weapon-revolutionizing-computing.

[52] https://siliconangle.com/2021/06/18/aws-secret-weapon-revolutionizing-computing.

[53] https://www.allthingsdistributed.com/2020/09/reinventing-virtualization-with-nitro.html.

[54] https://images.nvidia.com/content/APAC/assets/in/Increasing-Data-Center-Power-Efficiency-with-the-NVIDIA-BlueField-DPU.pdf.

111. On information and belief, accelerated performance of Amazon's cloud servers with its Nitro DPUs could allow public cloud providers like Amazon to realize more than double the cost savings, or at least $30,000 per server in increased revenue over the three-year lifespan of a server.

112. As of November 2023, Amazon had more than 20 million Nitro DPUs in use, each of which delivers these extraordinary financial and business benefits to Amazon.[55]

113. In short, Amazon's mega cloud distributed computing business now benefits from the very advantages that Dr. Dalal and Xockets long ago foresaw and shared.

114. The ML/AI revolution underway today has only added further value to this groundbreaking technology that Amazon took for itself. It is no secret that DPUs' "ability to manage massive data flows, reduce power inefficiencies, and enable scalability will be essential to meeting AI factories' growing demands."[56]

115. Amazon's ML/AI-driven deployment of Dr. Dalal's DPU technologies has caused AWS's revenues and operating income to skyrocket. For example, in 2024, AWS generated $107.6 billion in net sales and $39.8 billion in operating income.[57] These financials are up from 2017 revenue of $17.46 billion and operating income of $4.33 billion[58]—a nearly 10x increase in operating income for AWS since its Deep Dive meeting.

[55] https://www.geekwire.com/2023/inside-the-ai-chip-race-how-a-pivotal-happy-hour-changed-amazons-strategy-in-the-cloud.

[56] https://techstrong.ai/articles/dpus-the-new-heroes-powering-ai-factories.

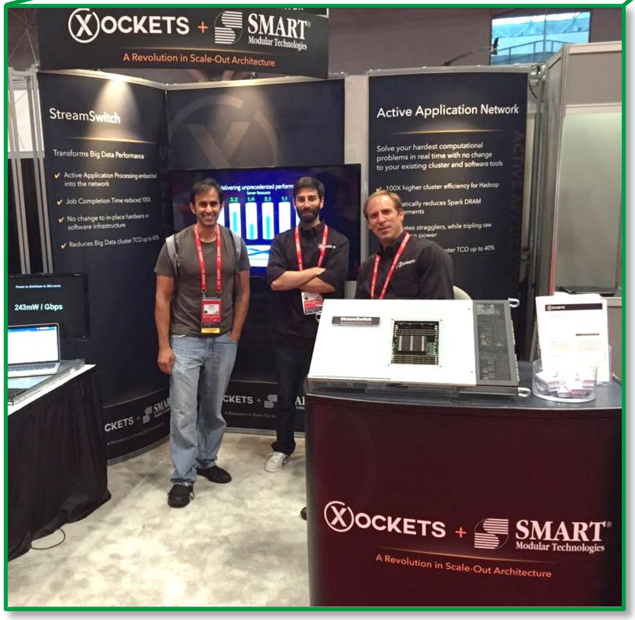
[57] https://s2.q4cdn.com/299287126/files/doc_financials/2024/q4/AMZN-Q4-2024-Earnings-Release.pdf.

[58] https://s2.q4cdn.com/299287126/files/doc_financials/annual/Amazon_AR.PDF.

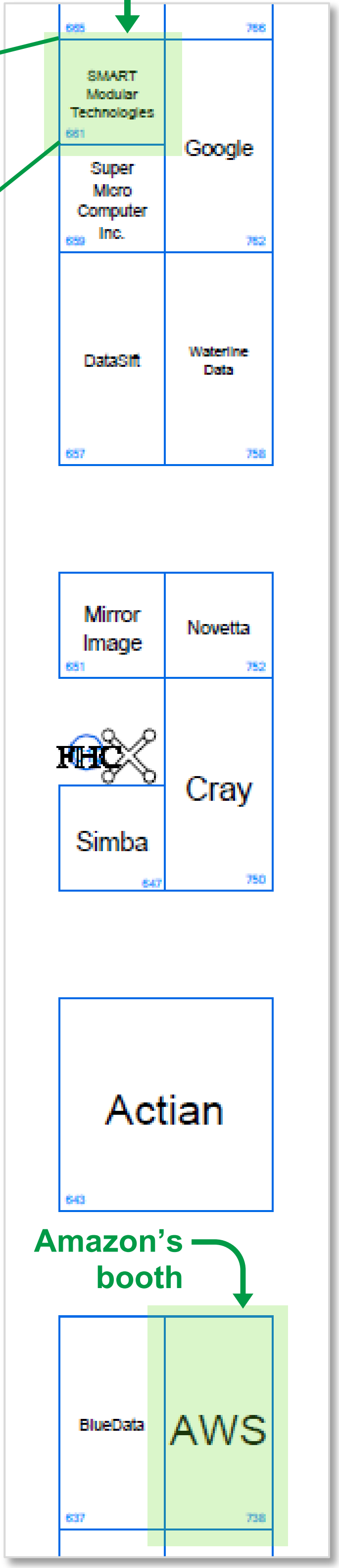
AMAZON'S INFRINGEMENT HAS BEEN WILLFUL

116. Amazon has infringed the Asserted Patents willfully.

117. Amazon first became aware of the essential and novel features of Xockets' DPU technology in September 2015, when Dr. Dalal and the Xockets team presented their groundbreaking DPU inventions at the Strata + Hadoop World Conference in New York City.


XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture
StreamSwitch
Active Application Network
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture

Xockets' booth



Amazon's booth

118. At Strata, Xockets demonstrated its revolutionary new DPU computing architecture and the unprecedented performance benefits it provided by offloading, accelerating, and isolating processing of data-intensive workloads from server processors and conventional computing, and by forming a new cloud network fabric of interconnected DPUs that can operate independent of the performance limitations of existing cloud networks.

119. Xockets' presentation at Strata included demonstrations of its DPUs in its debut product, the StreamSwitch.

120. Representatives of AWS attended Strata in 2015, where they saw and learned about Xockets' groundbreaking DPU technologies. Indeed, AWS's Strata booth was only a few booths down from the booth hosted by Xockets.

121. Amazon's knowledge of Xockets' patents and the willfulness of its infringement of Xockets' patents is further reinforced by Amazon's own patents' repeated citations to Xockets' patents and patent applications.

122. For example, Amazon's U.S. Patent No. 12,164,445 ("Coherent Agents for Memory Access") cites Xockets' U.S. Patent No. 9,665,503, which shares a parent application with the '297 Patent.

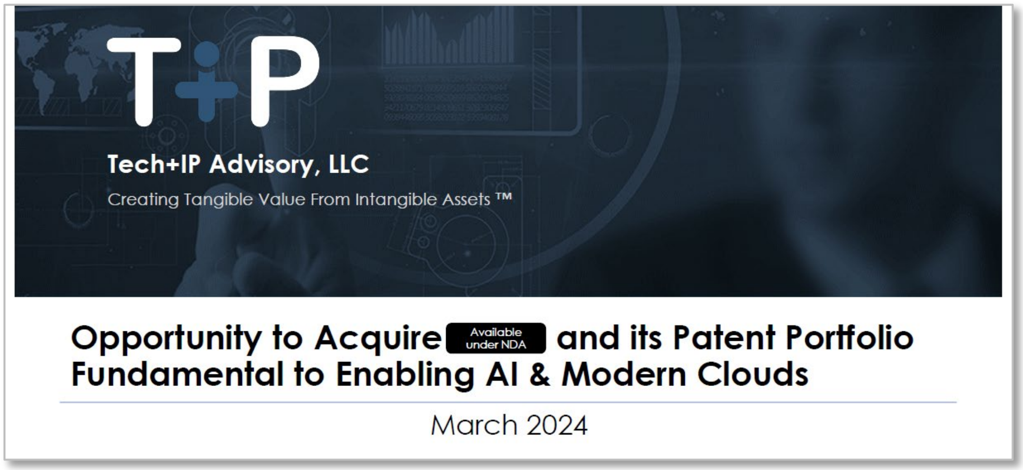
123. Amazon also cites Xockets patent applications that share parent applications with Xockets' '640 Patent in Amazon's U.S. Patent Nos. 9,703,951 ("Allocation of Shared System Resources"), 9,754,103 ("Micro-Architecturally Delayed Timer"), 9,378,363 ("Noise Injected Virtual Timer"), and 9,864,636 ("Allocating Processor Resources Based on a Service-Level Agreement").

124. Xockets put Amazon on clear notice of its patents at least as early as March 2024. From March through June of 2024, Xockets and Amazon engaged in correspondence regarding Xockets' patented DPU technologies, including the Asserted Patents.

125. Specifically, on March 27, 2024, Xockets, acting through an investment bank, contacted Amazon about its "early cloud technology and IP powering the meteoric rise of AI/ML and modern clouds - the Data Processing Unit (DPU)."

126. Xockets informed Amazon that "[t]he original DPU technology is covered with a strong set of patents, dating from 2012 to today," *i.e.*, March 27, 2024.

127. With this correspondence, Xockets sent Amazon a detailed presentation, titled "Opportunity to Acquire [REDACTED[59]] and its Patent Portfolio Fundamental to Enabling AI & Modern Clouds."



Tech+IP Advisory, LLC

Creating Tangible Value From Intangible Assets ™

Opportunity to Acquire Available under NDA and its Patent Portfolio Fundamental to Enabling AI & Modern Clouds

March 2024

128. The presentation described how the Xockets team "invented a series of foundational innovations that laid the groundwork for the widespread deployment of a new processor in clouds known today as the Data Processing Unit (DPU)." The presentation explained to Amazon that Xockets "built a breakthrough new DPU computing architecture and cloud switching fabric to show clouds the path forward."

129. The presentation further explained that Xockets "was granted pioneering patents covering the underlying technologies which have transformed yesterday's cloud to enable today's ML/AI revolution."

130. The presentation notified Amazon that Xockets' "core technologies [were] covered by best-in-class patents recognized across the industry," and that, "[a]s of March 2024, the patent

[59] Xockets' identity was originally withheld pending an NDA negotiation, but was later disclosed to Amazon, as explained below.

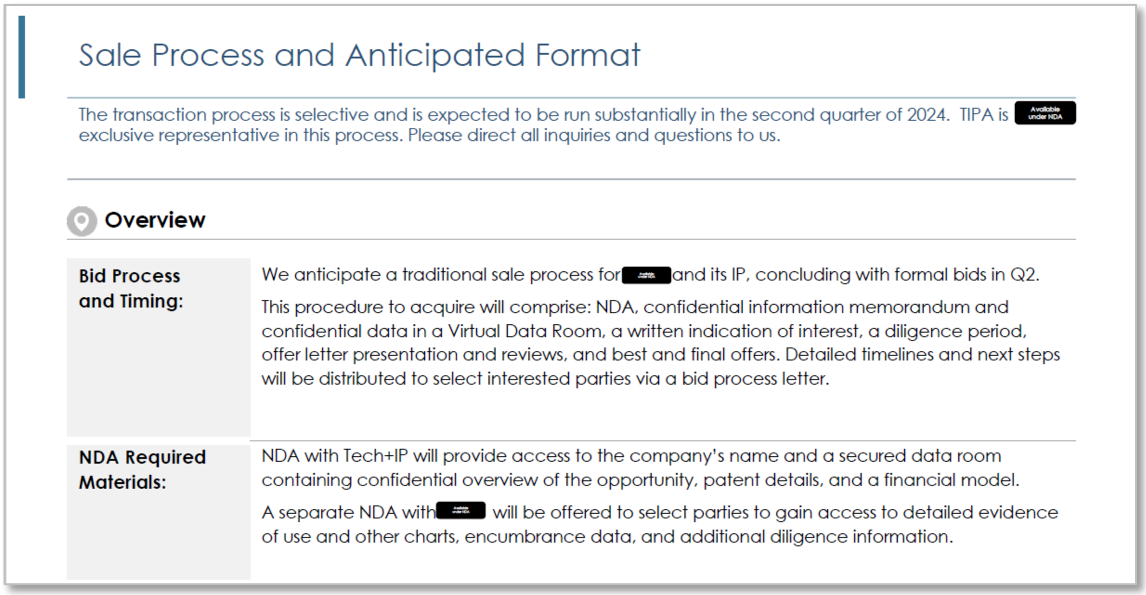
portfolio has >60 world-class claim sets ready, prepared to expand to over 100 assets, covering all key use cases of DPUs."

131. It further explained that "The DPU has Become a Ubiquitous Component Across the Entire Spectrum of Cloud Industry Players," including cloud infrastructure providers like Amazon.

132. The presentation advertised a selective transaction process for the sale of Xockets and its IP. Interested bidders would be required to execute an NDA to access Xockets' name and, in turn, its patent portfolio.

133. Specifically, an NDA would "provide access to the company's name and a secured data room containing confidential overview of the opportunity, patent details, and a financial model." Amazon would have been granted access to this data room had it signed an NDA. The data room contained, among other materials, Xockets' IP information and exemplary evidence-of-use charts.

134. A separate NDA would provide access to a second data room containing "detailed evidence of use and other charts, encumbrance data, and additional diligence information."



Sale Process and Anticipated Format

The transaction process is selective and is expected to be run substantially in the second quarter of 2024. TIPA is Available under NDA exclusive representative in this process. Please direct all inquiries and questions to us.

Overview

Bid Process and Timing:	We anticipate a traditional sale process for Available under NDA and its IP, concluding with formal bids in Q2. This procedure to acquire will comprise: NDA, confidential information memorandum and confidential data in a Virtual Data Room, a written indication of interest, a diligence period, offer letter presentation and reviews, and best and final offers. Detailed timelines and next steps will be distributed to select interested parties via a bid process letter.
NDA Required Materials:	NDA with Tech+IP will provide access to the company's name and a secured data room containing confidential overview of the opportunity, patent details, and a financial model. A separate NDA with Available under NDA will be offered to select parties to gain access to detailed evidence of use and other charts, encumbrance data, and additional diligence information.

135. The presentation's detailed descriptions of Xockets' patented DPU technologies, its widespread, growing use in the cloud computing industry, and statement that entry into NDAs would grant access to "patent details" and "detailed evidence of use and other charts," put Amazon on clear notice that its own cloud products, including its AWS Nitro products, infringe Xockets' patents, including the Asserted Patents.

136. On May 14, 2024, Xockets' agent followed up to offer more information so Amazon could "see if this is a right fit, either for [its] corporate development department or maybe even only the patent acquisition side, which[ever] the company would prefer."

137. That same day, Amazon asked Xockets' agent to "share the patent numbers." In response, Xockets' agent gave Amazon the name of the entity it represented—*i.e.*, Xockets—so that Amazon could obtain the patents and prosecution history of each patent from the public online records of the United States Patent and Trademark Office and review Xockets' patents.

138. On information and belief, Amazon did in fact locate and review Xockets' patents, including the Asserted Patents. For example, Amazon told Xockets on June 11, 2024 that it was "still looking at the materials," which, on information and belief, referred to Xockets' patents, including the Asserted Patents.

139. Ultimately, Amazon declined to acquire Xockets or any of its IP, or to take a license in any of Xockets' patents. Indeed, Amazon declined to participate in the sales process at all.

140. Amazon has continued to infringe Xockets' patents, including the Asserted Patents.

* * *

141. Xockets brings this patent-infringement action to obtain a finding that Amazon has infringed its patents and to recover the significant damages it has incurred as a result of Amazon's

years of willful infringement, and to enjoin Amazon from continued infringement of Xockets' exclusive IP rights.

COUNT ONE
Infringement of the '297 Patent

142. Xockets repeats and incorporates by reference each preceding paragraph as if fully set forth herein and further states:

143. On March 5, 2019, the United States Patent and Trademark Office duly and legally issued the '297 Patent, entitled "Offloading of Computation for Servers Using Switching Plane Formed by Modules Inserted Within Such Servers." A true and correct copy of the '297 Patent is attached as Exhibit 1 to this Complaint.

144. The '297 Patent relates to server systems in cloud data centers, and more particularly to computation modules, or DPUs, also called offload processor modules, in such systems that are connected to form a new switching plane or new cloud fabric that can operate independent of server processors.

145. Xockets holds all substantial rights in and to the '297 Patent, including the exclusive right to assert all causes of action under the '297 Patent and the exclusive right to any remedies for the infringement of the '297 Patent.

146. Amazon is not licensed under the '297 Patent, either expressly or implicitly.

147. Amazon has and continues, without authorization, to operate and use, and/or to induce and contribute to the operation and use by others of equipment and services that practice one or more claims of the '297 Patent literally or under the doctrine of equivalents (hereafter "'297 Accused Instrumentalities"). At a minimum, the '297 Accused Instrumentalities include

(1) Amazon's AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA),

(2) Amazon's AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and

(3) Amazon's AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

148. Amazon has directly infringed and continues to directly infringe, literally and/or under the doctrine of equivalents, at least claims 1 and 7 of the '297 Patent under 35 U.S.C. § 271(a) by operating and using the '297 Accused Instrumentalities in the United States.

149. For example, Amazon infringes at least claim 1 of the '297 Patent. Claim 1 discloses:

> A system, comprising:
>
> a plurality of first server modules interconnected to one another via a communication network, each first server module including
>
> > a first switch,
> >
> > at least one main processor, and
> >
> > at least one computation module coupled to the main processor by a bus, each computation module including
> >
> > > a second switch, and
> > >
> > > a plurality of computation elements; wherein
>
> the second switches of the first server modules form a switching plane for the ingress and egress of network packets independent of any main processors of the first server modules, and
>
> each computation module is insertable into a physical connector of the first server module.

150. Amazon infringes at least claim 1 of the '297 Patent through its AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

151. On information and belief, the Amazon EC2 (Elastic Compute Cloud) instances including Amazon compute instances (i.e., a plurality of server modules) form a system.

152. On information and belief, the Amazon EC2 compute instances are interconnected by AWS Cloud. For example, the Amazon EC2 compute instances, forming a system, are interconnected by Elastic Fabric Adapter (EFA) networking, allowing high levels of inter-node communications between Amazon cloud servers. As another example, the Amazon EC2 compute instances, forming a system, are interconnected by Elastic Network Adapter (ENA) networking, allowing enhanced networking capabilities in Amazon cloud servers.

153. On information and belief, the AWS Nitro System including a Nitro Controller and Nitro Cards (i.e., computation module) for each EC2 compute instance in an Amazon data center operate as a DPU to securely offload infrastructure services from the server processor of the EC2 compute instance and accelerate the data plane infrastructure services of its data centers, including HPC workloads and distributed ML/AI services. Furthermore, Amazon's Elastic Fabric Adapter (EFA) is built on the AWS Nitro DPU, to provide high levels of inter-node communications between Amazon cloud servers. As another example, Amazon's Elastic Network Adapter (ENA) is also built on the AWS Nitro DPU, to provide enhanced networking capabilities in Amazon cloud servers.

154. On information and belief, as another example, the Amazon EC2 compute instances, including Trainium instances (including Trn1, Trn1n, Trn2 and Trn2u instances), form a system that includes the AWS Nitro DPU. The Amazon EC2 Trainium instances enable training and deploying models with hundreds of billions to trillion+ parameters for generative AI. The Amazon EC2 Trainium instances are interconnected by Amazon's EFA networking.

155. On information and belief, the Amazon EC2 compute instances include a first switch, for example, implemented by top of rack (ToR) switches.

156. On information and belief, the Amazon EC2 compute instances include host processors (i.e., at least one main processor) on the system main board.

157. On information and belief, the AWS Nitro DPU is coupled to the host processor (i.e., the main processor) via a PCIe bus (i.e., a bus).

158. On information and belief, the AWS Nitro DPU operates as a virtual switch (i.e., a second switch) to implement software-defined networking, thereby offloading data plane functions of the host processors.

159. On information and belief, the AWS Nitro DPU includes dedicated hardware components with 64-bit processing capabilities and specialized application-specific integrated circuits (ASICs) that function as hardware accelerators (i.e., computation elements). The hardware accelerators of the AWS Nitro DPU include specific-purpose hardware accelerators and general-purpose hardware accelerators (ARM cores).

160. On information and belief, the second switches implemented by the AWS Nitro DPU of each EC2 compute instance facilitate communication by forming a switching plane for the ingress and egress of network packets. As an example, the virtual switches form a switching plane for the ingress and egress of network packets of machine learning/AI flows for distributed machine learning processing across a cluster of GPUs to offload brokering of communication collective operations from server modules. Furthermore, the Elastic Fabric Adapter of the Amazon EC2 compute instances, that is built on the AWS Nitro DPU, facilitates inter-node communications by forming a switching plane for the ingress and egress of network packets for distributed machine learning/AI applications.

161. On information and belief, the AWS Nitro DPU offloads from the host processor (i.e., independent of the main processor) of the EC2 compute instances, including forming a switching plane for ingress and egress of ML/AI and other network packet flows.

162. On information and belief, the AWS Nitro DPU is inserted into the physical PCIe interfaces (i.e., physical connector) on the EC2 compute instance.

163. Amazon thus infringes at least claim 1 of the ’297 Patent through its AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

164. Amazon also infringes at least claim 1 of the ’297 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

165. On information and belief, Amazon EC2 (Elastic Cloud Compute) server platforms, including Amazon EC2 UltraClusters, form a system. Amazon EC2 server system utilizes NVIDIA’s server platforms, including Blackwell-based server platforms with GPUs connected by NVLink Switch DPUs, to accelerate complex ML/AI workloads. As another example, the Amazon EC2 instances, including Amazon EC2 UltraClusters, also utilizes NVIDIA Hopper-based server platforms, including NVIDIA GH200 NVL32, to accelerate training for complex LLMs and generative AI applications.

166. On information and belief, Amazon EC2 server systems include multiple server modules interconnected by a communication network. Amazon EC2 server systems include multiple NVIDIA GB200 compute racks (i.e., first server modules) interconnected via Amazon’s Elastic Fabric Adapter (EFA) networking. Furthermore, Amazon EC2 server systems include

multiple NVIDIA GB200 compute racks (i.e., first server modules) interconnected via Amazon's Elastic Network Adapter (ENA) providing enhanced networking capabilities.

167. On information and belief, each server module in Amazon's EC2 UltraCluster includes a top of rack (ToR) switch to implement a first switch.

168. For example, on information and belief, each Blackwell-based server module includes at least one GB200 Grace Blackwell Superchip (i.e., a main processor). As another example, each Hopper-based server module also includes a Grace Hopper Superchip (i.e., a main processor).

169. On information and belief, each server module includes 5th Generation NVLink and NVLink Switch DPUs (i.e., computation modules) that provide high-speed, seamless communication between every GPU in server clusters. On information and belief, an NVLink Switch DPU is coupled to the GB200 Grace Blackwell Superchip by NVLink Cable Cartridge bus.

170. On information and belief, each NVLink Switch DPU includes Core Logic and Management Logic elements (i.e., a plurality of computation elements), including hardware accelerator engines for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP) for performing in-network computing operations including reductions and multicast acceleration, used for offloading communication collective operations from the main processors.

171. On information and belief, the NVLink Switch DPUs include a second switch that is programmable for switching together multiple SHARP hardware accelerator engines and output ports into software-defined hardware accelerator pipelines to process data packets according to their CUDA context and enable in-network reductions offloaded from server processors.

172. On information and belief, the second switches implemented in the NVLink Switch DPUs of the server modules fully connect all Superchips in the server system and allow passing

of traffic to and from the Superchips by forming a non-blocking compute fabric (i.e., switching plane) for the ingress and egress of network packets across the cluster of Superchips, within a single rack and between racks. The second switches implemented in the NVLink Switch DPUs of the server modules form a non-blocking compute fabric (i.e., switching plane) for the ingress and egress of network packets across the cluster of Superchips, within a single rack and between racks, to offload brokering of communication collective operations from server processors.

173. On information and belief, the switching plane is formed by the NVLink Switch DPUs in an NVLink domain independent of the Grace Blackwell Superchips in the server modules.

174. On information and belief, each NVLink Switch DPU is insertable into a physical connector of the NVIDIA server module.

175. Amazon thus infringes at least claim 1 of the ’297 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

176. Amazon also infringes at least claim 1 of the ’297 Patent through its AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

177. On information and belief, Amazon storage server systems comprise AWS Elastic Block Storage (EBS) servers (i.e., a plurality of servers). The Amazon storage server system includes AWS Nitro System with Nitro Controller and Nitro Cards to enable the offloading of data plane infrastructure functions, including security, networking, storage, and ML/AI services.

178. On information and belief, in an AWS EBS cluster (i.e., a server system), the AWS EBS servers (i.e., a plurality of servers) are interconnected by Elastic Fabric Adapter (EFA) network for high-performance and lower-latency communication.

179. As another example, on information and belief, the Amazon storage server systems comprise AWS storage-optimized servers (i.e., a plurality of server modules), that further include the AWS Nitro DPU to enable the offloading of data plane infrastructure functions, including security, networking, storage, and ML/AI services.

180. On information and belief, the AWS storage-optimized servers are interconnected by AWS Cloud. For example, the AWS storage-optimized servers are interconnected by Elastic Fabric Adapter (EFA) networking, allowing high levels of inter-node communications between the AWS storage-optimized servers.

181. As another example, on information and belief, the AWS storage-optimized servers, are interconnected by Elastic Network Adapter (ENA) networking, allowing enhanced networking capabilities in the AWS storage-optimized servers.

182. On information and belief, the Amazon EC2 compute instances include a first switch, for example, implemented by top of rack (ToR) switches.

183. On information and belief, Amazon storage servers, including AWS EBS servers and AWS storage-optimized servers, further include host processors (i.e., at least one main processor) on the system main board.

184. On information and belief, the AWS Nitro DPU (i.e., computation module) for each storage server in an Amazon data center securely offloads infrastructure services from the server processor of the storage server and accelerates the data plane infrastructure services of its data centers, including distributed ML/AI services.

185. Furthermore, on information and belief, Amazon's Elastic Fabric Adapter (EFA) is built on the AWS Nitro DPU, to provide high levels of inter-node communications between AWS storage servers.

186. As another example, on information and belief, Amazon's Elastic Network Adapter (ENA) is also built on the AWS Nitro DPU, to provide enhanced networking capabilities in AWS storage servers.

187. On information and belief, the AWS Nitro DPU is coupled to the host processor (i.e., the main processor) via a PCIe bus (i.e., a bus).

188. On information and belief, the AWS Nitro DPU operates as a virtual switch (i.e., a second switch) to implement software-defined networking, thereby offloading data plane functions of the host processors.

189. On information and belief, the AWS Nitro DPU includes dedicated hardware components with 64-bit processing capabilities and specialized application-specific integrated circuits (ASICs) that function as hardware accelerators (i.e., computation elements). The hardware accelerators of the AWS Nitro DPU include specific-purpose hardware accelerators and general-purpose hardware accelerators (ARM cores).

190. On information and belief, the second switches implemented by the AWS Nitro DPU of each storage server facilitate communication by forming a switching plane for the ingress and egress of network packets.

191. As an example, on information and belief, the virtual switches form a switching plane for the ingress and egress of network packets of machine learning/AI flows for distributed machine learning processing across a cluster of GPUs to offload brokering of communication collective operations from server modules. Furthermore, the Elastic Fabric Adapter of the AWS storage server, that is built on the AWS Nitro DPU, facilitates inter-node communications by forming a switching plane for the ingress and egress of network packets for distributed machine learning/AI applications.

192. On information and belief, the AWS Nitro DPU offloads from the host processor (i.e., independent of the main processor) of the AWS storage servers, including forming a switching plane for ingress and egress of ML/AI and other network packet flows.

193. On information and belief, the AWS Nitro DPU is inserted into the physical PCIe interfaces (i.e., physical connector) on the AWS storage server.

194. Amazon thus infringes at least claim 1 of the ’297 Patent through its AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System (including AWS EFA and ENA).

195. As a further example, Amazon also infringes at least claim 7 of the ’297 Patent. Claim 7 discloses:

> The system of claim 1, wherein the second switch is a virtual switch comprising computation elements on the computation module.

196. Amazon infringes at least claim 7 of the ’297 Patent through its AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

197. On information and belief, the second switch implemented by the AWS Nitro DPU in each EC2 compute instance is a virtual switch that is programmable (for example, by P4 programming language) for switching together different software-defined pipelines of hardware accelerators (i.e., computation elements) on the AWS Nitro DPU.

198. Amazon also infringes at least claim 7 of the ’297 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

199. On information and belief, the NVLink Switch DPUs include a second switch that is a virtual switch comprising multiple SHARP accelerator engines (i.e., computation elements)

that are switched together with output ports into software-defined hardware accelerator pipelines to process data packets according to their CUDA context and enable in-network reductions offloaded from server processors.

200. Amazon also infringes at least claim 7 of the ’297 Patent through its AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

201. On information and belief, the second switch implemented by the AWS Nitro DPU in each AWS storage server is a virtual switch that is programmable (for example, by P4 programming language) for switching together different software-defined pipelines of hardware accelerators (i.e., computation elements) on the AWS Nitro DPU.

202. Amazon operates and sells within the United States access to its (1) AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), its (2) AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and its (3) AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), thereby directly infringing at least claims 1 and 7 of the ’297 Patent.

203. The infringing aspects of the ’297 Accused Instrumentalities can be used only in a manner that infringes the ’297 Patent and thus have no substantial non-infringing uses. The infringing aspects of those instrumentalities otherwise have no meaningful use, let alone any meaningful non-infringing use.

204. Amazon has had actual or constructive knowledge and notice of the ’297 Patent and its infringement since at least March of 2024, when Xockets communicated with Amazon about the Asserted Patents, and through the filing and service of the Complaint. Despite this knowledge, Amazon continued to commit the infringing acts described herein.

205. Amazon makes and sells hardware and/or software components (e.g., its (1) AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), its (2) AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and its (3) AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA)) especially made or especially adapted to practice the invention claimed in the ’297 Patent, including at least claims 1 and 7, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

206. On information and belief, Amazon’s infringement of the ’297 Patent is willful and deliberate, entitling Xockets to the recovery of enhanced damages under 35 U.S.C. § 284. Amazon has infringed and continues to infringe the ’297 Patent despite the risk of infringement being either known or so obvious that it should have been known to Amazon.

207. Amazon’s acts of infringement have caused and continue to cause damage to Xockets, and Xockets is entitled to recover from Amazon the damages it has sustained as a result

of those wrongful acts in an amount subject to proof at trial, but in no event less than a reasonable royalty for the use made of the invention in the ’297 Patent, together with interest and costs as fixed by the Court.

COUNT TWO
Infringement of the ’092 Patent

208. Xockets repeats and incorporates by reference each preceding paragraph as if fully set forth herein and further states:

209. On February 19, 2019, the United States Patent and Trademark Office duly and legally issued the ’092 Patent, entitled “Architectures and Methods for Processing Data in Parallel Using Offload Processing Modules Insertable into Servers.” A true and correct copy of the ’092 Patent is attached as Exhibit 2 to this Complaint.

210. The ’092 Patent relates to a cloud distributed computing architecture for executing at least first and second computing operations in parallel for processing data-intensive workloads of server processors, including CPUs, GPUs, or hybrids of these server processors.

211. Xockets holds all substantial rights in and to the ’092 Patent, including the exclusive right to assert all causes of action under the ’092 Patent and the exclusive right to any remedies for the infringement of the ’092 Patent.

212. Amazon is not licensed under the ’092 Patent, either expressly or implicitly.

213. Amazon has and continues, without authorization, to operate and use, and/or to induce and contribute to the operation and use by others of equipment and services that practice one or more claims of the ’092 Patent literally or under the doctrine of equivalents (hereafter “’092 Accused Instrumentalities”). At a minimum, the ’092 Accused Instrumentalities include

(1) Amazon’s AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA),

(2) Amazon’s AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA, ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and

(3) Amazon’s AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

214. Amazon has directly infringed and continues to directly infringe, literally and/or under the doctrine of equivalents, at least claim 1 of the ’092 Patent under 35 U.S.C. § 271(a) by operating and using the ’092 Accused Instrumentalities in the United States.

215. For example, Amazon infringes at least claim 1 of the ’092 Patent. Claim 1 discloses:

> A distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data, the architecture comprising:
>
> a plurality of servers, including first servers that each include
>
> > at least one central processing unit (CPU), and
> >
> > at least one offload processing module coupled to the at least one CPU by a bus, each offload processing module including a plurality of computation elements, the computation elements configured to
> >
> > > operate as a virtual switch, and
> > >
> > > execute the second computing operations on first processed data to generate second processed data; wherein
>
> the virtual switches form a switch fabric for exchanging data between the offload processing modules,
>
> the first computing operations generate the first processed data and are not executed by the offload processing modules, and the second computing operations are executed on a plurality of the offload processing modules in parallel.

216. Amazon infringes at least claim 1 of the ’092 Patent through its AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

217. On information and belief, the Amazon EC2 (Elastic Compute Cloud) instances including Amazon EC2 compute instances (i.e., a plurality of servers) form a distributed computing architecture for various applications, including big data and ML/AI workloads.

218. On information and belief, the Amazon EC2 compute instances, including for example, Trainium instances (including Trn1, Trn1n, Trn2 and Trn2u instances), form a distributed computing architecture that include AWS Nitro System with Nitro Controller and Nitro Cards. Trainium instances, deployed in EC2 Ultraclusters, form a distributed computing architecture to enable distributed training.

219. On information and belief, the Amazon EC2 compute instances include host processors (i.e., at least one central processing unit (CPU)) on the system main board.

220. On information and belief, the Amazon EC2 compute instances include AWS Nitro DPU (i.e., offload processor modules) to enable the offloading of data plane infrastructure functions, and accelerate the data plane infrastructure services of its data centers, including big data and distributed ML/AI services. Furthermore, Amazon's Elastic Fabric Adapter is built on the AWS Nitro DPU, to provide high levels of inter-node communications between Amazon cloud servers and accelerate the data plane infrastructure services including ML/AI services. As another example, Amazon's Elastic Network Adapter is also built on the AWS Nitro DPU, to provide enhanced networking capabilities in Amazon cloud servers.

221. On information and belief, the AWS Nitro DPU is coupled to the host processor (i.e., the main processor) via a PCIe bus (i.e., a bus).

222. On information and belief, the AWS Nitro DPU includes dedicated hardware components with 64-bit processing capabilities and specialized application-specific integrated circuits (ASICs) that function as hardware accelerators (i.e., computation elements). The hardware

accelerators of the AWS Nitro DPU include specific-purpose hardware accelerators and general-purpose hardware accelerators (ARM cores).

223. On information and belief, the AWS Nitro DPU implements a virtual switch for software-defined networking that is programmable (for example, by P4 programming language) for switching together different software-defined pipelines of hardware accelerators to address the different needs of network packets flows of different cloud users or services, to offload the data plane functions of the host processors.

224. On information and belief, the AWS Nitro DPU in each Amazon EC2 compute instance facilitates communication by forming a switching fabric for exchanging data between them.

225. For example, on information and belief, the AWS Nitro DPU in Amazon EC2 compute instances, interconnected by Elastic Fabric Adapter (EFA) networking facilitates communication by forming a switching fabric for exchanging data.

226. On information and belief, the Amazon EC2 compute instances are supported by the AWS Nitro DPU in running different customer applications, including code and data processing operations. The Amazon EC2 compute instances are configured to generate requests to the AWS Nitro DPU (e.g., NVMe functions for reading and writing application data to EBS storage in the Amazon cloud) when running the code and data processing operations of a customer application (i.e., the first computing operations generate first processed data) that initialize microservices of the AWS Nitro DPU for offloading infrastructure services (e.g., storage) from the Amazon EC2 compute instances. These requests for the microservice initialization are not initiated by the AWS Nitro DPU (i.e., not executed by the offload processing modules).

227. On information and belief, the AWS Nitro DPU execute second computing operations in parallel on the first processed data that serve to execute the requested microservice for the customer application. For example, the AWS Nitro DPU maps NVMe functions for logical EBS volumes to the Amazon EC2 compute instances, thereby allowing reading and writing of application data to memory addresses of the EBS volumes (i.e., executing second computing operations on the first processed data).

228. Amazon thus infringes at least claim 1 of the '092 Patent through its AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

229. Amazon also infringes at least claim 1 of the '092 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA, ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

230. On information and belief, Amazon EC2 (Elastic Cloud Compute) server platforms, including Amazon EC2 UltraClusters, form a distributed computing architecture. Amazon EC2 UltraClusters utilizes NVIDIA's server platforms, including Blackwell-based server platforms with GPUs connected by NVLink Switch DPUs that provide an accelerated platform for complex ML/AI workloads.

231. As another example, on information and belief, the Amazon EC2 instances, including Amazon EC2 UltraClusters, also utilizes Hopper-based server platforms, including NVIDIA GH200 NVL32 platform, to accelerate training for complex LLMs and generative AI applications.

232. On information and belief, Amazon EC2 UltraClusters forming a distributed computing architecture include a plurality of servers. Amazon EC2 UltraClusters comprising

NVIDIA GB200 NVL72 servers include multiple compute nodes, interconnected via Amazon's Elastic Fabric Adapter (EFA) networking. Furthermore, Amazon EC2 UltraClusters comprising NVIDIA GB200 NVL72 servers include multiple compute nodes, interconnected via Amazon's Elastic Network Adapter (ENA) providing enhanced networking capabilities.

233. On information and belief, the compute nodes of Grace Blackwell server, e.g., GB200 NVL72 server, includes Grace Blackwell Superchips with Grace CPUs (i.e., central processing units (CPUs)).

234. As another example, on information and belief, each compute node of the Grace Hopper server also includes a Grace Hopper Superchip with Grace CPUs.

235. On information and belief, the Blackwell-based server includes 5th Generation NVLink and NVLink Switch DPUs (i.e., offload processing modules) that provide high-speed, seamless communication between every GPU in server clusters.

236. On information and belief, the NVLink Switch DPU is coupled to the Grace Blackwell Superchip including the CPU by a NVLink Cable Cartridge bus.

237. On information and belief, each NVLink Switch DPU of the GB200 NVL72 server includes Core Logic and Management Logic elements (i.e., a plurality of computation elements), including hardware accelerator engines for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP) for performing in-network computing operations including reductions and multicast accelerations, used for offloading communication collective operations from server processors.

238. On information and belief, the NVLink Switch DPUs include a virtual switch that is programmable for switching together multiple SHARP hardware accelerator engines and output

ports into software-defined hardware accelerator pipelines to process data packets according to their CUDA context and enable in-network reductions offloaded from server processors.

239. On information and belief, the virtual switches fully connect all Superchips in the server system and form a non-blocking compute fabric (i.e., switch fabric) for exchanging data between the NVLink Switch DPUs, within a single rack and between racks, to offload brokering of communication collective operations from server processors.

240. On information and belief, the virtual switches implemented in the NVLink Switch DPUs of the server modules form a non-blocking compute fabric (i.e., switching plane) for exchanging data across the cluster of Superchips.

241. On information and belief, ML/AI training processes with large datasets involve AllReduce operations, where each individual GPU in the network performs arithmetic operations (i.e., first computing operations) to generate partial local gradients (i.e., first processed data). These arithmetic operations are not executed by the NVLink Switch DPUs.

242. On information and belief, the NVLink Switch DPUs utilize the hardware accelerator engines for SHARP to offload and accelerate communication collective operations (i.e., second computing operations) from server processors. Instead of distributing the data to each GPU, the GPUs send the first processed data to the NVLink Switch DPUs, which perform in-network computing operations including reductions (i.e., second computing operations) to generate results (i.e., second processed data) and multicast acceleration to share the results back to GPUs, thereby significantly reducing the number of operations and time required for large-scale model training in distributed ML/AI processing.

243. Amazon thus infringes at least claim 1 of the ’092 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA, ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

244. Amazon also infringes at least claim 1 of the ’092 Patent through its AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

245. On information and belief, Amazon storage server systems forming a distributed computing architecture comprise AWS Elastic Block Storage (EBS) servers (i.e., a plurality of servers).

246. As another example, on information and belief, Amazon storage server systems forming a distributed computing architecture comprise AWS storage-optimized servers (i.e., a plurality of servers).

247. On information and belief, AWS storage servers, including AWS EBS servers and AWS storage-optimized servers include host processors (i.e., at least one central processing unit (CPU)) on the system main board.

248. On information and belief, the AWS storage servers include AWS Nitro System with Nitro Controller and Nitro Cards (i.e., offload processor modules) to enable the offloading of data plane infrastructure functions, and accelerate the data plane infrastructure services of its data centers, including big data and distributed ML/AI services.

249. Furthermore, on information and belief, Amazon’s Elastic Fabric Adapter is built on the AWS Nitro DPU, to provide high levels of inter-node communications between AWS storage servers.

250. As another example, on information and belief, Amazon's Elastic Network Adapter is also built on the AWS Nitro DPU, to provide enhanced networking capabilities in AWS storage servers.

251. On information and belief, the AWS Nitro DPU is coupled to the host processor (i.e., the main processor) via a PCIe bus (i.e., a bus).

252. On information and belief, the AWS Nitro DPU includes dedicated hardware components with 64-bit processing capabilities and specialized application-specific integrated circuits (ASICs) that function as hardware accelerators (i.e., computation elements). The hardware accelerators of the AWS Nitro DPU includes specific-purpose hardware accelerators and general-purpose hardware accelerators (ARM cores).

253. On information and belief, the AWS Nitro DPU implements a virtual switch for software-defined networking that is programmable (for example, by P4 programming language) for switching together different software-defined pipelines of hardware accelerators to address the different needs of network packets flows of different cloud users or services, to offload the data plane functions of the host processors.

254. On information and belief, the AWS Nitro DPU in each AWS storage server facilitates communication by forming a switching fabric for exchanging data between them.

255. For example, on information and belief, the AWS Nitro DPU, interconnected by Elastic Fabric Adapter (EFA) networking, facilitates communication by forming a switching fabric for exchanging data between the Nitro Controllers.

256. On information and belief, the AWS storage servers are supported by the AWS Nitro DPU in running different customer applications, including code and data processing operations. For example, the AWS storage servers are configured to generate requests to the AWS

Nitro DPU (e.g., NVMe functions for reading and writing application data to EBS storage in the Amazon cloud) when running the code and data processing operations of a customer application (i.e., the first computing operations generate first processed data) that initialize microservices of the AWS Nitro DPU for offloading infrastructure services (e.g., storage) from the Amazon EC2 storage-optimized instances. These requests for the microservice initialization are not initiated by the AWS Nitro DPU (i.e., not executed by the offload processing modules).

257. On information and belief, the AWS Nitro DPU executes second computing operations in parallel on the first processed data that serve to execute the requested microservice for the customer application. For example, the AWS Nitro DPU maps NVMe functions for logical EBS volumes to the AWS storage servers, thereby allowing reading and writing of application data to memory addresses of the EBS volumes (i.e., executing second computing operations on the first processed data).

258. Amazon thus infringes at least claim 1 of the ’092 Patent through its AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA).

259. Amazon operates and sells within the United States access to its (1) AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), its (2) AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and its (3) AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), thereby directly infringing at least claim 1 of the ’092 Patent.

260. The infringing aspects of the ’092 Accused Instrumentalities can be used only in a manner that infringes the ’092 Patent and thus have no substantial non-infringing uses. The infringing aspects of those instrumentalities otherwise have no meaningful use, let alone any meaningful non-infringing use.

261. Amazon has had actual or constructive knowledge and notice of the ’092 Patent and its infringement since at least March of 2024, when Xockets communicated with Amazon about the Asserted Patents, and through the filing and service of the Complaint. Despite this knowledge, Amazon continued to commit the infringing acts described herein.

262. Amazon makes and sells hardware and/or software components (e.g., its (1) AWS Compute Server Systems (including AWS Compute Servers, AWS Trainium Servers, and AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA), its (2) AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), and its (3) AWS Storage Server Systems (including AWS EBS Servers and AWS Storage-Optimized Servers) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA)) especially made or especially adapted to practice the invention claimed in the ’092 Patent, including at least claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

263. On information and belief, Amazon’s infringement of the ’092 Patent is willful and deliberate, entitling Xockets to the recovery of enhanced damages under 35 U.S.C. § 284. Amazon

has infringed and continues to infringe the ’092 Patent despite the risk of infringement being either known or so obvious that it should have been known to Amazon.

264. Amazon’s acts of infringement have caused and continue to cause damage to Xockets, and Xockets is entitled to recover from Amazon the damages it has sustained as a result of those wrongful acts in an amount subject to proof at trial, but in no event less than a reasonable royalty for the use made of the invention in the ’092 Patent, together with interest and costs as fixed by the Court.

COUNT THREE
Infringement of the ’161 Patent

265. Xockets repeats and incorporates by reference each preceding paragraph as if fully set forth herein and further states:

266. On June 28, 2016, the United States Patent and Trademark Office duly and legally issued the ’161 Patent, entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors.” A true and correct copy of the ’161 Patent is attached as Exhibit 3 to this Complaint.

267. The ’161 Patent relates to improved systems, hardware, and methods for cloud data centers by creating a rack level server system having offload processor modules, or DPUs, connected together in a novel switching architecture to form a new switching plane or cloud fabric.

268. Xockets holds all substantial rights in and to the ’161 Patent, including the exclusive right to assert all causes of action under the ’161 Patent and the exclusive right to any remedies for the infringement of the ’161 Patent.

269. Amazon is not licensed under the ’161 Patent, either expressly or implicitly.

270. Amazon has and continues, without authorization, to operate and use, and/or to induce and contribute to the operation and use by others of equipment and services that practice

one or more claims of the ’161 Patent literally or under the doctrine of equivalents (hereafter “’161 Accused Instrumentalities”). At a minimum, the ’161 Accused Instrumentalities include Amazon’s AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

271. Amazon has directly infringed and continues to directly infringe, literally and/or under the doctrine of equivalents, at least claim 1 of the ’161 Patent under 35 U.S.C. § 271(a) by operating and using the ’161 Accused Instrumentalities in the United States.

272. For example, Amazon infringes at least claim 1 of the ’161 Patent. Claim 1 discloses:

> A rack server system for a packet processing, comprising:
>
> a plurality of servers mountable in a rack;
>
> a top of rack (TOR) unit having connections to each of the servers;
>
> a plurality of offload processor modules, each offload processor module having at least one input-output (IO) port and multiple offload processors, including at least a first offload processor module connected directly to a second offload processor module through their respective IO ports, the offload processor modules are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module; and
>
> a memory controller configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

273. Amazon infringes at least claim 1 of the ’161 Patent through its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

274. On information and belief, Amazon EC2 (Elastic Cloud Compute) server platforms, including Amazon EC2 UltraClusters, form a rack server system. Amazon EC2 server system

utilizes Blackwell-based server platforms including GB200 NVL72 platform, in a rack scale system, that provide an accelerated platform for complex ML/AI workloads. As another example, on information and belief, the Amazon EC2 instances, including Amazon EC2 UltraClusters, also utilize Hopper-based server platforms, including GH200 NVL32 platform, as a rack scale system, allowing to accelerate training for complex LLMs and generative AI applications.

275. On information and belief, Amazon EC2 rack server system utilizes Blackwell-based server platforms to perform packet processing. The Blackwell Platform delivers trillion-parameter Large Language Model (LLM) training and real-time inference. The Blackwell server system includes 72 Blackwell GPUs connected by NVLink Switch DPUs that provide an accelerated platform for complex ML/AI workloads.

276. On information and belief, Amazon EC2 rack server systems utilizing Blackwell-based platform includes multiple servers mountable in a rack. The GB200 NVL72 rack server system includes multiple compute nodes (i.e., plurality of servers) mountable in a rack. Furthermore, Amazon EC2 server platforms utilizing the GB200 NVL72 rack server system including multiple compute nodes (i.e., plurality of servers) mountable in a rack are interconnected via Amazon's Elastic Fabric Adapter (EFA) networking. As a further example, Amazon EC2 server platforms utilizing the GB200 NVL72 rack server system including multiple compute nodes (i.e., plurality of servers) mountable in a rack are interconnected via Amazon's Elastic Network Adapter (ENA) providing enhanced networking capabilities

277. On information and belief, Amazon EC2 rack server systems including the GB200 NVL72 servers comprise a top of rack (TOR) switch that connects to each of the compute server nodes.

278. On information and belief, Amazon EC2 rack server systems utilizing the Blackwell-based platform, including GB200 NVL72 platform, comprises multiple NVLink Switch DPUs (i.e., offload processor modules), connected to the compute nodes. The NVLink Switches provide high-speed, seamless communication between every GPU in server clusters. The GB200 NVL72 server system includes multiple NVLink Switch DPUs (i.e., offload processor modules). The NVLink Switch DPU is coupled to the Grace Blackwell Superchip including the CPU by a NVLink Cable Cartridge bus.

279. On information and belief, each NVLink Switch DPU includes Core Logic and Management Logic elements, including hardware accelerator engines (i.e., offload processors) for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP) for performing in-network computing operations including reductions and multicast acceleration, used for offloading communication collective operations from server processors. Each NVLink Switch DPU includes IO ports for direct connectivity.

280. On information and belief, the NVLink Switch DPUs are connected directly with each other and with Grace Blackwell Superchips to form a non-blocking switching fabric. An NVLink Switch DPU (i.e., a first offload processor module) is connected directly to another NVLink Switch DPU (i.e., second offload processor module) through their respective IO ports.

281. On information and belief, the NVLink Switch DPUs are connected to a memory bus and are further configured to receive network packets from the server through the memory bus and from the IO port on the NVLink Switch DPU to perform hardware acceleration of packet flows for NVIDIA SHARP in-network reductions.

282. On information and belief, the NVLink Switch DPU is connected to memory associated with the Grace Blackwell Superchips such as LPDDR5X and/or HBM3e which include

a memory controller configured to stream the network packet data directly to at least one NVLink Switch DPU via the memory bus to which the NVLink Switch DPU is attached.

283. Amazon thus infringes at least claim 1 of the ’161 Patent.

284. Amazon operates and sells within the United States access to its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), thereby directly infringing at least claim 1 of the ’161 Patent.

285. The infringing aspects of the ’161 Accused Instrumentalities can be used only in a manner that infringes the ’161 Patent and thus have no substantial non-infringing uses. The infringing aspects of those instrumentalities otherwise have no meaningful use, let alone any meaningful non-infringing use.

286. Amazon has had actual or constructive knowledge and notice of the ’161 Patent and its infringement since at least March of 2024, when Xockets communicated with Amazon about the Asserted Patents, and through the filing and service of the Complaint. Despite this knowledge, Amazon continued to commit the infringing acts described herein.

287. Amazon makes and sells hardware and/or software components (e.g., its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster)) especially made or especially adapted to practice the invention claimed in the ’161 Patent, including at least claim 1, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

288. On information and belief, Amazon’s infringement of the ’161 Patent is willful and deliberate, entitling Xockets to the recovery of enhanced damages under 35 U.S.C. § 284. Amazon has infringed and continues to infringe the ’161 Patent despite the risk of infringement being either known or so obvious that it should have been known to Amazon.

289. Amazon’s acts of infringement have caused and continue to cause damage to Xockets, and Xockets is entitled to recover from Amazon the damages it has sustained as a result of those wrongful acts in an amount subject to proof at trial, but in no event less than a reasonable royalty for the use made of the invention in the ’161 Patent, together with interest and costs as fixed by the Court.

COUNT FOUR
Infringement of the ’640 Patent

290. Xockets repeats and incorporates by reference each preceding paragraph as if fully set forth herein and further states:

291. On September 6, 2016, the United States Patent and Trademark Office duly and legally issued the ’640 Patent, entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors.” A true and correct copy of the ’640 Patent is attached as Exhibit 4 to this Complaint.

292. The ’640 Patent generally relates to systems, hardware, and methods for cloud data centers to create a rack server system with offload processor modules, or DPUs, for in-network reduction/combining of data-intensive workloads using map/reduce data processing, which is critical in training large language models for AI.

293. Xockets holds all substantial rights in and to the ’640 Patent, including the exclusive right to assert all causes of action under the ’640 Patent and the exclusive right to any remedies for the infringement of the ’640 Patent.

294. Amazon is not licensed under the ’640 Patent, either expressly or implicitly.

295. Amazon has and continues, without authorization, to operate and use, and/or to induce and contribute to the operation and use by others of equipment and services that practice one or more claims of the ’640 Patent literally or under the doctrine of equivalents (hereafter "’640 Accused Instrumentalities"). At a minimum, the ’640 Accused Instrumentalities include Amazon’s AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster).

296. Amazon has directly infringed and continues to directly infringe, literally and/or under the doctrine of equivalents, at least claim 9 of the ’640 Patent under 35 U.S.C. § 271(a) by operating and using the ’640 Accused Instrumentalities in the United States.

297. For example, Amazon infringes at least claim 9 of the ’640 Patent. Claim 9 discloses:

> A rack server system for a map/reduce data processing, comprising:
>
> a plurality of servers arranged in a rack,
>
> a plurality of offload processor modules supported on at least two of the servers, each offload processor module having an input-output (IO) port and multiple offload processors, a first offload processor module configured to execute map steps of the map/reduce data processing, and being connected directly to a second offload processor through their respective IO ports to define a midplane switch, and
>
> a top of rack (TOR) unit connected to each of the servers that does not transfer map/reduce data, wherein
>
> a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module.

298. On information and belief, Amazon EC2 (Elastic Cloud Compute) server platforms, including Amazon EC2 UltraClusters, form a rack server system. Amazon EC2 server platform utilizes Blackwell-based server systems, including GB200 NVL72 platform, arranged in a rack

scale system, that provide an accelerated platform for complex ML/AI workloads. As another example, on information and belief, the Amazon EC2 instances, including Amazon EC2 UltraClusters, also utilize Hopper-based server platforms, including GH200 NVL32 platform, to form a rack-scale system, allowing accelerated training for complex LLMs and generative AI applications.

299. On information and belief, Amazon EC2 rack server systems including GB200 NVL72 system performs map/reduce data processing. The GB200 NVL72 system within the Blackwell Platform delivers trillion-parameter Large Language Model (LLM) training and real-time inference. The NVIDIA GB200 NVL72 server system includes 72 NVIDIA Blackwell GPUs connected by NVIDIA's NVLink Switch DPUs that provide an accelerated platform for complex ML/AI workloads including ML/AI collective operations such as AllReduce used in ML/AI training.

300. On information and belief, Amazon EC2 rack server systems utilizing Blackwell-based platform includes multiple servers arranged in a rack. The GB200 NVL72 platform comprises multiple NVIDIA GB200 compute nodes (i.e., plurality of servers) arranged in a rack. Furthermore, Amazon EC2 server platforms utilizing the GB200 NVL72 rack server system including multiple compute nodes (i.e., plurality of servers) mountable in a rack are interconnected via Amazon's Elastic Fabric Adapter (EFA) networking. As a further example, Amazon EC2 server platforms utilizing the GB200 NVL72 rack server system including multiple compute nodes (i.e., plurality of servers) mountable in a rack are interconnected via Amazon's Elastic Network Adapter (ENA) providing enhanced networking capabilities.

301. On information and belief, Amazon EC2 rack server systems utilizing Blackwell-based server platform, including GB200 NVL72 platform, comprises multiple NVLink Switch

DPUs (i.e., offload processor modules), connected to the compute servers. The NVLink Switch DPUs provide high-speed, seamless communication between every GPU in server clusters.

302. On information and belief, the GB200 NVL72 server system comprises multiple NVLink Switch DPUs (i.e., offload processor modules) supported on at least two of compute servers. The NVLink Switch DPU is coupled to the Grace Blackwell Superchip including the CPU by a NVLink Cable Cartridge bus.

303. On information and belief, each NVLink Switch DPU includes Core Logic and Management Logic elements, including hardware accelerator engines (i.e., offload processors) for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP) for performing in-network computing operations including reductions and multicast acceleration, used for offloading communication collective operations from server processors. Each NVLink Switch DPU includes IO ports for direct connectivity.

304. On information and belief, the NVLink Switch DPUs are connected directly with each other and with Grace Blackwell Superchips to form a non-blocking switching fabric and define a midplane switch. An NVLink Switch DPU (i.e., a first offload processor module) is connected directly to another NVLink Switch DPU (i.e., second offload processor module) through their respective IO ports to define a midplane switch.

305. On information and belief, Amazon EC2 rack server systems, including GB200 NVL72 servers, comprise a top of rack (TOR) switch that connects to each of the compute server nodes. The NVLink Switch DPUs execute map/reduce data processing using offload engines for in-network reductions and multicast acceleration and enable all-to-all communication between GPUs using their direct connections. As a result, the ToR switch is not used to transfer map/reduce data among compute node servers arranged in the rack.

306. On information and belief, the NVLink Switch DPUs accelerate ML/AI collective communications including AllReduce operations used in ML/AI training. In AllReduce, a first NVLink Switch DPU is configured to execute map steps of the map/reduce data processing on messages received from GPUs by assigning each message to appropriate destination queue(s) of a second NVLink Switch DPU in the communication collective group for data exchange.

307. On information and belief, the NVLink Switch DPUs of the compute nodes of the GB200 NVL72 server also include the second NVLink Switch DPU configured to execute reduce steps of the AllReduce operations by performing reduction operations on messages provided from the first NVLink Switch DPU, thereby significantly reducing the number of operations and training time required for large-scale model training in distributed ML/AI processing.

308. Amazon thus infringes at least claim 9 of the ’640 Patent.

309. Amazon operates and sells within the United States access to its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster), thereby directly infringing at least claim 9 of the ’640 Patent.

310. The infringing aspects of the ’640 Accused Instrumentalities can be used only in a manner that infringes the ’640 Patent and thus have no substantial non-infringing uses. The infringing aspects of those instrumentalities otherwise have no meaningful use, let alone any meaningful non-infringing use.

311. Amazon has had actual or constructive knowledge and notice of the ’640 Patent and its infringement since at least March of 2024, when Xockets communicated with Amazon about the Asserted Patents, and through the filing and service of the Complaint. Despite this knowledge, Amazon continued to commit the infringing acts described herein.

312. Amazon makes and sells hardware and/or software components (e.g., its AWS Compute Server Systems (including AWS UltraCluster) with the AWS Nitro System or Nitro DPU (including AWS EFA and ENA) and NVLink Switch DPU (Blackwell/Hopper Cluster)) especially made or especially adapted to practice the invention claimed in the ’640 Patent, including at least claim 9, and that (i) is a material part of the invention and (ii) is not a staple article or commodity of commerce suitable for substantial non-infringing use at least because it is specifically designed to perform the claimed functionality. Any other use of such hardware and/or software would be unusual, far-fetched, illusory, impractical, occasional, aberrant, or experimental.

313. On information and belief, Amazon’s infringement of the ’640 Patent is willful and deliberate, entitling Xockets to the recovery of enhanced damages under 35 U.S.C. § 284. Amazon has infringed and continues to infringe the ’640 Patent despite the risk of infringement being either known or so obvious that it should have been known to Amazon.

314. Amazon’s acts of infringement have caused and continue to cause damage to Xockets, and Xockets is entitled to recover from Amazon the damages it has sustained as a result of those wrongful acts in an amount subject to proof at trial, but in no event less than a reasonable royalty for the use made of the invention in the ’640 Patent, together with interest and costs as fixed by the Court.

INJUNCTIVE RELIEF

315. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

316. As a remedy for Amazon’s willful infringement of Xockets’ patents, Xockets, seeks to enjoin Amazon from making, using, selling, and offering to sell the Accused Instrumentalities.

317. Xockets seeks to uphold its constitutional promise of exclusive rights to its patented inventions, as protected by Congress in the United States patent laws. These exclusive rights are

protected for limited periods of time following the grant of a patent to incentivize innovators to innovate and investors to invest in commercializing those innovations.[60]

318. Xockets has preserved its exclusive rights and only licensed its intellectual property rights on an exclusive basis for making and selling Xockets' StreamSwitch.

319. As explained above, Amazon has benefited from the immense value of Dr. Dalal's groundbreaking DPU and virtual switch computing architecture for a new cloud processor, and its DPU switching plane architecture for a new cloud fabric, which have revolutionized the world of distributed computing in data centers, and led a transition of data centers to machine learning and artificial intelligence.

320. Amazon continues to expand the scope of its widespread and willful infringement. Amazon CEO Andy Jassy recently announced Amazon's plans to spend $100 billion on AI infrastructure in 2025.[61] Jassy told investors in February 2025 that this is a "once-in-a-lifetime type of business opportunity."[62]

321. According to media reporting, Amazon's imminent plans "cover[] the full stack of infrastructure, focusing on components designed to improve training and inference tasks."[63] "Digital advancements include custom silicon chips, such as Tranium for training and Inferentia

[60] https://www.sonecon.com/docs/studies/Value_of_Intellectual_Capital_in_American_Economy.pdf ("Innovation is widely recognized by economists as the most powerful factor that can drive changes in an economy's underlying rates of productivity and growth. The quality of the new ideas embodied in those innovations and the pace at which innovations are developed and applied, therefore, significantly affect a nation's prosperity. . . . One legal aspect is especially critical to the development and broad application of economically-powerful ideas – the strict protection and enforcement of intellectual property rights. Without such protections and enforcement, innovators have little incentive, especially to develop new technologies, materials and production processes.").

[61] https://www.cnbc.com/2025/02/06/amazon-expects-to-spend-100-billion-on-capital-expenditures-in-2025.html.

[62] https://www.cnbc.com/2025/02/06/amazon-expects-to-spend-100-billion-on-capital-expenditures-in-2025.html.

[63] https://voip.review/2025/02/10/amazon-unveils-ambitious-100b-ai-infrastructure-investment-plan.

for inference," which "address customer needs for cost-effective AI workloads and higher performance." Amazon's planned infringement includes the development of a data center in Round Rock, Texas, in this District.[64]

322. Amazon's widespread and unauthorized infringement casts a cloud over Xockets' exclusive rights to its patented inventions, resulting in a material diminution in their market value and Xockets' bargaining power. This harm is irreparable, and an injunction is necessary to stop continued infringement. No buyer or exclusive licensee will pay for the value of an exclusive right to the property when the largest cloud, Amazon's AWS, is sitting on the property and unwilling to leave.

323. As a result of Amazon's unauthorized infringement, Xockets has struggled to attract investors, effectively foreclosing Xockets from capitalizing on years of research and development.

324. Amazon's massive and imminent AI-infrastructure plans will cause additional irreparable harm to Xockets through (i) further eradication of Xockets' footprint in this incredibly valuable, emerging technology market, (ii) further devaluation of Xockets' exclusive patent rights by Amazon's unauthorized and extensive use of Xockets' intellectual property, and (iii) further destruction of Xockets' business opportunities.

325. In 2024, Amazon declined to engage in good-faith business discussions with Xockets.[65] Rather than pay fair value for Xockets' exclusive patent rights before incorporating the

[64] https://communityimpact.com/austin/round-rock/development/2025/04/01/mass-grading-begins-on-amazon-site-in-round-rock.

[65] *Wall Street Journal* reporter Dana Mattioli, in discussing her book, *The Everything War: Amazon's Ruthless Quest to Own the World and Remake Corporate Power*, describes Amazon's "pattern of lying, cheating, copying, their way to the top and using their leverage in all these different industries to crush competition." *See* https://podcasts.apple.com/us/podcast/amazons-everything-war-with-dana-mattioli/id730188152?i=1000680976211. AWS's Director of Product Management for the Amazon EC2 Accelerated Computing Portfolio described the pressure to innovate in the AI space and remarked,

patented DPU technologies in the Accused Products, Amazon chose to employ "an 'infringe now, pay later' strategy," irreparably harming Xockets.[66]

326. Here, Amazon extensively used Xockets' patented technologies, in deliberate contravention of Xockets' legitimate patent rights, for which Xockets has been granted exclusive rights under United States patent laws. Amazon then refused to engage in negotiations with Xockets but has continued to reap extraordinary profits from its infringing sales in a trillion-dollar cloud industry, dwarfing any court-ordered reasonable royalty that Xockets could possibly recover after years of protracted litigation.

327. It is impossible to calculate the difference in the value Xockets would receive from a negotiated exclusive license to its patented technology as compared to the amounts Xockets would collect in compulsory license payments through litigation if no injunctive relief is granted. This is the epitome of irreparable harm.

328. Awarding injunctive relief also serves the public interest because "[a]bsent the threat of a permanent injunction, the incentives to 'engage in the toils of scientific and technological research', are reduced if not eliminated."[67]

329. Amazon's wrongful infringement has caused Xockets to suffer irreparable harm, resulting from the loss of its lawful patent rights to exclude others from making, using, selling,

"if we can leverage an IP from a third party, from a partner, then great. That's going to enable us to get to market quicker." https://redmonk.com/videos/a-redmonk-conversation-ai-and-custom-silicon-at-annapurna-labs-with-aws.

[66] Kristen J. Osenga, *The Loss of Injunctions Under* eBay*: Evidence of the Negative Impact on the Innovation Economy, Hudson Institute* (Feb. 28, 2024), https://www.hudson.org/regulation/loss-injunctions-under-ebay-evidence-negative-impactinnovation-economy ("In patent law, this is now known as 'predatory infringement', in which defendants choose a commercial strategy of 'infringe now, pay later,' at worst, or, at best, they get away with infringement through a lengthy legal battle of attrition in which the patent owner ultimately just gives up. . . . [A] very big problem because it reduces the ability of patent owners to obtain returns on their investments that drive growth in the innovation economy.").

[67] Acri née Lybecker, Kristina M.L., *Injunctive Relief in Patent Cases: the Impact of* eBay (June 14, 2024), available at https://ssrn.com/abstract=4866108; *see also supra* n.64.

offering for sale, and importing the patented technology, as set forth in detail in the preceding paragraphs.

330. The balance of hardships weighs strongly in Xockets' favor. Amazon had notice of Xockets' Asserted Patents at least as early as March 2024. Amazon thus had an opportunity to negotiate to buy or take an exclusive license to the Asserted Patents, but chose not to participate in the sales process and instead expanded the scope of its willful infringement. Under the circumstances, an injunction is critical to stop the continuing harm.

DEMAND FOR JURY TRIAL

331. Xockets hereby demands a jury trial pursuant to Federal Rule of Civil Procedure 38.

FEES AND COSTS

332. To the extent that Amazon's willful and deliberate infringement or litigation conduct supports a finding that this is an "exceptional case," an award of attorney's fees and costs to Xockets is justified pursuant to 35 U.S.C. § 285.

PRAYER FOR RELIEF

WHEREFORE, Xockets prays for relief against Amazon as follows:

a. Declaring that Amazon has directly infringed the Asserted Patents;

b. Awarding Xockets damages arising out of this infringement of the Asserted Patents, including enhanced damages pursuant to 35 U.S.C. § 284 and supplemental damages for any continuing post-verdict infringement through entry of the final judgment, in an amount according to proof;

c. Awarding Xockets prejudgment and post-judgment interest, in an amount according to proof;

d. Awarding Xockets a compulsory ongoing royalty, in an amount according to proof;

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

f. Declaring that Amazon's infringement of the Asserted Patents is willful;

g. Enjoining Amazon's infringement of the Asserted Patents;

h. Awarding to Xockets such other costs, equitable relief, and any other relief to which Xockets is entitled and as the Court deems just and proper.

Dated: June 30, 2025

Respectfully submitted,

SUSMAN GODFREY LLP

/s/ Kalpana Srinivasan
Joseph S. Grinstein
Texas State Bar No. 24002188
Hunter A. Vance
Texas State Bar No. 24102596
Allen J. Hernandez
Texas State Bar No. 24132804
1000 Louisiana Street, Suite 5100
Houston, TX 77002
(713) 651-9366
jgrinstein@susmangodfrey.com
hvance@susmangodfrey.com
ahernandez@susmangodfrey.com

Kalpana Srinivasan
California State Bar No. 237460
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
(310) 789-3100
ksrinivasan@susmangodfrey.com

Tamar Lusztig
New State York Bar No. 5125174
One Manhattan West
New York, NY 10001
(212) 336-8330
tlusztig@susmangodfrey.com

Attorneys for Plaintiff Xockets, Inc.