Michael Matulewicz-Crowley*
mmc@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
650 Fifth Avenue, Suite 2320
New York, NY 10019
Telephone: (212) 381-1965

Christine E. Lehman*
clehman@reichmanjorgensen.com
Zachary L. Jacobs*
zjacobs@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
1909 K Street, NW, Suite 800
Washington, DC 20006
Telephone: (202) 894-7311

Laura Carwile (CA Bar No. 291906)
lcarwile@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
100 Marine Parkway, Suite 300
Redwood Shores, CA 94065
Telephone: (650) 623-1401

**Pro hac vice* forthcoming

Attorneys for Plaintiffs
Concurrent Ventures, LLC and
XtreamEdge, Inc.

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA

Concurrent Ventures, LLC and XtreamEdge, Inc.,	§	
	§	
Plaintiffs,	§	
v.	§	**CIVIL ACTION NO. 4:26-cv-7455**
	§	
NVIDIA Corporation, Mellanox Technologies, Ltd., and Mellanox Technologies, Inc.	§	**JURY TRIAL DEMANDED**
	§	
Defendants.	§	

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiffs Concurrent Ventures, LLC ("Concurrent Ventures") and XtreamEdge, Inc. ("XtreamEdge") (collectively, "Plaintiffs") file this Complaint against Defendants NVIDIA Corporation ("NVIDIA") and Mellanox Technologies, Ltd. and Mellanox Technologies, Inc. ("Mellanox") (collectively, "Defendants"), and in support thereof alleges as follows:

I.
NATURE OF THE ACTION

1. Data storage, manipulation, and use serves as the foundation for the global economy and have all increased exponentially over the past years. Data is integral to how our society operates—how we communicate, interact, and engage with entertainment on personal computing devices; how we do business; and how we power the artificial intelligence ("AI") revolution.

2. As data becomes even more firmly entrenched in society, data processing and data functions require more resources. The traditional solution to this problem is to add more processors, for example in the form of additional Central Processing Units ("CPUs") or Graphics Processing Units ("GPUs"). CPUs are general purpose hardware that implement functionality in software. This allows for a flexible approach, but these software solutions are slower and more resource-intensive than hardware-based solutions, and do not work well at scale.

3. GPUs are processors that were originally created to process and display graphics but now form the cornerstone of AI and Large Language Model ("LLM") processing. The artificial intelligence ("AI") revolution has compounded the data processing problem. GPUs form the backbone of the current generative AI boom and as AI usage explodes, so too does the number of GPUs necessary to keep up. "Whether you're looking to solve business problems in deep learning and AI, HPC, graphics, or virtualization in the data center or at the edge, NVIDIA GPUs provide the ideal solution. Now, you can realize breakthrough performance with fewer, more powerful servers, while driving faster time to insights

and reducing costs."[1] GPU sales have correspondingly skyrocketed in recent years. Defendants themselves advertise that their GPUs and CPUs are specially designed for AI uses.[2]

4. Adding more processors is not sustainable. Additional CPUs lead to higher capital expenditures and operational expenditures, lower flexibility, and increased latency. Simply using CPUs and GPUs worked as a stopgap for years because the power of CPUs increased at a sufficient pace in accordance with Moore's Law.[3] But we are nearing the end of Moore's Law. Similarly, as the demands of AI have grown exponentially, the demands on GPUs have grown exponentially, and adding more GPUs is insufficient. In particular, the number of GPUs cannot keep up with the needs of generative AI, both because of the financial and logistical strain of building, purchasing, and implementing massive numbers of GPUs in datacenters and also because of the environment impacts of power and water consumption from those GPU datacenters. As data and AI use and consumption continues to increase, there is a need for more efficient data processing

5. The inventors of the Asserted Patents addressed this need through a number of innovations that increase the efficiency of communications between processors and other network components. Plaintiffs' innovations, among other benefits, solves long-standing problems by offloading data-centric computations, which are estimated to be between 24-70% of the workload,[4] to more efficient units designed to handle data computations. By offloading the networking, communication, and other tasks from other processors, those processors are freed up to focus on what they do best.

---

[1] Exhibit 6, https://resources.nvidia.com/l/en-us-gpu?ncid=no-ncid; *see also* Exhibit 7, https://www.bitdeer.ai/en/blog/nvidia-gpu-evolution-and-the-road-ahead/.

[2] Exhibit 6; Exhibit 8, https://www.nvidia.com/en-us/data-center/grace-cpu/ ("NVIDIA CPU platforms support flexible AI factory designs, from tightly integrated CPU-GPU architectures like NVIDIA Vera Rubin and NVIDIA Grace Blackwell to standalone single- and dual-socket CPU servers. Vera extends the platform across agentic AI, reinforcement learning, analytics, orchestration, and storage, while Grace provides the proven CPU foundation for accelerated computing and power-efficient computing."); Exhibit 9, https://nvidianews.nvidia.com/news/nvidia-unveils-vera-the-cpu-for-agents ("AI agents will be the largest users of computing," said Jensen Huang, founder and CEO of NVIDIA. "Vera is the first CPU designed for that future — built to run agentic AI at hyperscale with extraordinary performance, efficiency and programmability."); Exhibit 10, https://developer.nvidia.com/blog/nvidia-vera-cpu-boosts-ai-factory-throughput-to-accelerate-agentic-workloads/.

[3] According to Moore's Law, the number of transistors in a dense integrated circuit doubles about every two years. Exhibit 11, https://en.wikipedia.org/wiki/Moore%27s_law; Exhibit 12, https://www.technologyreview.com/2020/02/24/905789/were-not-prepared-for-the-end-of-moores-law.

[4] Exhibit 13, https://blogs.nvidia.com/blog/bluefield-dpus-energy-efficiency/.

6. Plaintiffs patented the innovations necessary to create and use this specialized unit, which it referred to as a Stream Processing Unit ("SPU"). Although it was the first, it was not the last. Mellanox pursued similar solutions as XtreamEdge—years after XtreamEdge—but called its product a Data Processing Unit ("DPU").[5] NVIDIA completed its acquisition of Mellanox in 2020,[6] leading to the integration of Mellanox's BlueField and ConnectX DPU and DPU-enabled systems into NVIDIA's products, including its Graphics Processing Units ("GPU"), hardware chips, systems, and software, for use in various markets, such as data centers, artificial intelligence, machine learning, high-performance computing, gaming, creative design, autonomous vehicles, and robotics.

7. Defendants and industry leaders have all noted the importance of DPUs in the evolution of data processing and its role as the third leg of processing, alongside CPUs and GPUs. NVIDIA's CEO Jensen Huang said in 2020 that DPUs were "going to represent one of the three major pillars of computing going forward[.]" Exhibit 14, https://blogs.nvidia.com/blog/whats-a-dpu-data-processing-unit/ (May 20, 2020) ("CPUs and GPUs [are] powering sprawling new hyperscale data centers. These data centers are knit together with a powerful new category of processors. The DPU has become the third member of the data-centric accelerated computing model."). *See also* Exhibit 15, https://nvidianews.nvidia.com/news/nvidia-bluefield-4-powers-new-class-of-ai-native-storage-infrastructure-for-the-next-frontier-of-ai (Huang stating, "AI is revolutionizing the entire computing stack — and now, storage… With BlueField-4 [DPUs], NVIDIA and our software and hardware partners are reinventing the storage stack for the next frontier of AI."); Exhibit 16, https://blogs.nvidia.com/blog/accelerated-computing-networking-supercomputing-ai/ ("NVIDIA BlueField-4 DPUs are powering the operating system of AI factories. By offloading, accelerating and isolating critical data center functions — networking, storage and security — they free up CPUs and GPUs to focus entirely on compute-intensive workloads."). Huang called NVIDIA's acquisition of Mellanox "a homerun deal," stating, "Man, I've been dreaming of this." Exhibit 17, https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html ("We're combining the leaders of AI computing and high speed networking and data processing into one company

[5] Exhibit 18, https://network.nvidia.com/files/doc-2020/pb-bluefield-dpu.pdf.
[6] Exhibit 17.

and, so, this is really quite extraordinary."). Defendants have also recognized the importance of adaptive computing engines—such as those implementing DPUs—in its AI space.[7]

8. The patented technology only continues to grow in importance. In particular, Defendants' sale of the Accused Products is key to their ambitious plans to sell other additional products, including many of their CPUs and GPUs. Recent advertisements from Defendants confirm that (a) they bundle DPUs with CPUs and GPUs,[8] (b) DPUs, CPUs, and GPUs are functionally interrelated,[9] and (c) Defendants leverage that functional relationship to promote sales of CPUs and GPUs.[10] This is unsurprising given that the purpose of DPUs, discussed above, is to offload work from CPUs and GPUs, making those CPUs and GPUs of greater value to customers (and thus promoting increased sales of those CPUs and GPUs through their interrelated functioning with DPUs).

9. Defendants rely on the Accused Products as a part of their "end-to-end" AI and security architecture.[11] For example, "[a]t NVIDIA GTC Washington 2025, NVIDIA announced its next-generation Data Processing Unit (DPU) – NVIDIA BlueField-4. As an important component of the full-stack BlueField platform, it provides system-level acceleration for computing, networking, and security in AI data centers with 800Gb/s throughput and deep integration of the NVIDIA ConnectX-9 SuperNIC and the DOCA microservice framework."[12] "BlueField-4 is positioned as the 'operating system for AI factories,' aiming to deeply integrate computing, networking, and security into a unified, adaptive optimization system."[13] In fact, NVIDIA acquired Mellanox explicitly to combine Mellanox's infringing

---

[7] Exhibit 19, https://resources.nvidia.com/en-us-accelerated-networking-resource-library/turbocharging-ai-factories-with-dpu-accelerated-service-proxy-for-kubernetes; Exhibit 20, https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center; Exhibit 14; Exhibit 21, https://www.crn.com/news/components-peripherals/2025/nvidia-reveals-bluefield-4-dpu-packed-with-64-core-grace-cpu-for-ai-data-centers.

[8] Exhibit 22, https://resources.nvidia.com/en-us-accelerated-networking-resource-library/whats-a-dpu-data-product; Exhibit 20; Exhibit 14; Exhibit 21.

[9] Exhibit 19; Exhibit 20; Exhibit 14; Exhibit 23, https://www.crn.com/news/components-peripherals/2025/nvidia-reveals-bluefield-4-dpu-packed-with-64-core-grace-cpu-for-ai-data-centers.

[10] Exhibit 20; Exhibit 14; Exhibit 23.

[11] Exhibit 24, https://developer.nvidia.com/blog/advancing-ai-infrastructure-for-agentic-ai-with-nvidia-doca-in-silicon-security/.

[12] Exhibit 25, https://www.naddod.com/blog/nvidia-bluefield-4-dpu-powers-gigascale-ai-factories?srsltid=AfmBOopcx-oK-r24CInJ3SAqVJWXs5XUWMMqisIW29uCuE57CyfHWF_s.

[13] Exhibit 25.

DPUs with NVIDIA's products: "With Mellanox, the new NVIDIA has end-to-end technologies from AI computing to networking, full-stack offerings from processors to software, and significant scale to advance next-generation data centers."[14]

10. Defendants explicitly advertise that the Accused Products work together with CPUs and GPUs to improve those processors.[15] On information and belief, Defendants advertise the functional interrelation of these products to leverage the efficiencies provided by DPUs to sell more CPUs and GPUs. In short, customers would be less likely to buy Defendants' CPUs and GPUs if Defendants did not sell compatible DPUs to improve those other products. For example, Defendants state that their DPU allows customers to "offload" specific tasks that can "perform the work of 30, 100, or even 300 CPU cores when the workload meets the DPU's hardware acceleration capabilities."[16] Similarly, in a July 22, 2024 blog post, Defendants boasted that their DPU can reduce power consumption by "up to 30% by offloading essential data center networking and infrastructure functions from less efficient CPUs."[17] Defendants' promotional materials show that they used these synergies to promote bundled sales of DPUs and CPUs and GPUs.[18]

11. The functional relationship between DPUs and CPUs and GPUs has also been covered extensively by industry commentators and analysts. For example, one blog entry observed that DPUs and CPUs "complement each other in data-centric tasks. The combination of the two increases system performance and optimizes resource consumption."[19] The point of DPUs is to function with CPUs and GPUs, to make them more efficient. These developments underscore that CPUs and GPUs are convoyed sales to the infringing sales of DPUs, indicate the high priority Defendants place on their ability to provide the patented technology, and, in part, demonstrate the significant license Defendants would have agreed

[14] Exhibit 26, https://nvidianews.nvidia.com/news/nvidia-to-acquire-mellanox-for-6-9-billion; Exhibit 27, https://nvidianews.nvidia.com/news/nvidia-completes-acquisition-of-mellanox-creating-major-force-driving-next-gen-data-centers.
[15] Exhibit 22; Exhibit 20; Exhibit 14.
[16] Exhibit 28, https://developer.nvidia.com/blog/offloading-and-isolating-data-center-workloads-with-bluefield-dpu/.
[17] Exhibit 29, https://blogs.nvidia.com/blog/accelerated-ai-energy-efficiency/.
[18] *Id.*; Exhibit 22; Exhibit 20; Exhibit 14.
[19] Exhibit 30, https://monovm.com/blog/cpu-vs-dpu/.

to at a hypothetical negotiation if they had sought to fairly license the patented technology, instead of using it without permission.

12. On September 9, 2024, Plaintiffs communicated to Defendants that Defendants have been and are infringing the Asserted Patents. On October 18, 2024, Plaintiffs sent Defendants charts that demonstrated Defendants' infringement of the Asserted Patents in detail. On information and belief, Defendants took no action on or after those dates to avoid or mitigate their infringement of the Asserted Patents.

13. Defendants' refusal to acknowledge Plaintiffs' patents and offer fair compensation for their use violates the patent laws and undermines the effort that Plaintiffs took to develop their innovations. Plaintiffs feel they have no recourse but to file this action to stop Defendants' unauthorized use of Plaintiffs' patents. Plaintiffs invented something groundbreaking and Defendants are taking advantage of those innovations without permission. Accordingly, Plaintiffs brings this action under the patent laws, 35 U.S.C. § 1 et seq., in order to stop Defendants' willful infringement of U.S. Patent Nos. 8,924,596; 10,873,753; 10,985,943; 10,944,634; 9,529,767 (collectively, the "Asserted Patents").

II.
PARTIES

14. XtreamEdge, Inc. is a corporation organized and existing under the laws of the State of Delaware, with its principal place of business in Johns Creek, Georgia. XtreamEdge is the assignee of the Asserted Patents.

15. Concurrent Ventures, LLC is a limited liability corporation organized and existing under the laws of the State of Georgia, with its principal place of business in Johns Creek, Georgia. Concurrent Ventures is the exclusive licensee to the Asserted Patents and has all substantial rights to the Asserted Patents.

16. NVIDIA Corporation ("Nvidia" or "NVIDIA") is a public corporation organized and existing under the laws of Delaware with its principal place of business and headquarters at 2701 San Tomas Expressway, Santa Clara, CA 95050. NVIDIA has a registered agent for service of process at 1226 Tiros Way, Sunnyvale, CA 94086.

17. Mellanox Technologies, Ltd. ("Mellanox Ltd.") is an Israeli company with its principal place of business at Hakidma 26, Ofer Industrial Park, Yokneam 2069200 Israel. Mellanox Ltd. is a wholly owned subsidiary of Nvidia.

18. Mellanox Technologies, Inc. ("Mellanox Inc.") is a public corporation organized and existing under the laws of California with its principal place of business and headquarters at 350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085. Mellanox Inc. is a wholly owned subsidiary of Mellanox Ltd. Hereinafter, Defendants Mellanox Inc. and Mellanox Ltd. will be referred to as "Mellanox."

III.
ASSERTED PATENTS

19. U.S. Patent No. 8,924,596 (the "'596 Patent"), titled "System and Method for Dividing and Synchronizing a Processing Task Across Multiple Processing Elements/Processors in Hardware," issued on December 30, 2014. XtreamEdge is the current assignee and owner of the '596 Patent and Concurrent Ventures has exclusively licensed all substantial rights to the '596 Patent. Plaintiffs own all necessary rights, title, and interest in the '596 Patent to bring this action, including the right to seek damages, including past damages, for any infringement thereof. A true and correct copy of the '596 Patent is attached hereto as Exhibit 1. The '596 Patent claims patent-eligible subject matter and is valid and enforceable.

20. The '596 Patent provides "a system and method for dividing and synchronizing a processing task across multiple processing elements/processors in hardware." '596 Patent (Ex. 1), 1:11-13. "In the prior art, the processing load (of a processing element) to determine what command to issue [wa]s relatively high." *Id.*, 1:15-16. Before the '596 Patent, the task of determining what command to issue was performed by either (1) the processing element that executes (or processes) the command or (2) the processing element that issues the command itself. The inventors of the '596 Patent discovered that it was preferable for the task of determining what command to issue to be performed by a different processing element, distinct from the processing elements that issues the command or that executes it. *Id.*, 5:3-7:58. This innovative division allowed such determinations to be performed quickly and efficiently.

21. The inventors of the '596 Patent did not only figure out that the determination of when and what command to issue should be done by a different, distinct processor; they also figured out how that determination should be performed so that the various processing elements remain synchronized. In

particular, the inventors faced a problem that is inherent in a system in which "commands [are] determined from multiple processing elements." *Id.*, 1:20-22. "It can become a challenge to ensure the module command queues (FIFOs) are not overfilled, thus synchronization is required amongst the multiplicity of processing elements. In the prior art, such synchronization is performed using software techniques, such as semaphores and mutexes to a shared available space count variable in memory. However, this takes additional processing time, making this process non-optimal. To further compound the problem, there may be a need to keep the time from command issue to execution minimized, which tends to keep the command queue shallow." *Id.*, 1:22-32, 3:11-29.

22. The '596 Patent solves this technological problem by reciting a hardware reservation register and then, most importantly, setting out very specific relationships and steps that the various processing elements must take vis-à-vis each other, the hardware reservation register, and the various queues implemented in hardware. '596 Patent (Ex. 1), 3:30-4:19. For example, in claim 1, the claimed "second processing element" is in control of when and what commands are issued, while it is the claimed "first processing element" that must issue those commands. And dependent claim 6 recites a "third processing element" that then executes the command. Thus, claim 1 and its dependent claims set forth a particular architecture, with three distinct processing elements, two queues, and a reservation register, each of which have very particular roles and responsibilities. Claims 7 and 13 set out different arrangements. For example, in claim 7, the second processing element initially receives the command from the first processing element, and only then accesses the reservation register to determine whether the command should be issued, and finally, if so, issues the command itself.

23. These different architectural arrangements all have different benefits and tradeoffs, but all are specific technological solutions to a problem unique to parallel processing: how to quickly and efficiently determine what command to issue, while synchronizing the processing element performing that determination with the processing elements actually issuing the command and executing it. The '596 Patent provides concrete solutions to this computer-specific problem.

24. U.S. Patent No. 10,873,753 ("the '753 Patent"), titled "Hardware Defined Anything In A Platform With Swappable Pods, Message Interface, Sandboxes And Memory Superposition," issued on December 22, 2020. XtreamEdge is the current assignee and owner of the '753 Patent and Concurrent

Ventures has exclusively licensed all substantial rights to the ’753 Patent. Plaintiffs own all necessary rights, title, and interest in the ’753 Patent to bring this action, including the right to seek damages, including past damages, for any infringement thereof. A true and correct copy of the ’753 Patent is attached hereto as Exhibit 2. The ’753 Patent claims patent-eligible subject matter and is valid and enforceable.

25. U.S. Patent No. 10,985,943 (“the ’943 Patent), titled “Hardware Defined Anything In A Platform With Swappable Pods, Message Interface, Sandboxes And Memory Superposition,” issued on April 20, 2021. XtreamEdge is the current assignee and owner of the ’943 Patent and Concurrent Ventures has exclusively licensed all substantial rights to the ’943 Patent. Plaintiffs own all necessary rights, title, and interest in the ’943 Patent to bring this action, including the right to seek damages, including past damages, for any infringement thereof. A true and correct copy of the ’943 Patent is attached hereto as Exhibit 3. The ’943 Patent claims patent-eligible subject matter and is valid and enforceable.

26. The ’753 and ’943 Patents have nearly identical specifications and claim a “platform for data flow processing” and a “programmable device for data flow processing” respectively. ’753 and ’943 Patents (Exs. 2-3) at Abstract, claim 1. In particular, while “[s]oftware-based systems are appropriate for many data flow and data processing applications,” “there are many situations where dedicated electronic hardware is faster at processing data than software.” *Id.*, 1:8-11. Yet, “redesigning dedicated hardware each time a new data processing application or data formatting requirement comes up is an expensive proposition. And, designing, debugging and manufacturing dedicated electronic hardware is time-consuming, especially when revision cycles are taken into account.” *Id.*, 1:11-21. The ’753 and ’943 Patents recite specific solutions that balance the flexibility of software with the efficiency of hardware, through an architecture that carefully balances the two, a unique messaging system, and combination of memory arrangements.

27. U.S. Patent No. 10,944,634 (“the ’634 Patent”), titled “Optimization for Network Connections,” issued on March 9, 2021. XtreamEdge is the current assignee and owner of the ’634 Patent and Concurrent Ventures has exclusively licensed all substantial rights to the ’634 Patent. Plaintiffs own all necessary rights, title, and interest in the ’634 Patent to bring this action, including the right to seek damages, including past damages, for any infringement thereof. A true and correct copy of the ’634 Patent

is attached hereto as Exhibit 4. The ’634 Patent claims patent-eligible subject matter and is valid and enforceable.

28. The ’634 Patent claims an innovative technological architecture that improves network connectivity. The specification explains that, before the ’634 Patent, “[a] TCP/IP (transmission control protocol/Internet protocol) network connection between two endpoints starts off without knowledge of the performance of the connection and the effect of bandwidth realized between the peers.” ’634 Patent (Ex. 4), 1:6-9. In other words, each new network connection was unaware of the network connections that had come before, and thus could not benefit from the lessons learned from those previous connections about the optimal bandwidth to use. *Id*. If the new network connection sends data too quickly (bandwidth is too high), then it will exceed the “capabilities in the network[] and/or the endpoint[], in which case packets get dropped.” *Id*., 3:35-41. If it sends data too slowly (bandwidth is too low), then “data throughput suffers” and the network has “long upload, download or other communication speeds,” which “are inadequate.” *Id*., 3:41-45. In order to determine the appropriate bandwidth before the ’634 Patent, “various algorithms” had to be used once a network connection was created. *Id*., 1:9-11. When a connection was terminated and re-created, the process to determine the optimal bandwidth had to begin again. *Id*. But performing these algorithmic calculations for each connection came with significant costs, namely in time and computational resources. *Id*., 1:11-23. “Many algorithms induce considerable time loss in order to average out to the effect of bandwidth, and may take many seconds or even minutes to reach full effective bandwidth capacity.” *Id*., 1:11-15. Other algorithms could reach full operational capacity sooner “but are compute intensive and more difficult to implement.” *Id*., 1:15-18. “These time delays and sub-optimal transmission bandwidth have significant effect in situations where the network connection involves set up, transfer of data, then tear down, and is repeated over and over again . . . .” *Id*., at 1:18-23. Network connections thus suffered from the combination of two inherent features: (1) each connection has no memory of previous connections and (2) connections are frequently “set up” and “t[orn] down.” *Id*., 1:6-23, 3:35-45. These features created a unique technological problem.

29. The ’634 Patent solves this problem with a specific networking architecture. In a typical network connection, the system only includes the network endpoints between which the connection runs. ’634 Patent (Ex. 4), 1:6-9. The ’634 Patent’s innovative solution adds a whole new piece to the network

system: a tuner server (also referred to as a "tuning" server) that is distinct from the network endpoints. Crucially, because the tuning server is distinct from the network endpoints, it "has a higher level view of the network traffic" which allows it to "offload[] the calculations" and "incorporate knowledge of the higher level view of the network topology to enhance congestion control over/among the network(s)." *Id.*, 5:5-9. The tuner server's novel position within the network architecture (i.e., a server distinct from the endpoints) means it has access to the necessary data, and can perform the required calculations and algorithms, to allow networks to achieve full bandwidth quickly. *Id.* "[T]he tuning server 302 . . . is better positioned and equipped to have a broader view of network connections than could any single endpoint." *Id.*, 4:44-46.

30. The '634 Patent's novel architecture improves network connection functionality by offloading data and bandwidth calculations to the tuner server, thereby improving network speeds and efficiency. In particular, the endpoints "continuously or periodically communicate relevant bandwidth information and/or statistics to the tuning server 302," keeping the "tuning server 302 up to date and well informed." *Id.*, 4:46-50, 2:61-65, 3:3-5. The tuner server "us[es] past or recent tuning results when the same network connection or another network connection in a similar geographic area . . . is made" and "communicates bandwidth information back to the endpoint that sent connection information to the tuning server, so that the endpoint can use the initial bandwidth when making the next connection to another endpoint." *Id.*, 2:52-3:2. As a result, "an endpoint can shorten the startup time." *Id.*, 2:52-57, 3:50-58 (the invention "improve[s] the rapidity with which the transmitting endpoint 103 achieves the desired average bandwidth . . . by storing and accessing information about past network connections to set the initial transmission bandwidth.").

31. The '634 Patent thus solves a problem inherent in network connections: how to efficiently determine optimal bandwidth for network connections that lack knowledge of past connections. This problem existed because the previous technology was deficient. The '634 Patent solves this problem with a technological solution: a tuner server distinct from the endpoints to match geographic areas of a network connection to determine bandwidth. This technological solution did not exist in the prior art, and represents a concrete, non-generic invention. *See* '634 Patent (Ex. 4), 1:6-29.

32. Each independent claim explicitly recites the ’634 Patent’s technological solution: (1) “a tuner server,” (2) “distinct from the first endpoint device and the further endpoint devices,” (3) at which there is a “determinati[on]” “that a next network connection from the first endpoint device to a second endpoint device matches a geographical area of a past network connection,” where (4) the “next network connection” is “initiat[ed] “based on the determination at the tuner server,” with (5) “a transmission bandwidth based on the parameter values for the past network connection.” *See* ’634 Patent (Ex. 4), 13:65-14:13; 8:51-67; 15:40-16:8.

33. U.S. Patent No. 9,529,767 (“the ’767 Patent”), titled “System And Method For Abstracting SATA And/Or SAS Storage Media Devices Via A Full Duplex Queued Command Interface To Increase Performance, Lower Host Overhead, And Simplify Scaling Storage Media Devices And Systems,” issued on December 27, 2016. XtreamEdge is the current assignee and owner of title of the ’767 Patent and Concurrent Ventures has exclusively licensed all substantial rights to the ’767 Patent. Plaintiffs own all necessary rights, title, and interest in the ’767 Patent to bring this action, including the right to seek damages, including past damages, for any infringement thereof. A true and correct copy of the ’767 Patent is attached hereto as Exhibit 5. The ’767 Patent claims patent-eligible subject matter and is valid and enforceable.

34. The ’767 Patent claims an invention in “the field of storage devices.” ’767 Patent (Ex. 5), 1:13-14. Specifically, it claims a method and system for a unique architecture and protocol for accessing and utilizing “storage media devices via a full duplex queued command interface to increase performance, lower host overhead, and simplify scaling storage media devices and systems.” *Id*. at 2:14-19. Before the invention of the ’767 Patent, due to unique issues with storage media protocols (such as Serial Advanced Technology Attachment, or SATA, and Serial Attached SCSI, or SAS), communications with storage media were “effectively half duplex as full command queuing and out of order responses [we]re not natively supported.” *Id*., 1:21-33. Further, the available prior art methods “only address[] some data transfers, not all commands and typically do[] not support interleaving read and write commands.” *Id*., 1:33-37. Moreover, prior art systems required the host to “support and manage both protocols when accessing a mixture of devices . . . . The host software layers interfacing to [the storage protocols] must support their high protocol overhead.” *Id*., 1:37-42. Crucially, “prior art systems . . . do not provide a

system and method for abstracting . . . storage media devices via a full duplex queued command interface to increase performance, lower host overhead, and simplify scaling storage media devices and systems." *Id.*, 2:14-19.

35. The '767 Patent addresses these problems in the prior art with a two-prong solution. First, the '767 Patent recites a particular architecture: (1) a host; (2) a master controller with (a) a master processing element, (b) a master controller interface to communicate with the host, and (c) at least one master controller control link interface; (3) an edge controller with (a) an edge processing element, (b) an edge controller control link interface to communicate with the master controller via the master controller control link interface, and (c) at least one storage media device interface to communicate with at least one storage media device; and (4) an additional edge controller for which the first edge controller acts as a master controller. The result of this architecture is a fabric of devices that allows the host to retrieve, store, and process data flexibly and efficiently. This innovative architecture was not present in the prior art.

36. Second, the '767 Patent requires that the master and edge controllers communicate via an "abstraction protocol." While the idea of an abstraction protocol was understood, the '767 Patent invented a particular abstraction protocol. The claims require that the abstraction protocol is a "full-duplex protocol supporting full command queuing for the at least one storage media device." '767 Patent (Ex. 5) at claim 1. This specification is, in part, what allows the unique architecture discussed above to work. The master controller can issue commands to the storage media devices indirectly, through the edge controllers, because the abstraction protocol is a full-duplex protocol supporting full command queuing. This ensures that communications can be bilaterial and understood and packets are not dropped, despite the indirect communications.

37. The specification further describes the abstraction protocol. In particular, it explains that it is "performance-centric;" "supports common high-level read and write access to a pool of storage media devices;" "is link-agnostic;" and "may be carried via a multiplicity of direct attach or networked interfaces." '767 Patent (Ex. 5) at Abstract, 4:4-63. These are further specifications for the abstraction protocol that are concrete innovations over the prior art. The prior art protocols did not meet these requirements and it is these requirements that help make the patented invention successful.

38. Of note, the problem addressed by the ’767 Patent is not a generic problem, but one that is rooted in computer technology. In particular, issues related to how a host communicates with different storage media only exists for computerized storage, which can communication via different protocols. Solutions for non-computerized storage are simply inapplicable in this context, here the considerations and requirements are completely different. For example, the claimed architecture requires multiple levels of devices (e.g., a master controller and two edge controllers) because that allows the computerized system to robustly, flexibly, and efficiently access and process stored data. These additional layers make no sense (and indeed might be detrimental) outside of the particular computerized storage context.

IV.
JURISDICTION AND VENUE

39. Plaintiffs incorporate by reference paragraphs 1-38 herein.

40. This civil action arises under the patent laws of the United States, 35 U.S.C. § 1 *et seq.*, including without limitation 35 U.S.C. §§ 271, 281, 283, 284, and 285. This is a patent infringement lawsuit over which this Court has subject matter jurisdiction under, *inter alia*, 28 U.S.C. §§ 1331 and 1338(a).

41. This District has general and specific personal jurisdiction over Defendants because Defendants have committed acts within this District giving rise to this action, including, on information and belief, developing, making, using, marketing, selling, and testing the infringing products; transact and conduct business in this District and the State of California; and transact and conduct business with residents of this District and the State of California.

42. Plaintiffs’ causes of action arise, at least in part, from Defendants’ contacts with and activities in and/or directed at this District and the State of California. Defendants have systematic and continuous business activities in this District. Defendants have committed acts of patent infringement giving rise to this action within this District. In particular, Defendants have infringed and continue to infringe the Asserted Patents within this District and the State of California by making, using, selling, licensing, offering for sale, and/or importing or exporting in, into, or out of this District and elsewhere in the State of California, products and services covered by claims in the Asserted Patents, including without limitation products that, when made or used, practice the claimed methods and systems of the Asserted

Patents. Defendants, directly and through intermediaries, make, use, sell, offer for sale, import, ship, distribute, advertise, promote, and/or otherwise commercialize such infringing products and services in or into this District and the State of California. Defendants regularly conduct and solicit business in, engage in other persistent courses of conduct in, and/or derive substantial revenue from goods and services provided to residents of this District and the State of California.

43. This Court has personal jurisdiction over Defendants pursuant to CAL. CIV. PROC. CODE § 410.10, the California Long Arm Statute. This Court has personal jurisdiction over Defendants at least because of Defendants' continuous and systematic contacts within the State of California and this District. Defendants create products and services that are and have been used, offered for sale, sold, and purchased in the Northern District of California, and Defendants have committed, and continue to commit, acts of infringement in the Northern District of California, have conducted business in the Northern District of California, and/or have engaged in continuous and systematic activities in the Northern District of California.

44. Venue is proper in this District under 28 U.S.C. §§ 1391 and 1400(b).

45. Defendants are doing business, either directly or through respective agents, on an ongoing basis in this District and elsewhere in the United States and have committed acts of infringement in this District. Defendants have a regular and established place of business in this Judicial District, including at its headquarters located at 2788 San Tomas Expressway, Santa Clara, CA 95051. On information and belief, Defendants make, use, sell, test, offer to sell, and/or import infringing products into and/or within this District, including at its Santa Clara headquarters. Defendants maintain a permanent and/or continuing presence within this District at its Santa Clara headquarters, and have the requisite minimum contacts with this District such that this venue is a fair and reasonable one. Upon information and belief, Defendants have transacted and, at the time of the filing of the Complaint, are continuing to transact business within this District.

46. Defendants purposefully direct or control the sale of the Accused Products, including, on information and belief, for sale in California and elsewhere in the United States, and expect and intend that the Accused Products will be so sold in this District. Defendants purposefully place the Accused Products—whether by itself or through subsidiaries, affiliates, or third parties—into an international

supply chain, knowing that the Accused Products will be sold in the United States, including California and in this District. Therefore, Defendants also facilitate the sale of the Accused Products in California. On information and belief, Defendants know that California is a termination point of its established distribution channels.

47. On information and belief, current and former Mellanox and NVIDIA employees with knowledge relevant to the claims in this case are located in this District and within the State of California. For all of these reasons, venue is appropriate and convenient in this District.

V.

DEFENDANTS ARE INTENTIONALLY AND KNOWINGLY USING PLAINTIFFS' TECHNOLOGY

48. Plaintiffs incorporate by reference paragraphs 1-47 herein.

49. On information and belief, Defendants have infringed and continue to infringe one or more claims of the Asserted Patents by, at a minimum, making, using, offering for sale, and selling infringing products and services in the United States and in this District.

50. The Accused Products are all products including or related to Defendants' DPU technology, including without limitation, Defendants' Data Processing Units including the BlueField-2 DPU, the BlueField-2X DPU, the BlueField-3 DPU, the BlueField-3X DPU, the BlueField-4 DPU, the BlueField-3 SuperNIC, BlueField-3 Liquid Cooled DPU, BlueField-4 STX, the H100 CNX, ConnectX, and the A100 EGX, as well as any products incorporating those items (collectively, the "Accused Products"). Each of these products and services practice at least one claim of each Asserted Patent.

51. On information and belief, Defendants have been on notice of the Asserted Patents and their infringement since at least since their issuance, as evidenced the similarities between Defendants' Accused Products and their marketing and the Plaintiffs' patented technology and marketing; and the small nature of the industry, Plaintiffs' pioneering role in that market, and Defendants' direct competition against Plaintiffs.

52. Defendants have had actual knowledge of each of the Asserted Patents, and of the infringing nature of the Accused Products, since no later than September 9, 2024. On that date, Plaintiffs sent Defendants written notice that specifically identified each of the Asserted Patents and informed

Defendants that the Accused Products infringe the Asserted Patents. Exhibit 31, Letter to Nvidia dated September 9, 2024. On October 18, 2024, Plaintiffs sent Defendants charts that demonstrated Defendants' infringement of the Asserted Patents in detail. Exhibits 32-36.

53. Defendants did not seek a license from Plaintiffs and, on information and belief, took no action on or after that date to investigate, avoid, or mitigate their infringement or their customers' infringement of the Asserted Patents in response to Plaintiffs' notice. Instead, on information and belief, Defendants continued without alteration to make, use, sell, offer for sale, and import the Accused Products, and to instruct and encourage their customers to deploy and use the Accused Products in the infringing manner described herein.

54. In the alternative, to the extent Defendants lacked actual knowledge of any Asserted Patent before September 9, 2024, Defendants were willfully blind to the existence of the Asserted Patents and to the infringing nature of the Accused Products. Defendants subjectively believed there was a high probability that the Accused Products practiced patents held by others in the small, specialized market, which Plaintiffs pioneered years before Defendants, in which Plaintiffs publicly marketed their patented SPU technology. On information and belief, Defendants took deliberate actions to avoid learning of Plaintiffs' patents and of their own infringement. On information and belief, NVIDIA conducted extensive intellectual property diligence in connection with its acquisition of Mellanox in 2020, including diligence concerning patents in the DPU field, and Defendants actively monitor patent activity in the fields in which they operate. On information and belief, notwithstanding that diligence and monitoring, Defendants deliberately declined to review or analyze Plaintiffs' patents in order to shield themselves from knowledge of their infringement.

55. At the latest, Defendants have had actual knowledge of each of the Asserted Patents, and of the infringing nature of the Accused Products, since the filing and service of this Complaint, which identifies each Asserted Patent and describes Defendants' infringement on a limitation-by-limitation basis. Defendants' continued manufacture, sale, distribution, and support of the Accused Products, and their continued instruction and encouragement of customers' infringing use of the Accused Products, after the filing and service of this Complaint constitute, at a minimum, knowing post-suit induced infringement and contributory infringement.

56. In addition to Defendants' direct infringement, Defendants' customers and end users—including, on information and belief, cloud service providers, data center operators, server OEMs, system integrators, and enterprise customers—directly infringe the Asserted Patents, literally or under the doctrine of equivalents, when they deploy, configure, provision, and use the Accused Products in the United States in the manner that Defendants instruct, recommend, and intend, including by using BlueField DPUs and Defendants' associated DOCA software to offload and accelerate isolate networking, storage, security, telemetry, and other infrastructure functions as described in this Complaint.

57. Defendants take affirmative, active steps, directed at their customers, to cause, instruct, and encourage the infringing deployment and use of the Accused Products. Among other things, Defendants: (a) author and publish detailed product documentation, programming, deployment guides, and configuration instructions like the DOCA documentation, BlueField DPU management and provisioning documentation, and storage protocol documentation cited in this Complaint which direct customers, step by step, to deploy, configure, and operate the accused functionality; (b) develop, distribute, and promote the DOCA documentation, drivers, and sample code for the purpose of enabling and causing customers to implement the accused functionality; (c) publish developer blog posts, product briefs, and reference architectures instructing and urging customers to offload data center workloads to BlueField DPUs, including the materials cited in this Complaint; (d) advertise and market the accused offload functionality as the principal purpose and benefit of the Accused Products, including through Defendants' statements that a DPU can "perform the work of 30, 100, or even 300 CPU cores" and can reduce power consumption by "up to 30%" by offloading data center functions from CPUs; and (e) on information and belief, provide training, technical support, professional services, and architecture assistance to customers deploying the Accused Products in the infringing configurations.

58. The infringing functionality of the Accused Products is their intended, advertised, recommended, and principal use. The very purpose of a DPU, as Defendants themselves promote, is to offload networking, storage, security, and other infrastructure functions from host processors. Defendants design, document, market, and support the Accused Products precisely so that customers will deploy and use them in the infringing manner. Defendants intend that their instructions and recommendations be

followed and Defendants know that customers in fact follow them. When customers deploy and use the Accused Products as Defendants direct, those customers directly infringe the Asserted Patents.

59. Defendants' infringement of the Asserted Patents before and after the filing of this Complaint has been and continues to be willful in view of the above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

VI.
FIRST CLAIM
Infringement of the '596 Patent

60. Plaintiffs incorporate by reference paragraphs 1-59 herein.

61. On information and belief, Defendants have infringed and continue to infringe the '596 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, offering for sale, and/or importing into the United States products and/or methods that practice at least claim 1 of the '596 Patent, including the Accused Products.

62. The claims of the '596 Patent are directed to non-abstract ideas and otherwise recite inventive concepts and technological improvements. Plaintiffs have complied with the requirements of 35 U.S. Code § 287 with regards to the '596 Patent.

63. Each of the Accused Products meet every limitation of claim 1 of the '596 Patent, which recites:

> 1. A system for dividing and synchronizing a processing task across a plurality of processing elements comprising:
>
> an input queue implemented in hardware;
>
> an output queue implemented in hardware;
>
> a first processing element having access to said input queue and said output queue;
>
> at least one second processing element in communication with said first processing element;
>
> a reservation register implemented in hardware storing a value indicative of available space in said input queue, said reservation register accessible by both said first processing element and said at least one second processing element;
>
> computer storage storing instructions, which when executed by said at least one second processing element:
>
>> accesses said reservation register and reads said stored value;

determines when said read value indicates available space in said input queue for said first processing element to issue a command;

notifies said first processing element to issue said command to said input queue; and

wherein said first processing element receives notification from said at least one second processing element regarding issuing the command, issues said command to said input queue, and receives a response corresponding to said command from said output queue.

64. The Accused Products, including Defendants' DPU technology and DPU-enabled systems, allow for hardware processing of specific functions on the data path with hardware queues supporting various operations implemented in hardware. The Accused Products include systems for dividing and synchronizing a processing task across a plurality of processing elements. For example, the NVIDIA BlueField DPU can "[t]ake over infrastructure tasks from the server CPU so more CPU power can be used to run applications," "[r]un infrastructure functions more quickly than the CPU can, using hardware acceleration in the DPU silicon," and "[m]ove key data plane and control plane functions to a separate domain on the DPU, both to relieve the server CPU from the work and to protect the functions in case the CPU or its software are compromised." Exhibit 28. *See also id.* ("a proper DPU must use hardware acceleration to add maximum benefit to the data center… The DPU is best suited for offloading workloads involving data movement, and security. For example, networking is an ideal task to offload to DPU silicon, along with remote direct memory access (RDMA), used to accelerate data movement between servers for AI, HPC, and big data, and storage workloads. When the DPU has acceleration hardware for specific tasks, it can offload and run those with much higher efficiency than a CPU core. A properly designed DPU can perform the work of 30, 100, or even 300 CPU cores when the workload meets the DPU's hardware acceleration capabilities."). BlueField-3 DPUs, for example, offload, accelerate, and isolate software-defined networking, storage, security, and management functions, significantly enhancing data center performance, efficiency, and security.[20] The NVIDIA BlueField supports what is called the data-path accelerator ("DPA"), which is an embedded subsystem designed to accelerate workloads that require high-performance access to the NIC engines in certain packet and I/O processing workloads. The DPA platform

[20] Exhibit 39, https://resources.nvidia.com/en-us-accelerated-networking-resource-library/datasheet-nvidia-bluefield.

supports multiple processes with each process having multiple threads. Each thread can be mapped to a different execution unit to achieve parallel execution.[21]

65. The Accused Products include an input queue implemented in hardware and an output queue implemented in hardware. For example, the NVIDIA BlueField-3 DPA is an embedded subsystem designed to accelerate workloads that require high-performance access to the NIC engines in certain packet and input/output ("I/O") processing workloads. Exhibit 37. The DPA can send and receive any kind of packet toward the NIC and utilize all the accelerators that reside on the BlueField DPU (e.g., encryption/decryption, hash computation, compression/decompression). *Id.* The queues within the Accused Products, including those associated with this process (e.g., the queues for the Accused Products and for their accelerators and other internal processors), are examples of the claimed queues implemented in hardware.

66. The Accused Products include a first processing element having access to said input queue and said output queue, for example a CPU on a host system having access to said input queue and said output queue. For example, the DPA can send and receive any kind of packet toward the NIC and utilize all the accelerators that reside on the BlueField DPU (e.g., encryption/decryption, hash computation, compression/decompression). Exhibit 37. The first processing elements in the Accused Product, including for example external hosts or other network devices or internal processing elements, such as those associated with the DPA, are examples of the claimed first processing element with access to the claimed queues.

[21] Exhibit 37, https://docs.nvidia.com/doca/sdk/dpa+subsystem/index.html.

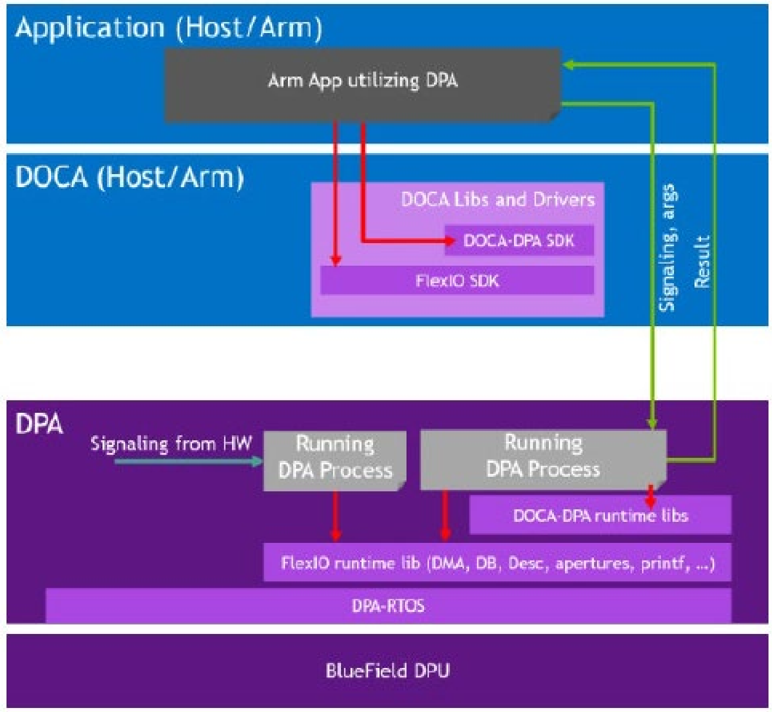


Exhibit 37

67. The Accused Products include at least one second processing element in communication with said first processing element. For example, the DPA process can use device side libraries to access the resources. The provided APIs support signaling of the DPA process from the host or DPU to explicitly pass control or to obtain results from the DPA. Exhibit 37. The second processing element in the Accused Products, including for example the processors associated with this scheduling/traffic management process within and for the Accused Products, are examples of the claimed second processing element.

68. The Accused Products contain a reservation register implemented in hardware storing a value indicative of available space in said input queue, said reservation register accessible by both said first processing element and said at least one second processing element accessible by the first and second processing element. For example, NVIDIA BlueField DPUs supporting the DPA platform also perform something called "FlexIO," which is a "low-level event-driven library to program and accelerate functions on the DPA."[22] A DPA process can initiate send and receive operations using the FlexIO outbox object. *Id.* The FlexIO outbox contains memory-mapped IO registers that enable the DPA application to issue

[22] Exhibit 38, https://docs.nvidia.com/doca/archive/doca-v2.2.0/dpa-subsystem-programming-guide/index.html.

device doorbells to manage the send and receive planes. The DPA outbox can be configured during run time to perform send and receive from a specific NIC function exposed by the DPU. *Id.* Those registers are one of many examples of the Accused Products containing the claimed reservation register implemented in hardware.

69. The Accused Products contain computer storage storing instructions, which when executed by said at least one second processing element, accesses the reservation register and reads the stored value using the scheduler. Exhibit 38. The Accused Products determine when said read value indicates available space in said input queue for said first processing element to issue a command. Exhibit 38. With the Accused Products, the second processing element notifies said first processing element to issue said command to said input queue. Exhibit 38. For example, the DPA outbox can be configured during run time to perform send and receive from a specific NIC function exposed by the DPU. *Id*. In general, the Accused Products' scheduling and traffic management processes include the second processing elements performing the claimed actions by executing computer storage storing instructions. These are examples of how the Accused Products completes the claimed process with computer storage storing instructions.

70. Ultimately, the first processing element receives notification from said at least one second processing element regarding issuing the command, issues said command (via the host interface) to said input queue and receives a response corresponding to said command from said output queue. Exhibit 38.

71. Plaintiffs incorporate by reference Exhibit 35, which further demonstrates Defendants' Accused Products' infringement of claim 1 of the '596 patent.

72. On information and belief, Defendants' infringement through its use of its DPU technology and DPU-enabled systems, described above, is exemplary of its direct infringement with respect to all the Accused Products.

73. Defendants have also indirectly infringed, and continue to indirectly infringe, the '596 Patent under 35 U.S.C. § 271(b) and (c).

74. Defendants knowingly and intentionally actively aided, abetted, and induced others to directly infringe at least claim 1 of the '596 Patent (such as its customers in this District and throughout the United States), and continue to do so, by, for example, selling and offering access to and encouraging and supporting use of the Accused Products. Defendants have had knowledge of the '596 Patent and of

the infringing nature of the Accused Products since no later than September 9, 2024, as alleged above, and, at the latest, since the filing and service of this Complaint. In the alternative, before September 9, 2024, Defendants were willfully blind to the existence of the ’596 Patent and to the infringing nature of the Accused Products, as alleged above.

75. Defendants’ customers and end users—including cloud service providers, data center operators, server OEMs, systems integrators, and enterprise customers—directly infringe at least claim 1 of the ’596 Patent, literally or under the doctrine of equivalents, when they deploy and use the Accused Products in the United States as Defendants instruct and intend, including when they deploy BlueField DPUs in host systems and use the DPA subsystem and FlexIO interfaces to divide and synchronize processing tasks across processing elements in the manner described above with respect to claim 1 of the ’596 Patent.

76. Defendants take affirmative steps, directed at their customers, to cause and encourage the infringing use, including by authoring and publishing the documentation cited above, including the DPA subsystem and FlexIO programming documentation cited above, which instructs customers, step by step, how to create and run DPA processes, map threads to execution units for parallel execution, signal between the host or DPU and the DPA, and configure the FlexIO outbox to issue device doorbells and manage the send and receive planes. Defendants also develop and distribute the DOCA SDK, FlexIO libraries, sample applications, and reference code that implement and enable that functionality and publish developer materials that instruct and urge customers to offload workloads to the DPU’s hardware acceleration. Defendants also market and that offload functionality is the principal benefit of the Accused Products. Defendants engage in these affirmative acts with the specific intent that their customers use the Accused Products in an infringing manner, and with knowledge or willful blindness to the fact that the induced acts constitute infringement of the ’596 Patent. On information and belief, the induced use is the intended, advertised, and recommended use of the Accused Products, and customers in fact deploy and use the Accused Products as Defendants direct.

77. Defendants have also contributed to the direct infringement of at least claim 1 of the ’596 Patent under 35 U.S.C. § 271(c), and continue to do so, by, for example, supplying, with knowledge of the ’596 Patent, a material part of a claimed invention, where the material part is not a staple article of

commerce and is incapable of substantial noninfringing use. For example, Defendants have provided, owned, operated, sold, offered to sell, leased, licensed, used, and/or imported, and continue to do so, various hardware and/or software that make up and enable the Accused Products, including as used in third-party (including customer) systems (including as discussed above), are a material part of the claimed invention, are not a staple article of commerce, and are incapable of substantial non-infringing uses. Defendants sell, offer to sell, and import into the United States components that constitute a material part of the invention claimed in the ’596 Patent, including BlueField DPUs incorporating the DPA subsystem, and the associated FlexIO whose function is to manage the issuance of commands to hardware queues through registers as described above. On information and belief, Defendants know that these components are specially made and adapted for use in infringing the ’596 Patent as the DPA subsystem and the FlexIO components exist to perform the accused queue synchronization functionality, and when deployed as Defendants instruct, there is no substantial use of the DPA subsystem and FlexIO components that does not practice the claimed invention.

78. On and information and belief, despite their knowledge of the ’596 Patent and of their customers’ infringement, Defendants have continued—including after September 9, 2024, and after the filing and service of this Complaint—to make, sell, offer for sale, and import the Accused Products, to distribute the software and documentation described above, and to instruct and encourage the infringing use, without redesigning the Accused Products, modifying their instructions, or taking any other action to avoid or mitigate infringement. Defendants’ induced infringement and contributory infringement of the ’596 Patent are both pre-suit and post-suit, and ongoing.

79. As explained above, Defendants’ infringement has been and continues to be willful in view of the facts asserted above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

VII.
SECOND CLAIM
(Infringement of the ’753 Patent)

80. Plaintiffs incorporate by reference paragraphs 1-79 herein.

81. On information and belief, Defendants have infringed and continue to infringe the ’753 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making,

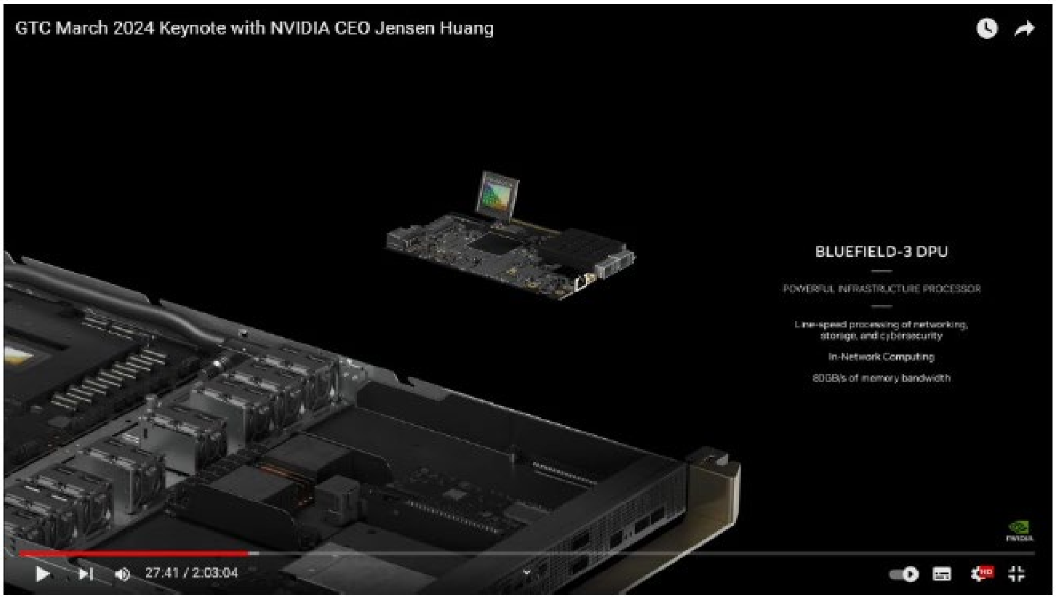
using, selling, offering for sale, and/or importing into the United States products and/or methods that practice at least claim 1 of the ’753 Patent, including the Accused Products.

82. The claims of the ’753 Patent are directed to non-abstract ideas and otherwise recite inventive concepts and technological improvements. Plaintiffs have complied with the requirements of 35 U.S. Code § 287 with regards to the ’753 Patent.

83. Each of the Accused Products meet every limitation of claim 1 of the ’753 Patent, which recites:

> 1. A platform for data flow processing, comprising:
>
> one or more swappable pods or cards in one or more chassis, coupled through a messaging interface network;
>
> each of the one or more swappable pods or cards having one or more hardware modules or one or more software modules;
>
> one or more of the plurality of swappable pods or cards having a portion for user-definable hardware modules or user-definable software modules;
>
> the plurality of swappable pods or cards being user-configurable to implement data flow processing architectures; and
>
> a network coupled to the one or more swappable pods or cards and supporting messaging-based communication using packets each having a header with a chassis identifier, a board identifier, a module identifier, an instance identifier, and a type identifier, so that each type of module, each instance of a type of module, and each module on each board in each chassis can be addressed through the header.

84. The Accused Products include a platform for data flow processing comprising one or more swappable pods or cards in one or more chassis coupled through a messaging interface network. For example, NVIDIA’s DGX GB200 NVL72 server system includes multiple compute trays in the rack, with each tray containing BlueField-3 DPUs. Furthermore, the NVIDIA BlueField DPU is connected to the network and hosts via Ethernet ports or NDR400 InfiniBand for network connectivity. This server system and its compute trays are one example of the claimed platform for data processing, and the BlueField-3 DPUs make up the one or more claimed swappable pods or cards, connected to the network, as outlined in the claim.



https://www.youtube.com/watch?v=Y2F8yisiS6E

The NVIDIA® BlueField®-3 networking platform is designed to accelerate data center infrastructure workloads and usher in the era of accelerated computing and AI. Supporting both Ethernet and InfiniBand connectivity, BlueField-3 offers speeds up to 400 gigabits per second (Gb/s). It combines powerful computing with software-defined hardware accelerators for networking, storage, and cybersecurity—all fully programmable through the NVIDIA DOCA™ software framework. Drawing on the platform's robust capabilities, BlueField data processing units (DPUs) and BlueField SuperNICs revolutionize traditional computing environments, transforming them into secure, high-performance, efficient, and sustainable data centers suitable for any workload at any scale.

Exhibit 39

Network and Host Interfaces

Network Interfaces

> 1 or 2 ports with up to 400Gb/s Ethernet or NDR InfiniBand connectivity

PCI Express Interface

> 32 lanes of PCIe Gen 5.0
> Flexible PCIe switch supporting self-hosting and server-hosting

Id.

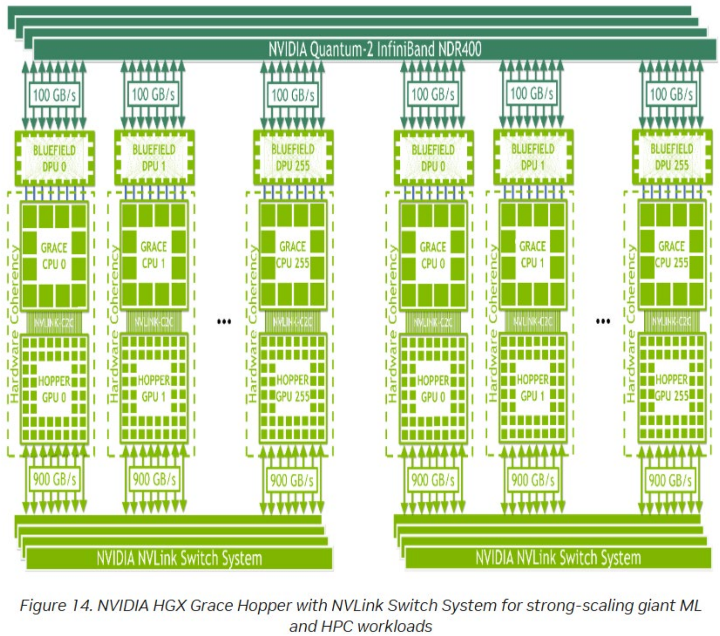


Figure 14. NVIDIA HGX Grace Hopper with NVLink Switch System for strong-scaling giant ML and HPC workloads

Exhibit 40, https://developer.nvidia.com/blog/Nvidia-grace-hopper-superchip-architecture-in-depth/

85. The Accused Products include swappable pods or cards which have hardware modules and software modules. For example, each of the one or more Nvidia Bluefield DPUs has multiple hardware modules or software modules, including embedded ARM cores, hardware accelerator engines to provide programmable security, network, and data path accelerators to offload data centric workloads in the data centers. These are examples of the claimed swappable pods or cards which have the claimed hardware and software modules.

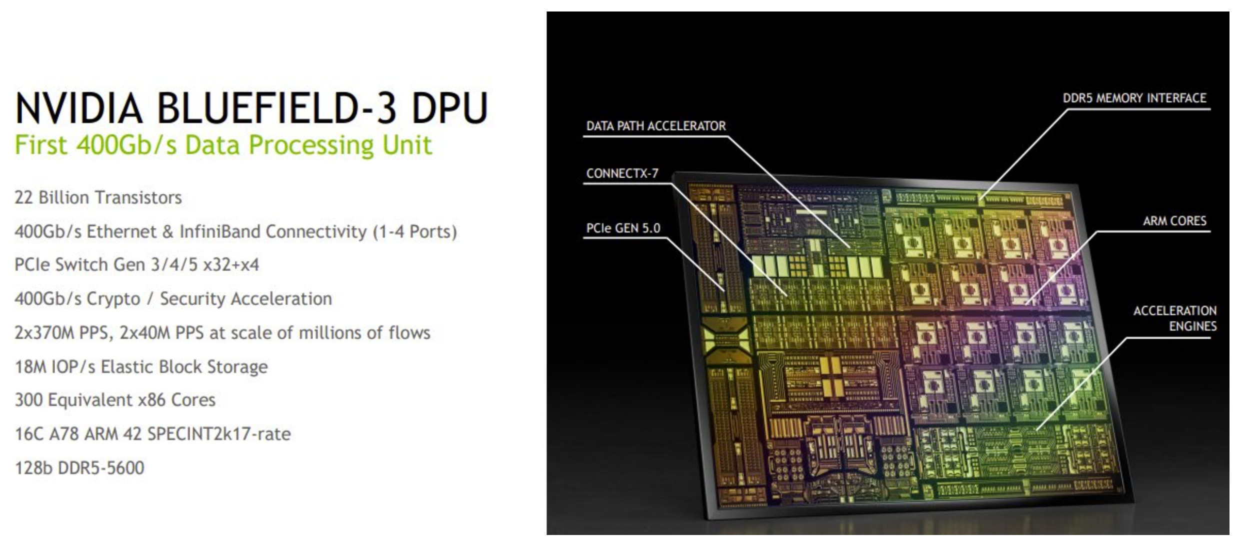
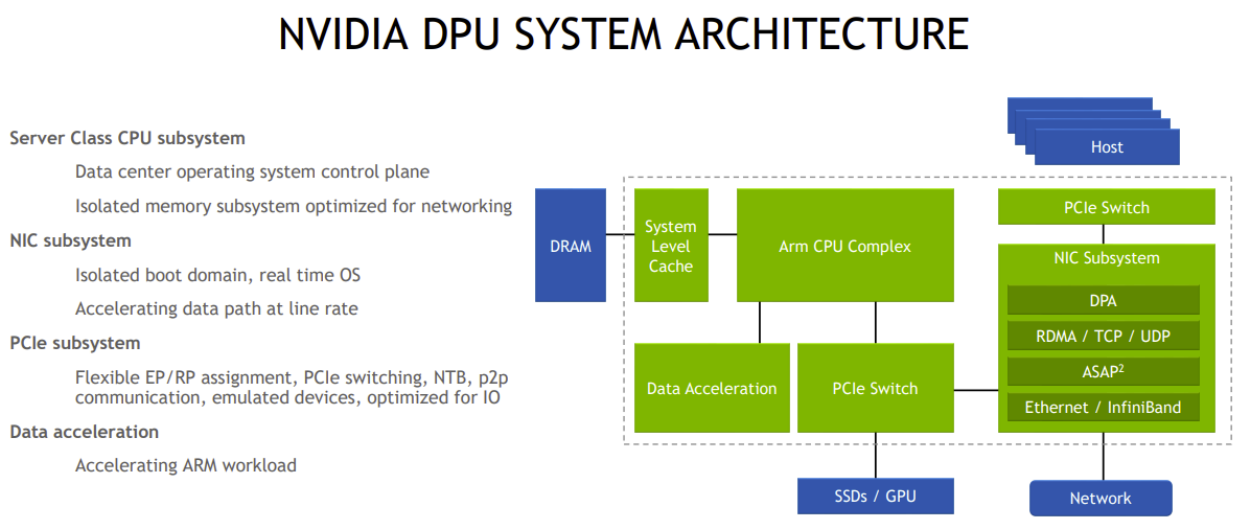


Exhibit 41, https://hc33.hotchips.org/assets/program/conference/day1/HC2021.NVIDIA.IdanBurstein.v08.norecording.pdf



Id.

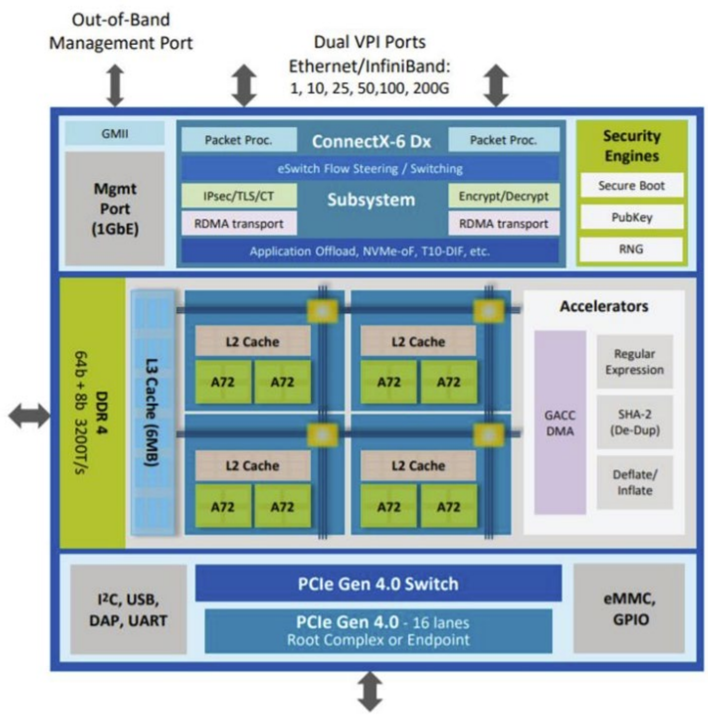


Figure 11: NVIDIA BlueField-2 Block Diagram [Ref. 8]

Exhibit 42, https://learning.dell.com/content/dam/dell-emc/documents/en-us/2021KS_Baligh-DPUs_Acceleration_Through_Disaggregation.pdf

86. The Accused Products include one or more of the plurality of swappable pods or cards that have a portion for user-definable hardware modules or user-definable software modules. For example, the NVIDIA servers containing multiple Bluefield DPUs also contain programmable ARM cores, which serve

as an example of the claimed swappable pods or cards that have a portion for user-defined hardware or software modules.

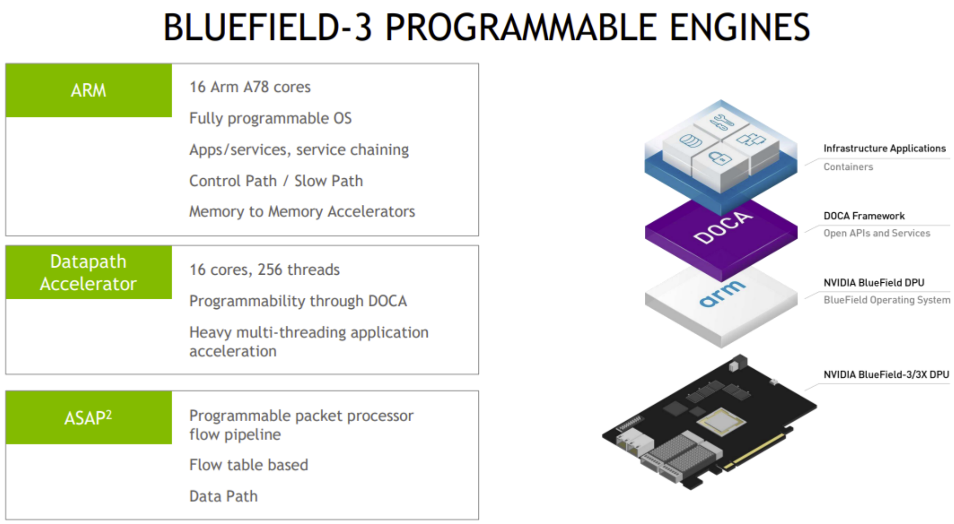


Exhibit 41

NVIDIA® BlueField® supports **ASAP^2 technology**. It utilizes the representors mentioned in the previous section. BlueField SW package includes OVS installation which already supports ASAP^2. The virtual switch running on the Arm cores allows us to pass all the traffic to and from the host functions through the Arm cores while performing all the operations supported by OVS. ASAP^2 allows us to offload the datapath by programming the NIC embedded switch and avoiding the need to pass every packet through the Arm cores. The control plane remains the same as working with standard OVS.

Exhibit 43, https://docs.nvidia.com/networking/display/bluefielddpubspv420/virtual+switch+on+dpu

Mellanox ASAP²

- SDN acceleration for:
  - Bare metal
  - Virtualization
  - Containers
- Full hardware offload for OVS data plane
- Flow update through RTE_Flow or TC_Flower
- OpenStack support
- Kubernetes support
- Rich classification engine (L2 to L4)
- Flex-Parser: user defined classification
- Hardware offload for:
  - Connection tracking (L4 firewall)
  - NAT
  - Header rewrite
  - Mirroring
  - Sampling
  - Flow aging
  - Hierarchial QoS
  - Flow-based statistics

Exhibit 44, https://network.nvidia.com/sites/default/files/doc-2020/pb-connectx-6-dx-en-dellemc.pdf

For another example, the NVIDIA Bluefield DPUs include user-definable hardware or software modules that are programmable using P4 pipelines to accelerate and customize cloud infrastructure services.

Compute and Memory

Arm CPU Cores

- Up to 16 Armv8.2+ A78 Hercules cores (64-bit)
- 8MB L2 cache
- 16MB LLC system cache

Programmable Datapath Accelerator

- 16 cores, 256 threads
- Programmability through DOCA
- Heavy multi-threading applications acceleration

Exhibit 45, https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-3-dpu.pdf

P4 is another language that can take advantage of the flexible, programmable, data path accelerators in the BlueField DPUs. P4 support is a component of DOCA enabling support for future VNFs that may be developed for P4. This support is integrated with other APIs that already have a rich ecosystem of VNF offerings. This architecture allows the simultaneous execution of programs written for Kernel, RoCE, DPDK, SPDK, P4, and P4.runtime interfaces, which can seamlessly coexist and take advantage of DPU data-path acceleration.

Exhibit 46, https://developer.nvidia.com/blog/programming-the-entire-data-center-infrastructure-with-the-Nvidia-doca-sdk/

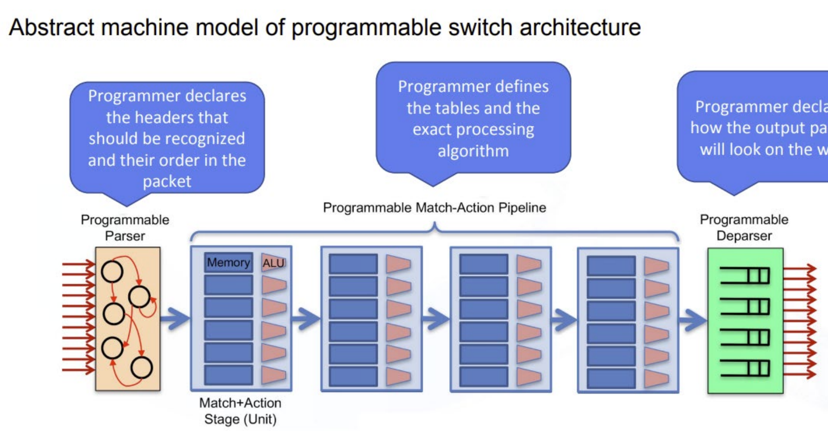


Exhibit 47, https://events19.linuxfoundation.org/wp-content/uploads/2017/12/Tutorial-P4-and-P4Runtime-Technical-Introduction-and-Use-Cases-for-Service-Providers-Carmelo-Cascone-Open-Networking-Foundation.pdf

87. The Accused Products are designed to have portions that are user-configurable to implement data flow processing architectures. For example, NVIDIA servers containing the NVIDIA Bluefield DPUs support fully programmable and accelerated offload management services. For example, programmable dataflow pipelines and customized processing logic for each data flow allowing a user to define a dataflow path or directing packets to specific actions or destinations making DPUs user-configurable, with user-definable hardware or software modules to implement data flow processing architectures. This is an example of the Accused Products containing portions that are user-configurable and are to implement data flow processing architectures, as the claim outlines.

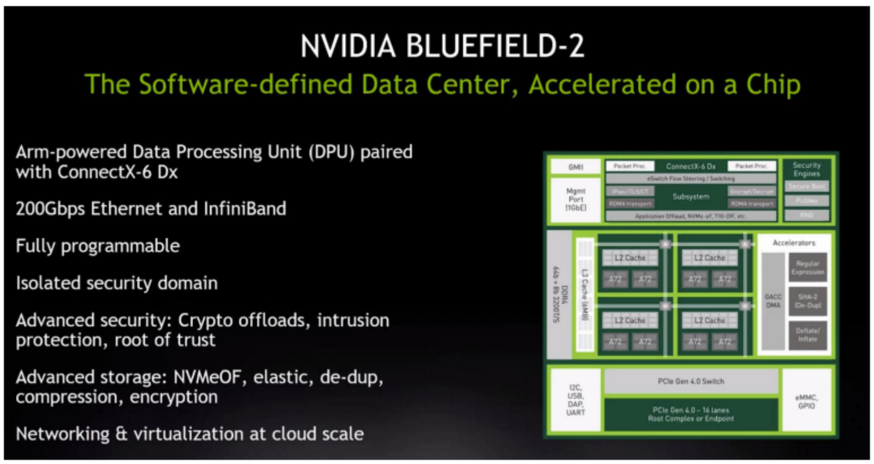


Figure 7: NVIDIA BlueField-2 DPU components & block diagram [Ref. 15]

Exhibit 42

NVIDIA® BlueField® supports **ASAP^2 technology**. It utilizes the representors mentioned in the previous section. BlueField SW package includes OVS installation which already supports ASAP^2. The virtual switch running on the Arm cores allows us to pass all the traffic to and from the host functions through the Arm cores while performing all the operations supported by OVS. ASAP^2 allows us to offload the datapath by programming the NIC embedded switch and avoiding the need to pass every packet through the Arm cores. The control plane remains the same as working with standard OVS.

Exhibit 43

P4 is another language that can take advantage of the flexible, programmable, data path accelerators in the BlueField DPUs. P4 support is a component of DOCA enabling support for future VNFs that may be developed for P4. This support is integrated with other APIs that already have a rich ecosystem of VNF offerings. This architecture allows the simultaneous execution of programs written for Kernel, RoCE, DPDK, SPDK, P4, and P4.runtime interfaces, which can seamlessly coexist and take advantage of DPU data-path acceleration.

Exhibit 46

Abstract machine model of programmable switch architecture

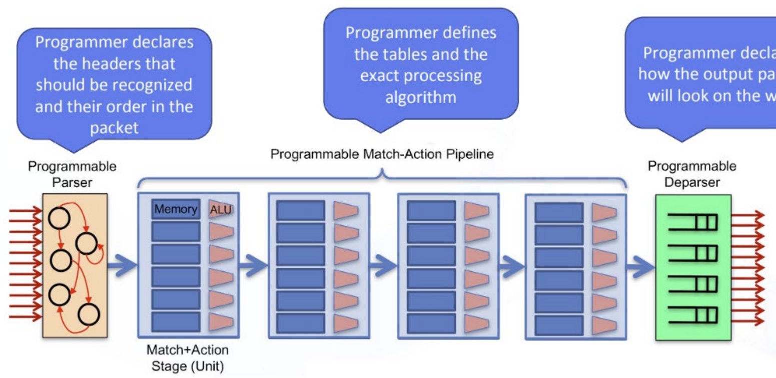


Exhibit 47

Pipes may be chained together through forwarding actions from one pipe to another to form a complete steering tree that defines paths for incoming packets. After executing a predefined action on a packet, the packet may be forwarded to another pipe for further action or inspection, to a software queue, to a hardware hairpin queue, or sent down the wire or dropped.

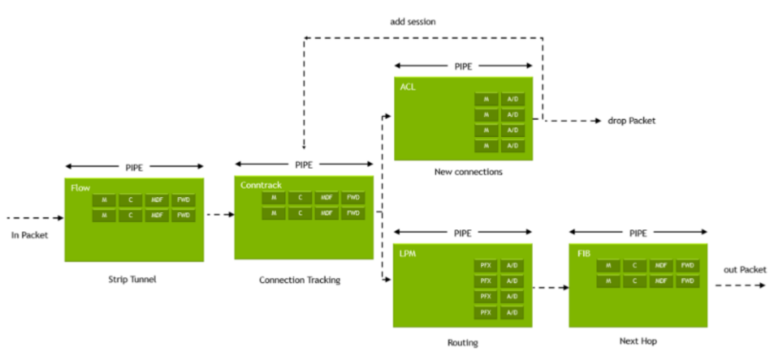


Figure 1. Pipe chaining constructs network logic, as each pipe handles unique criteria, actions, and forwarding, shaping a steering tree for packet control

Exhibit 48, https://developer.nvidia.com/blog/harness-dpu-accelerated-packet-steering-logic-with-Nvidia-doca-flow/

Steering trees

A steering tree can be used to create hardware-based network applications on a DPU or NIC by implementing common network function logic. This enables packets to be effectively categorized so the appropriate action can be applied to each one. Using the steering tree concept provides multiple benefits, including:

- Customized processing logic for each data flow
- Versatility in directing packets to specific actions or destinations
- Adaptable structure that's easily resized for changing conditions
- Flexible framework that allows adding new pipe types to meet evolving requirements
- Optimized resource use, minimizing redundancy and enabling shared matches and actions

Exhibit 48

88. The Accused Products include a network coupled to the one or more swappable pods or cards and supporting messaging-based communication using packets each having a header with a chassis identifier, a board identifier, a module identifier, an instance identifier, and a type identifier, so that each type of module, each instance of a type of module, and each module on each board in each chassis can be addressed through the header. For example, the NVIDIA BlueField DPUs can be deployed in clusters within data centers and allow header information such that the type of module, instance of a type of module, and module on each board in each chassis can be addressed. These DPUs typically use the LLDP protocol to discover neighboring devices on the network. LLDP advertisements include identifiers such as chassis ID and port ID, which help in mapping the network topology. Further, based on the packet flow, Nvidia BlueField DPU supports communication using packets each having a header with a chassis identifier, a board identifier, a module identifier, an instance identifier, and a type identifier, etc.

Exhibit 49, https://developer.nvidia.com/blog/Nvidia-gb200-nvl72-delivers-trillion-parameter-llm-training-and-real-time-inference

Link Layer Discovery Protocol (LLDP)

The Link Layer Discovery Protocol (LLDP) is a vendor-neutral Link Layer protocol in the Internet Protocol Suite used by network devices for advertising their identity, capabilities, and neighbors on a IEEE 802 LAN. The protocol is formally defined in IEEE 802.1AB. From version 3.6.2000, LLDP is now enabled by default.

Exhibit 50, https://networking-docs.nvidia.com/onyxum/3104006/link-layer-discovery-protocol-lldp

3. Display LLDP local information.

```
switch (config) # show lldp local

LLDP is Enabled

Local global configuration
Chassis sub type: macAddress (4)
Chassis id: 00:11:22:33:44:55
System Name: "switch-111111"
System Description: my-system-description
Supported capabilities: B
Supported capabilities enabled: B
```

4. Display LLDP remote information.

```
switch (config)# show lldp interfaces ethernet 1/1 remote

Ethernet 1/1
Remote Index: 1
Remote chassis id: 00:11:22:33:44:55 ; chassis id subtype: mac
Remote port-id: ethenret 1/2; port id subtype: local
Remote port description: ethernet 1/2
Remote system name: remote-system
Remote system description: remote-system-description
Remote system capabilities supported: B ; B
```

Id.

89. Plaintiffs incorporate by reference Exhibit 32, which further demonstrates Defendants' Accused Products' infringement of claim 1 of the '753 patent.

90. On information and belief, Defendants' infringement through its use of its DPU technology and DPU-enabled systems, described above, is exemplary of its direct infringement with respect to all the Accused Products.

91. Defendants have also indirectly infringed, and continue to indirectly infringe, the '753 Patent under 35 U.S.C. § 271(b) and (c).

92. Defendants knowingly and intentionally actively aided, abetted, and induced others to directly infringe at least claim 1 of the '753 Patent (such as its customers in this District and throughout the United States), and continue to do so, by, for example, selling and offering access to and encouraging and supporting use of the Accused Products. Defendants have had knowledge of the '753 Patent and of the infringing nature of the Accused Products since no later than September 9, 2024, as alleged above, and, at the latest, since the filing and service of this Complaint. In the alternative, before September 9, 2024, Defendants were willfully blind to the existence of the '753 Patent and to the infringing nature of the Accused Products, as alleged above.

93. Defendants' customers and end users—including cloud service providers, data center operators, server OEMs, systems integrators, and enterprise customers—directly infringe at least claim 1 of the '753 Patent, literally or under the doctrine of equivalents, when they deploy and use the Accused Products in the United States as Defendants instruct and intend, including when they deploy BlueField DPUs in multi-node server systems and racks, connect them through messaging interface networks, and program and configure them to implement data flow processing architectures in the manner described above with respect to claim 1 of the '753 Patent.

94. Defendants take affirmative steps, directed at their customers, to cause and encourage the infringing use, including by authoring and publishing the documentation and developer materials cited above that instruct customers how to deploy BlueField DPUs in their server systems, program the DPUs user-definable modules using DOCA, and configure data flow processing architectures. Defendants also cause and encourage the infringing use by developing and distributing the DOCA SDK, DOCA Flow,

reference applications, and sample code that implement and enable that functionality, and by marketing multi-DPU, rack-scale deployments as the intended use of the Accused Products. Defendants engage in these affirmative acts with the specific intent that their customers use the Accused Products in the infringing manner, and with knowledge or willful blindness to the fact that the induced acts constitute infringement of the ’753 Patent. On information and belief, the induced use is the intended, advertised, and recommended use of the Accused Products, and customers in fact deploy and use the Accused Products as Defendants direct.

95. Defendants contributed to the direct infringement of at least claim 1 of the ’753 Patent under 35 U.S.C. § 271(c), and continue to do so, by, for example, supplying, with knowledge of the ’753 Patent, a material part of a claimed invention, where the material part is not a staple article of commerce and is incapable of substantial noninfringing use. For example, Defendants have provided, owned, operated, sold, offered to sell, leased, licensed, used, and/or imported, and continue to do so, various hardware and/or software that make up and enable the Accused Products, including as used in third-party (including customer) systems (including as discussed above), are a material part of the claimed invention, are not a staple article of commerce, and are incapable of substantial non-infringing uses. Defendants sell, offer to sell, and import into the United States components that constitute a material part of the invention claimed in the ’753 Patent, including BlueField DPU cards and the associated DOCA software whose function is to provide the user-definable, user-configurable data flow processing modules and messaging-based communication described above. On information and belief, Defendants know that these components are specially made and adapted for use in an infringement of the ’753 Patent, and when deployed as Defendants instruct, there is no substantial use of that software that does not practice the claimed invention.

96. On and information and belief, despite their knowledge of the ’753 Patent and of their customers’ infringement, Defendants have continued—including after September 9, 2024, and after the filing and service of this Complaint—to make, sell, offer for sale, and import the Accused Products, to distribute the software and documentation described above, and to instruct and encourage the infringing use, without redesigning the Accused Products, modifying their instructions, or taking any other action to

avoid or mitigate infringement. Defendants' induced infringement and contributory infringement of the '753 Patent are both pre-suit and post-suit, and ongoing.

97. As explained above, Defendants' infringement has been and continues to be willful in view of the facts asserted above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

VIII.

<u>THIRD CLAIM</u>
(Infringement of the '943 Patent)

98. Plaintiffs incorporate by reference paragraphs 1-97 herein.

99. On information and belief, Defendants have infringed and continue to infringe the '943 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, offering for sale, and/or importing into the United States products and/or methods that practice at least claim 1 of the '943 Patent, including the Accused Products.

100. The claims of the '943 Patent are directed to non-abstract ideas and otherwise recite inventive concepts and technological improvements. Plaintiffs have complied with the requirements of 35 U.S. Code § 287 with regards to the '943 Patent.

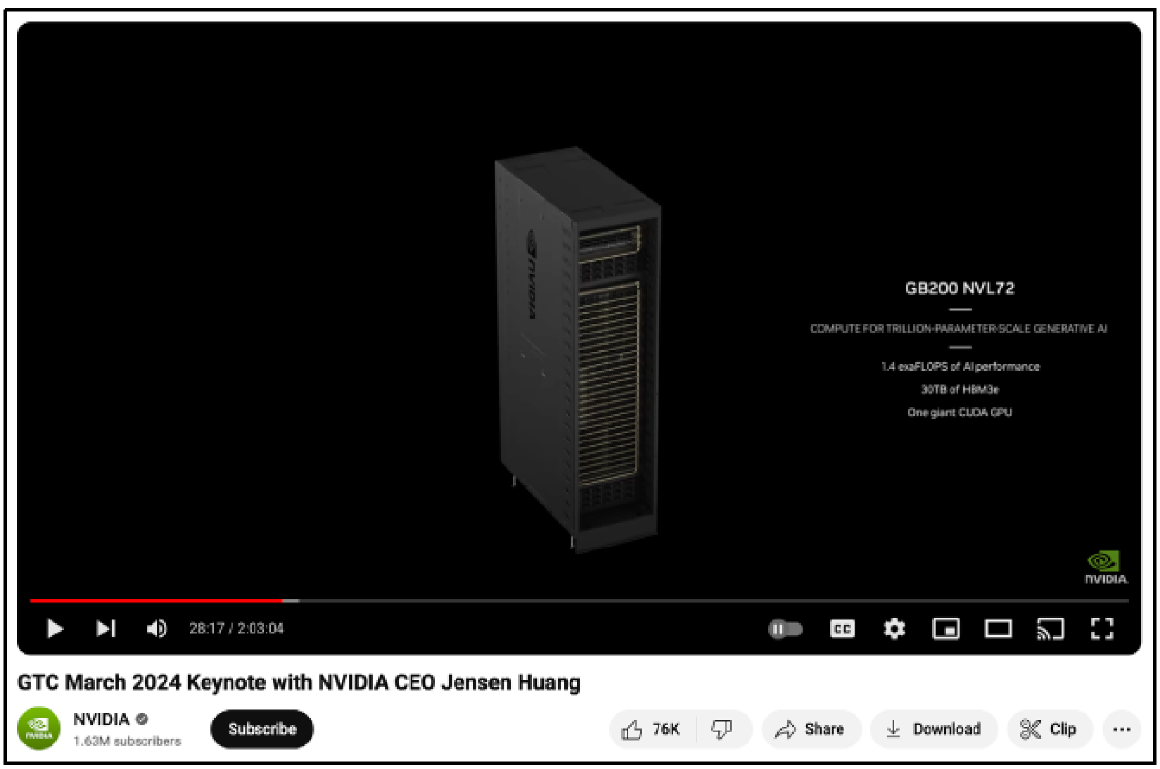
101. Each of the Accused Products meet every limitation of claim 1 of the '943 Patent, which recites:

> 1. A programmable device for data flow processing in a user-configurable server with swappable pods or cards, comprising:
>
> a programmable logic device (PLD) having a first region and a second region;
>
> the first region comprising a hardware-based or firmware-based router with a port enabling communication between the router and the second region; and the second region comprising one or more sandboxes with user-definable programmable electronic circuits of the PLD,
>
> wherein the port comprises: a bridge being lockable to prevent user access, and unlockable to enable user access, to each of the sandboxes and the user-definable programmable electronic circuits therein.

102. The Accused Products include a programmable device for data flow processing in a user-configurable server with swappable pods or cards. For example, NVIDIA's DGX GB200 NVL72 server

system includes a plurality of servers (e.g., multiple NVIDIA GB200 NVL72 compute nodes which include NVIDIA Bluefield-3 DPUs). These servers are examples of the Accused Products including the claimed programmable device with a user-configurable server, as required by the claim.

Exhibit 49



https://www.youtube.com/watch?v=Y2F8yisiS6E

NVIDIA GB200 NVL72 Platforms Arrive Equipped With NVIDIA BlueField-3 DPUs

Flagship, rack-scale platforms powered by the Grace Blackwell accelerator and NVIDIA BlueField-3 networking innovations enable breakthroughs in data processing for the most compute-intensive workloads.

Exhibit 51, https://www.nvidia.com/en-us/networking/products/data-processing-unit/

The GB200 is a key component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink. Additionally, GB200 NVL72 includes NVIDIA BlueField®-3 data processing units to enable cloud network acceleration, composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds. The GB200 NVL72 provides up to a 30x performance increase compared to the same number of NVIDIA H100 Tensor Core GPUs for LLM inference workloads, and reduces cost and energy consumption by up to 25x.

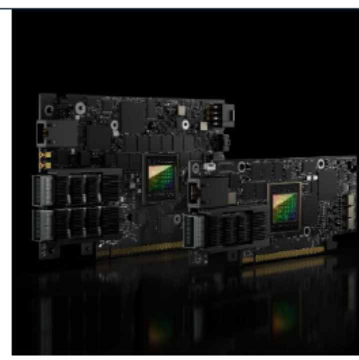
Exhibit 52, https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

Datasheet

NVIDIA.

NVIDIA BlueField-3 Networking Platform

The 400Gb/s infrastructure compute platform for powering the world's data centers.



The NVIDIA® BlueField®-3 networking platform is designed to accelerate data center infrastructure workloads and usher in the era of accelerated computing and AI. Supporting both Ethernet and InfiniBand connectivity, BlueField-3 offers speeds up to 400 gigabits per second (Gb/s). It combines powerful computing with software-defined hardware accelerators for networking, storage, and cybersecurity—all fully programmable through the NVIDIA DOCA™ software framework. Drawing on the platform's robust capabilities, BlueField data processing units (DPUs) and BlueField SuperNICs revolutionize traditional computing environments, transforming them into secure, high-performance, efficient, and sustainable data centers suitable for any workload at any scale.

The **BlueField-3 DPU** is a cloud infrastructure processor that empowers organizations to build software-defined, hardware-accelerated data centers from the cloud to the edge. BlueField-3 DPUs offload, accelerate, and isolate software-defined networking, storage, security, and management functions, significantly enhancing data center performance, efficiency, and security. By decoupling data center infrastructure from business applications, BlueField-3 creates a secure, zero-trust data center infrastructure, streamlines operations, and reduces the total cost of ownership.

Portfolio

- 1 or 2 ports with up to 400Gb/s connectivity
- 32GB on-board DDR5 memory
- Form factors: HHHL, FHHL
- 1GbE out-of-band management port
- Integrated BMC

Exhibit 39

103. The Accused Products contain a programmable logic device ("PLD") having a first region and a second region. For example, NVIDIA Bluefield DPUs are programmable logic devices, each of which supports a broad set of accelerated software-defined networking, storage, security, and management services that are fully programmable and allow for a user definable data path pipelines between its hardware modules for accelerated dataflow path within the DPU. Moreover, as shown below, the Bluefield DPU has various regions, such as regions for ARM cores, acceleration engines, memory interface, data path accelerators, etc. These are examples of the claimed first and second region.

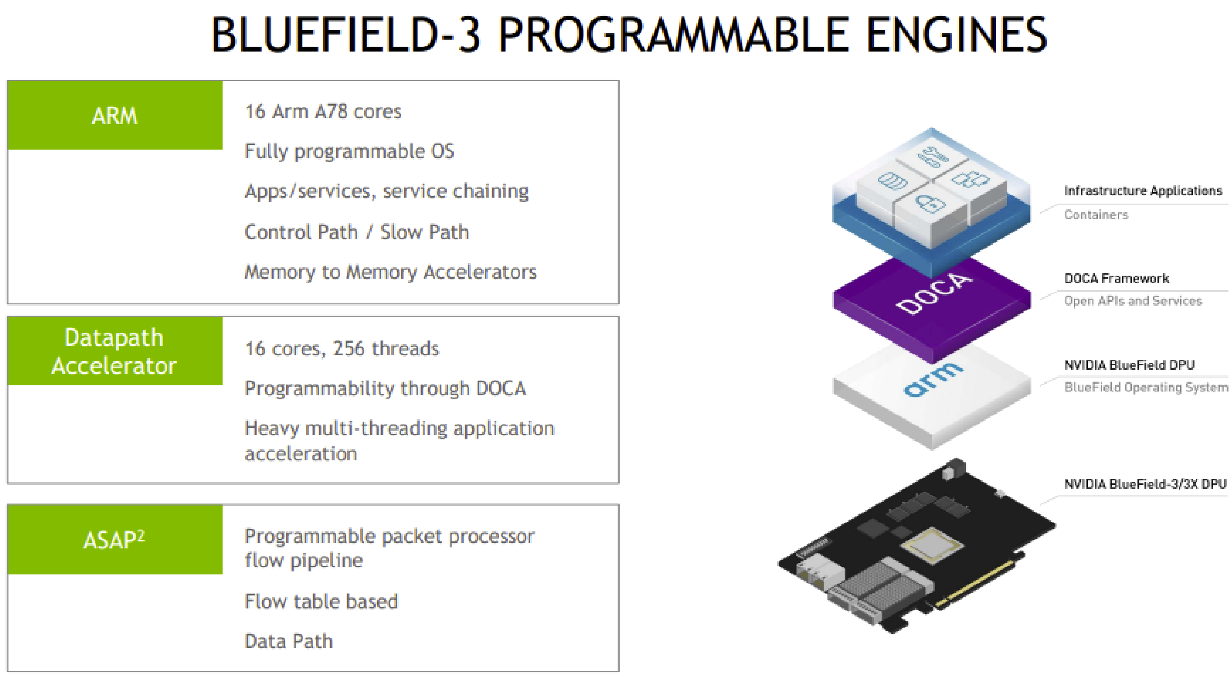


Exhibit 41

NVIDIA BLUEFIELD-3 DPU
First 400Gb/s Data Processing Unit

22 Billion Transistors
400Gb/s Ethernet & InfiniBand Connectivity (1-4 Ports)
PCIe Switch Gen 3/4/5 x32+x4
400Gb/s Crypto / Security Acceleration
2x370M PPS, 2x40M PPS at scale of millions of flows
18M IOP/s Elastic Block Storage
300 Equivalent x86 Cores
16C A78 ARM 42 SPECINT2k17-rate
128b DDR5-5600

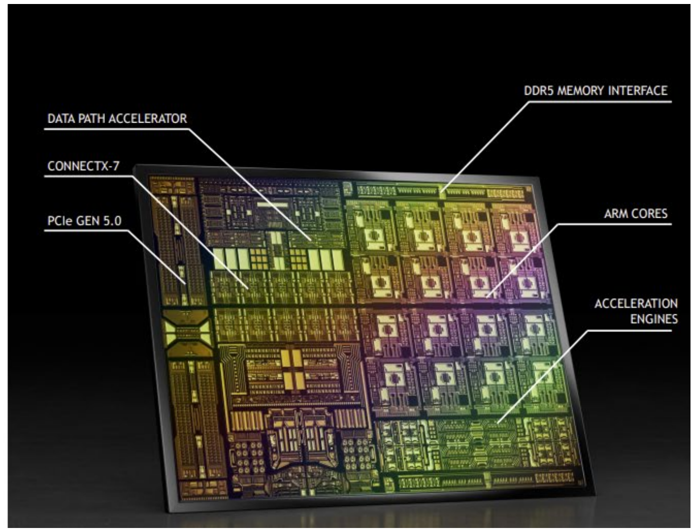


Exhibit 41

NVIDIA DPU SYSTEM ARCHITECTURE

Server Class CPU subsystem
Data center operating system control plane
Isolated memory subsystem optimized for networking
NIC subsystem
Isolated boot domain, real time OS
Accelerating data path at line rate
PCIe subsystem
Flexible EP/RP assignment, PCIe switching, NTB, p2p communication, emulated devices, optimized for IO
Data acceleration
Accelerating ARM workload

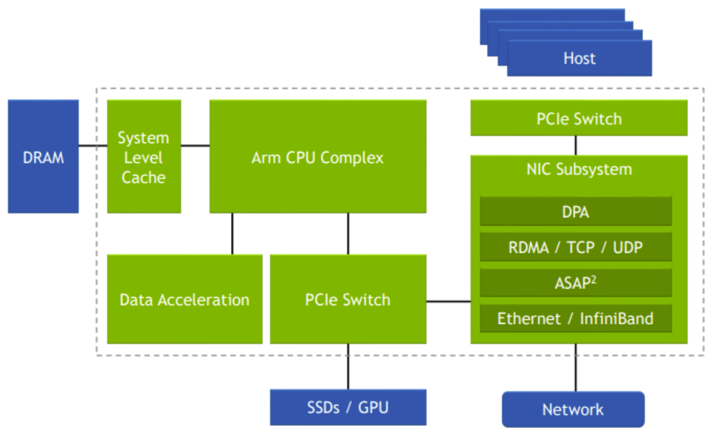


Exhibit 41

NVIDIA® BlueField® supports **ASAP^2 technology**. It utilizes the representors mentioned in the previous section. BlueField SW package includes OVS installation which already supports ASAP^2. The virtual switch running on the Arm cores allows us to pass all the traffic to and from the host functions through the Arm cores while performing all the operations supported by OVS. ASAP^2 allows us to offload the datapath by programming the NIC embedded switch and avoiding the need to pass every packet through the Arm cores. The control plane remains the same as working with standard OVS.

Exhibit 43

Mellanox ASAP²

- SDN acceleration for:
  - Bare metal
  - Virtualization
  - Containers
- Full hardware offload for OVS data plane
- Flow update through RTE_Flow or TC_Flower
- OpenStack support
- Kubernetes support
- Rich classification engine (L2 to L4)
- Flex-Parser: user defined classification
- Hardware offload for:
  - Connection tracking (L4 firewall)
  - NAT
  - Header rewrite
  - Mirroring
  - Sampling
  - Flow aging
  - Hierarchial QoS
  - Flow-based statistics

Exhibit 44

For another example, the multiple Nvidia Bluefield DPUs include user-definable hardware or software modules that are programmable using P4 pipelines to accelerate and customize cloud infrastructure services. Exhibit 45; Exhibit 46; Exhibit 47.

104. The Accused Products contain a first region comprising a hardware-based or firmware-based router with a port enabling communication between the router and the second region. For example, the first region of the Accused Products has hardware-based or firmware-based modules or routers with at least a port that enable communication between the router and various other regions of the Bluefield DPU (e.g., including the P4 pipeline), including the regions having the embedded ARM cores, hardware accelerator engines to provide programmable security, network, and data path accelerators to offload data centric workloads in the data centers, etc.

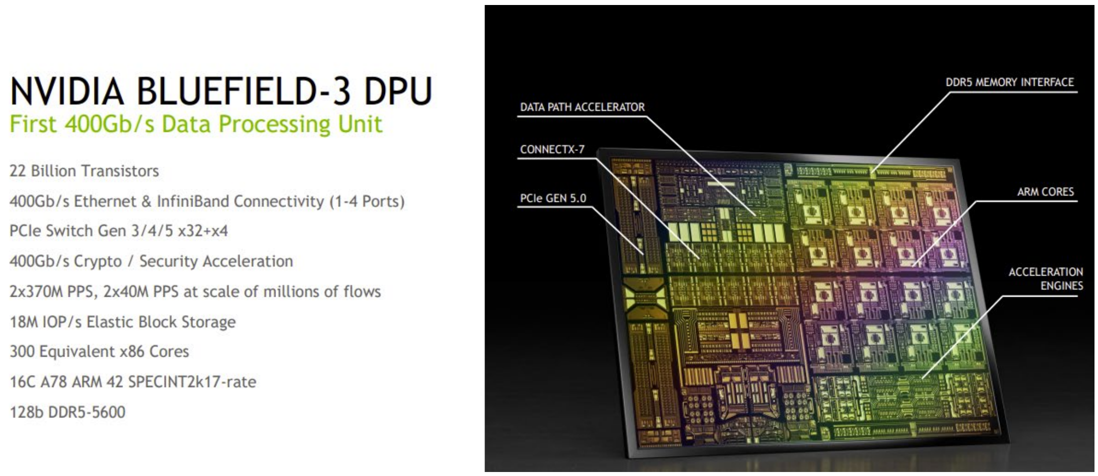
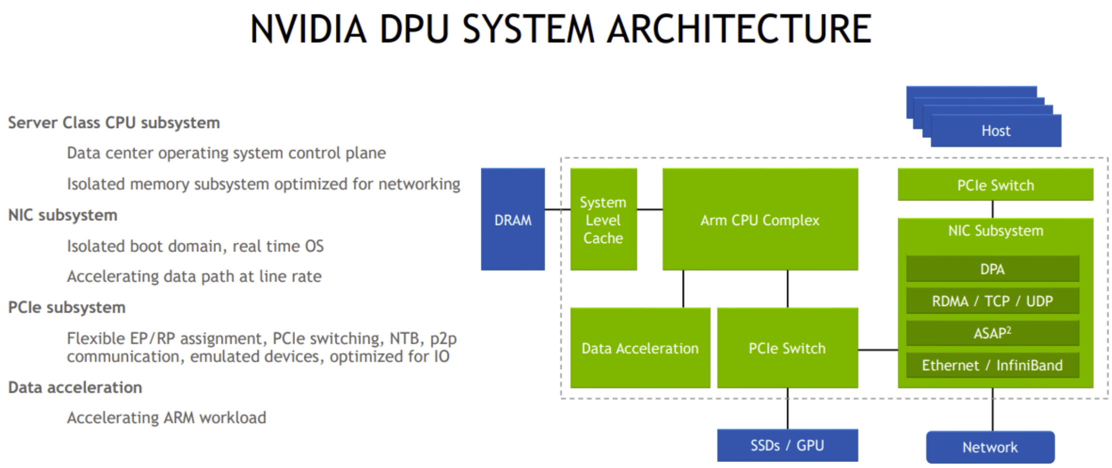


Exhibit 41



Id.

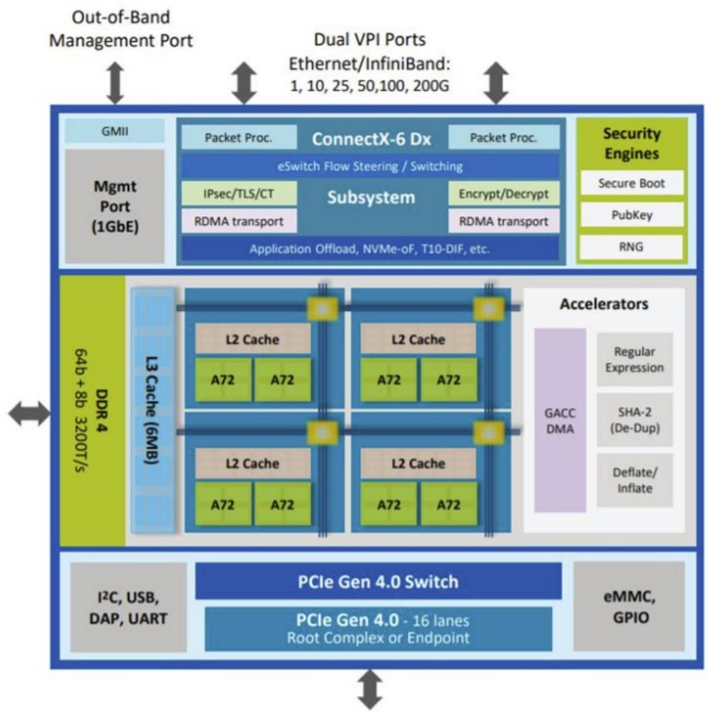


Figure 11: NVIDIA BlueField-2 Block Diagram [Ref. 8]

Exhibit 42

105. The second region of the Accused Products comprise one or more sandboxes with user-definable programmable electronic circuits of the PLD. For example, the Accused Products have one or more sandboxes with user-definable programmable electronic circuits of the PLD. For example, BlueField DPUs have various regions that include one or more sandboxes with user-definable programmable circuits of the DPU, which provide secure environments in isolation for different types of applications. Moreover, for example, the BlueField DPU is configured such that access to the sandbox(es) can be allowed or disallowed.[23]

106. The port of the Accused Products comprises: a bridge being lockable to prevent user access, and unlockable to enable user access, to each of the sandboxes and the user-definable programmable electronic circuits therein. For example, "BlueField DPU contains a programmable CPU based on Arm cores, a state-of-the-art NVIDIA® ConnectX®, and an enhanced set of security, storage, and networking accelerators that can be configured to perform multiple software-defined, hardware-accelerated functions.

---

[23] Exhibit 53, https://docs.nvidia.com/networking/display/initialdpumgmtandprovisioning/introduction#:%7E:text=The%20BlueField%20DPU%20contains%20a,isolated%20environment%20within%20the%20DPU.

With a BlueField DPU, a software-defined network, and/or software-defined storage solution can be deployed and offloaded from the main host CPU in the server. Similarly, other dedicated services (e.g., distributed firewall, deep packet inspection, malware detection) can run on the BlueField DPU and can be accelerated with zero CPU overheads. The BlueField DPU resembles a server embedded within the server itself, creating a secure environment where an infrastructure stack can operate independently from the primary (i.e., host) CPU, effectively isolating it from the untrusted tenant applications. This is the recommended mode for utilizing the DPU in which software running on the host CPU has no direct access to the DPU. For instance, in a scenario where a cloud service provider is responsible for managing both networking and storage in a cloud infrastructure stack, it can establish an isolated environment within the DPU." Exhibit 53. This is an example of a port on the Accused Product comprising the claimed bridge to each of the claimed sandboxes and circuits.

107. Plaintiffs incorporate by reference Exhibit 34, which further demonstrates Defendants' Accused Products' infringement of claim 1 of the '943 patent.

108. On information and belief, Defendants' infringement through its use of its DPU technology and DPU-enabled systems, described above, is exemplary of its direct infringement with respect to all the Accused Products.

109. Defendants have also indirectly infringed, and continue to indirectly infringe, the '943 Patent under 35 U.S.C. § 271(b) and (c).

110. Defendants have had knowledge of the '943 Patent and of the infringing nature of the Accused Products since no later than September 9, 2024, as alleged above, and, at the latest, since the filing and service of this Complaint. In the alternative, before September 9, 2024, Defendants were willfully blind to the existence of the '943 Patent and to the infringing nature of the Accused Products, as alleged above

111. Defendants knowingly and intentionally actively aided, abetted, and induced others to directly infringe at least claim 1 of the '943 Patent (such as its customers in this District and throughout the United States), and continue to do so, by, for example, selling and offering access to and encouraging and supporting use of the Accused Products. Defendants' customers and end users—including cloud service providers, data center operators, server OEMs, systems integrators, and enterprise customers—

directly infringe at least claim 1 of the ’943 Patent, literally or under the doctrine of equivalents, when they deploy and use the Accused Products in the United States as Defendants instruct and intend, including when they deploy BlueField DPUs in user-configurable servers and configure the DPU locking functionality that enables and disables host access to the DPU’s isolated environments in the manner described above with respect to claim 1 of the ’943 Patent.

112. Defendants take affirmative steps, directed at their customers, to cause and encourage the infringing use, including by authoring and publishing the BlueField DPU management and provisioning documentation cited above, which instructs customers how to provision BlueField DPUs, establish isolated environments within the DPU, and enable or restrict host access to those environments, and which affirmatively recommends the accused configuration, stating that operating the DPU so that “software running on the host CPU has no direct access to the DPU” is “the recommended mode for utilizing the DPU.” Defendants also cause and encourage the infringing use by developing and distributing the firmware and software that implement and enable that functionality, and by marketing the DPU isolation and security functionality as a principal benefit of the Accused Products. Defendants engage in these affirmative acts with the specific intent that their customers use the Accused Products in the infringing manner, and with knowledge or willful blindness to the fact that the induced acts constitute infringement of the ’943 Patent. On information and belief, Defendants themselves recommend, and customers in fact deploy and use the Accused Products, as Defendants direct.

113. Defendants contributed to the direct infringement of at least claim 1 of the ’943 Patent under 35 U.S.C. § 271(c), and continue to do so, by, for example, supplying, with knowledge of the ’943 Patent, a material part of a claimed invention, where the material part is not a staple article of commerce and is incapable of substantial noninfringing use. For example, Defendants have provided, owned, operated, sold, offered to sell, leased, licensed, used, and/or imported, and continue to do so, various hardware and/or software that make up and enable the Accused Products, including as used in third-party (including customer) systems (including as discussed above), are a material part of the claimed invention, are not a staple article of commerce, and are incapable of substantial non-infringing uses. Defendants sell, offer to sell, and import into the United States components that constitute a material part of the invention claimed in the ’943 Patent, including BlueField DPUs and the associated firmware and software whose

function is to implement the accused isolated environments and the lockable and unlockable functionality described above. On information and belief, Defendants know that these components are specially made and adapted for use in an infringement of the ’943 Patent, and when deployed as Defendants instruct and recommend, there is no substantial use of that isolation functionality that does not practice the claimed invention.

114. On and information and belief, despite their knowledge of the ’943 Patent and of their customers’ infringement, Defendants have continued—including after September 9, 2024, and after the filing and service of this Complaint—to make, sell, offer for sale, and import the Accused Products, to distribute the software and documentation described above, and to instruct and encourage the infringing use, without redesigning the Accused Products, modifying their instructions, or taking any other action to avoid or mitigate infringement. Defendants’ induced infringement and contributory infringement of the ’943 Patent are both pre-suit and post-suit, and ongoing.

115. As explained above, Defendants’ infringement has been and continues to be willful in view of the facts asserted above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

IX.
<u>FOURTH CLAIM</u>
(Infringement of the ’634 Patent)

116. Plaintiffs incorporate by reference paragraphs 1-115 herein.

117. On information and belief, Defendants have infringed and continue to infringe the ’634 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, offering for sale, and/or importing into the United States products and/or methods that practice at least claim 8 of the ’634 Patent, including with the Accused Products.

118. The claims of the ’634 Patent are directed to non-abstract ideas and otherwise recite inventive concepts and technological improvements. Plaintiffs have complied with the requirements of 35 U.S. Code § 287 with regards to the ’634 Patent.

119. Each of the Accused Products meet every limitation of claim 8 of the ’634 Patent, which recites:

> 8. A tangible, non-transitory, computer-readable media having instructions thereupon which, when executed by a processor, cause the processor to perform a method comprising:
>
> collecting, at a first endpoint device coupled to a network, parameter values for determination of bandwidth of network connections to further endpoint devices;
>
> determining at a tuner server, distinct from the first endpoint device and the further endpoint devices, that a next network connection from the first endpoint device to a second endpoint device matches a geographical area of a past network connection; and
>
> initiating the next network connection, from the first endpoint device to the second endpoint device, based on the determination at the tuner server, with a transmission bandwidth based on the parameter values for the past network connection.

120. The Accused Products are tangible, non-transitory, computer-readable media having instructions thereupon which, when executed by a processor, cause the processor to perform a method. For example, Nvidia provides NVIDIA DOCA, which is a software framework that facilitates and accelerates the development of applications for the NVIDIA BlueField DPUs. The DOCA is the software infrastructure for the BlueField networking platform, which is a computing platform for data centers. NVIDIA DOCA is used to deliver accelerated software-defined networking, storage, security, and management services at massive scale. Some of the services include network connection optimization, orchestration, telemetry, flow integrator, etc. That DOCA software framework is one example of tangible, non-transitory, computer-readable media having instructions thereupon which, when executed by a processor, cause the processor to perform a method.

121. The Accused Products collect, at a first endpoint device coupled to a network, parameter values for determination of bandwidth of network connections to further endpoint devices. For example, NVIDIA servers (or various datacenter customer servers) in a data center (or data centers) include the NVIDIA BlueField DPUs, which execute and support the NVIDIA DOCA framework. These servers that have the NVIDIA BlueField DPUs and running/executing NVIDIA DOCA (e.g., a server different from the various endpoints devices) perform various services, including NVIDIA DOCA Telemetry Service (DTS), which runs on top of the NVIDIA BlueField DPUs.

NVIDIA DOCA Telemetry Service Guide

This guide provides instructions on how to use the DOCA Telemetry Service (DTS) container on top of NVIDIA® BlueField® DPU.

Introduction

DOCA Telemetry Service (DTS) collects data from built-in providers and from external telemetry applications. The following providers are available:

- Data providers:
  - sysfs
  - ethtool
  - tc (traffic control)
- Aggregation providers:
  - fluent_aggr
  - prometheus_aggr

Exhibit 54, https://networking-docs.nvidia.com/doca/archive/2-8-0/nvidia-doca-telemetry-service-guide#id-(2.8.0)NVIDIADOCATelemetryServiceGuide-Description.

For another example, the NVIDIA DOCA DTS "collects data from built-in providers and from external telemetry applications." The data providers may provide data on traffic control (e.g., parameter values for bandwidth determinations) as well as other network speed, congestion related data. *Id.*; Exhibit 55, https://networking-docs.nvidia.com/doca/archive/2-8-0/nvidia-doca-telemetry-service-guide. These are examples of the ways that the Accused Products collect the claimed parameter values.

122. The Accused Products determine at a tuner server, distinct from the first endpoint device and the further endpoint devices, that a next network connection from the first endpoint device to a second endpoint device matches a geographical area of a past network connection. For example, the NVIDIA DOCA DTS operating on a processer (e.g., a claimed tuner server) "collects data from built-in providers and from external telemetry applications." The data providers may provide data on traffic control (e.g., parameter values for bandwidth determinations) as well as other network speed, congestion related data. Exhibit 54. Further, DTS supports on-board collection from "sysf, ethtool, and tc providers." There are various counters, such as hardware and device counters that collect, analyze, and use traffic rate and other bandwidth related data to provide for the telemetry services in DTS. Moreover, NVIDIA DOCA DTS also provides "traffic control info," which allows the NVIDIA BlueField DPU to analyze/determine various traffic control related data including transmit and receive locations in order to optimize network traffic using the BlueField DPUs.

123. These all include a determination that a next network connection from the first endpoint device to a second endpoint device matches a geographical area of a past network connection.

Traffic Control Info

The following TC objects are supported and reported regarding the ingress filters:

- Filters
  - **flower**
- Actions
  - **mirred**
  - **tunnel_key**

The info is provided as one of the following events:

- Basic filter event
- Flower/IPv4 filter event
- Flower/IPv6 filter event
- Basic action event
- Mirred action event
- Tunnel_key/IPv4 action event
- Tunnel_key/IPv6 action event

General notes:

- Actions always belong to a filter, so action events share the filter event's ID via the `event_id` data member
- Basic filter event only contains textual *kind* (so users can see which real life objects' support they are lacking)
- Basic action event only contains textual *kind* and some basic common statistics if available

Amber Provider

Amber data for both InfiniBand and Ethernet MST devices in amBER format.

Exhibit 55

124. The Accused Products initiate the next network connection, from the first endpoint device to the second endpoint device, based on the determination at the tuner server, with a transmission bandwidth based on the parameter values for the past network connection. For example, the NVIDIA DOCA DTS "collects data from built-in providers and from external telemetry applications." The data providers may provide data on traffic control (e.g., parameter values for bandwidth determinations) as well as other network speed, congestion related data. Exhibit 54. Further, DTS supports on-board collection from "sysf, ethtool, and tc providers." There are various counters, such as hardware and device counters that collect, analyze, and use traffic rate and other bandwidth related data to provide for the telemetry services in DTS. Moreover, NVIDIA DOCA DTS also provides "traffic control info," which allows the NVIDIA BlueField DPU to analyze/determine various traffic control related data including transmit and receive locations in order to optimize network traffic using the BlueField DPUs. These all include a determination that a next network connection from the first endpoint device to a second endpoint

device matches a geographical area of a past network connection. Exhibit 55. These are examples of how the Accused Products initiate the claimed network connection.

125. Plaintiffs incorporate by reference Exhibit 33, which further demonstrates Defendants' Accused Products' infringement of claims 1 and 8 of the '634 patent.

126. On information and belief, Defendants' infringement through its use of its DPU technology and DPU-enabled systems, described above, is exemplary of its direct infringement with respect to all the Accused Products.

127. Defendants have also indirectly infringed, and continue to indirectly infringe, the '634 Patent under 35 U.S.C. § 271(b) and (c).

128. Defendants knowingly and intentionally actively aided, abetted, and induced others to directly infringe at least claim 8 of the '634 Patent (such as its customers in this District and throughout the United States), and continue to do so, by, for example, selling and offering access to and encouraging and supporting use of the Accused Products. Defendants have had knowledge of the '634 Patent and of the infringing nature of the Accused Products since no later than September 9, 2024, as alleged above, and, at the latest, since the filing and service of this Complaint. In the alternative, before September 9, 2024, Defendants were willfully blind to the existence of the '634 Patent and to the infringing nature of the Accused Products, as alleged above.

129. Defendants' customers and end users—including cloud service providers, data center operators, server OEMs, systems integrators, and enterprise customers—directly infringe at least claim 8 of the '634 Patent, literally or under the doctrine of equivalents, when they deploy BlueField DPUs and enable, configure, and run at least the NVIDIA DOCA Telemetry Service (DTS) in their networks in the United States as Defendants instruct and intend, thereby executing instructions that perform the method recited in claim 8 in the manner described above with respect to claim 8 of the '634 Patent.

130. Defendants take affirmative steps, directed at their customers, to cause and encourage the infringing use, including by authoring and publishing at least the NVIDIA DOCA Telemetry Service documentation cited above, which instructs customers, step by step, how to deploy, enable, and configure DTS on BlueField DPUs, including how to configure the data providers, counters, and traffic control information used to monitor and optimize network connections. Defendants also cause and encourage the

infringing use by developing and distributing DTS itself as part of the DOCA software stack loaded onto and used with the Accused Products and by promoting the use of DTS and BlueField DPUs to monitor and optimize network traffic. On information and belief, Defendants engage in these affirmative acts with the specific intent that their customers use the Accused Products in the infringing manner, and with knowledge or willful blindness to the fact that the induced acts constitute infringement of the ’634 Patent when customers in fact deploy and use DTS on the Accused Products as Defendants direct.

131. Defendants contributed to the direct infringement of at least claim 8 of the ’634 Patent under 35 U.S.C. § 271(c), and continue to do so, by, for example, supplying, with knowledge of the ’634 Patent, a material part of a claimed invention, where the material part is not a staple article of commerce and is incapable of substantial noninfringing use. For example, Defendants have provided, owned, operated, sold, offered to sell, leased, licensed, used, and/or imported, and continue to do so, various hardware and/or software that make up and enable the Accused Products, including as used in third-party (including customer) systems (including as discussed above), are a material part of the claimed invention, are not a staple article of commerce, and are incapable of substantial non-infringing uses. Defendants sell, offer to sell, and import into the United States components that constitute a material part of the invention claimed in the ’634 Patent, including the DOCA Telemetry Service software and the computer-readable media of the Accused Products on which that software is stored and distributed. The function of DTS is to collect and provide the telemetry, traffic control, and bandwidth information used to monitor and optimize network connections as described above. On information and belief, Defendants know that DTS is specially made and adapted for use in an infringement of the ’634 Patent and when deployed as Defendants instruct there is no substantial use of DTS that does not practice the claimed invention.

132. On and information and belief, despite their knowledge of the ’634 Patent and of their customers’ infringement, Defendants have continued—including after September 9, 2024, and after the filing and service of this Complaint—to make, sell, offer for sale, and import the Accused Products, to distribute the software and documentation described above, and to instruct and encourage the infringing use, without redesigning the Accused Products, modifying their instructions, or taking any other action to avoid or mitigate infringement. Defendants’ induced infringement and contributory infringement of the ’634 Patent are both pre-suit and post-suit, and ongoing.

133. As explained above, Defendants' infringement has been and continues to be willful in view of the facts asserted above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

X.
FIFTH CLAIM
(Infringement of the '767 Patent)

134. Plaintiffs incorporate by reference paragraphs 1-133 herein.

135. On information and belief, Defendants have infringed and continue to infringe the '767 Patent in violation of 35 U.S.C. § 271(a), either literally or through the doctrine of equivalents, by making, using, selling, offering for sale, and/or importing into the United States products and/or methods that practice at least claim 1 of the '767 Patent, including with the Accused Products.

136. The claims of the '767 Patent are directed to non-abstract ideas and otherwise recite inventive concepts and technological improvements. Plaintiffs have complied with the requirements of 35 U.S. Code § 287 with regards to the '767 Patent.

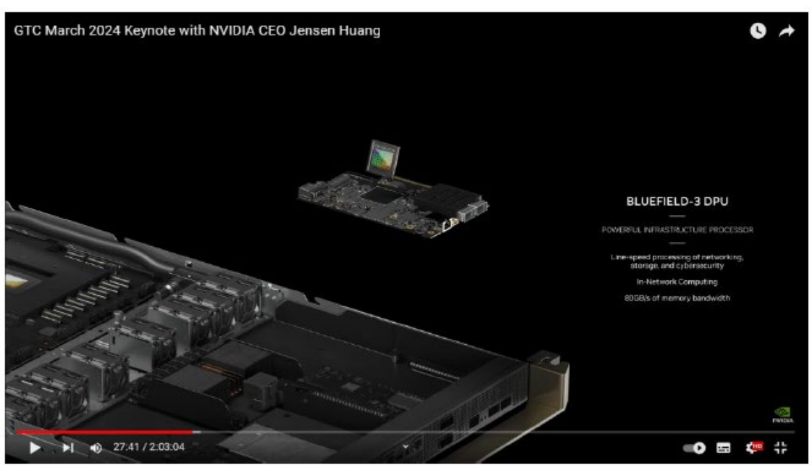
137. Each of the Accused Products meet every limitation of claim 1 of the '767 Patent, which recites:

> 1. A routable packet-switched network supported by an abstraction protocol comprising:
>
> at least one host;
>
> a master controller, the master controller comprising a master processing element, a master controller interface to communicate with the host, and at least one master controller control link interface; and
>
> at least one edge controller, the edge controller comprising an edge processing element, an edge controller control link interface to communicate with the master controller via the master controller control link interface, and at least one storage media device interface to communicate with at least one storage media device,
>
> wherein the master controller and the edge controller communicate via the abstraction protocol, the abstraction protocol comprising a full-duplex protocol supporting full command queuing for the at least one storage media device; and
>
> wherein the edge controller further comprises at least one other master controller further in communication with at least one other edge controller, when the edge processing

element associated with the edge controller authors abstract protocol messages intended for the other edge controller.

138. The Accused Products, including NVIDIA's DPU-enabled systems, make up a routable packet-switched network supported by an abstraction protocol. For example, Nvidia's DGX GB200 NVL72 server system includes a plurality of servers (e.g., multiple Nvidia GB200 NVL72 compute nodes which include Nvidia Bluefield-3 DPUs). These server systems are supported by InfiniBand or RoCE, both of which are abstraction technologies and languages at least because they provide abstraction through their layered architecture, verbs interface, and support for upper layer protocols. This is one example of how the Accused Products make up the claimed routable packet-switched network system.

Exhibit 49



https://www.youtube.com/watch?v=Y2F8yisiS6E

NVIDIA GB200 NVL72 Platforms Arrive Equipped With NVIDIA BlueField-3 DPUs

Flagship, rack-scale platforms powered by the Grace Blackwell accelerator and NVIDIA BlueField-3 networking innovations enable breakthroughs in data processing for the most compute-intensive workloads.

Exhibit 51

The GB200 is a key component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system for the most compute-intensive workloads. It combines 36 Grace Blackwell Superchips, which include 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink. Additionally, GB200 NVL72 includes NVIDIA BlueField®-3 data processing units to enable cloud network acceleration, composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds. The GB200 NVL72 provides up to a 30x performance increase compared to the same number of NVIDIA H100 Tensor Core GPUs for LLM inference workloads, and reduces cost and energy consumption by up to 25x.

Exhibit 52

BlueField-3 DPU – 2x 200Gb/s FHHL form factor

Exhibit 45

Network and Host Interfaces

Network Interfaces

- Ethernet - 1, 2, 4 ports with up to 400 Gb/s connectivity
- InfiniBand - Single port of NDR (400Gb/s), or dual ports of NDR200 / HDR (200Gb/s)

PCI Express Interface

- 32 lanes of PCIe Gen 5.0
- PCIe switch bi-furcation of up to 16 downstream ports
- Non-transparent bridging (NTB) support

Compute and Memory

Arm CPU Cores

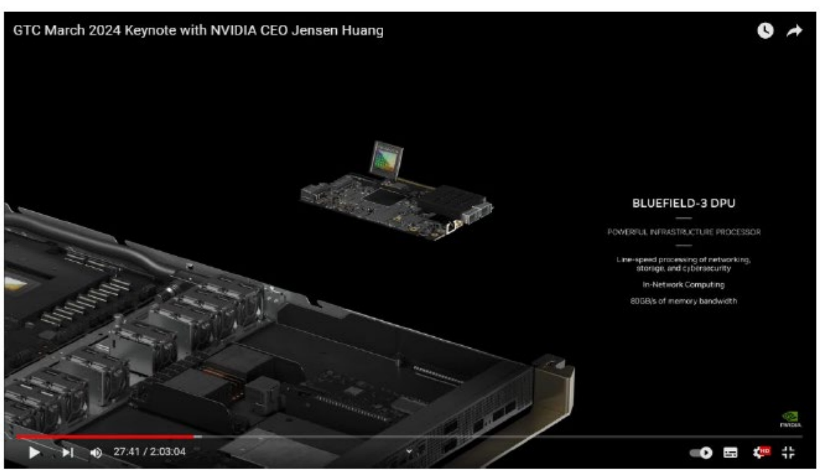
- Up to 16 Armv8.2+ A78 Hercules cores (64-bit)
- 8MB L2 cache
- 16MB LLC system cache

Programmable Datapath Accelerator

- 16 cores, 256 threads
- Programmability through DOCA
- Heavy multi-threading applications acceleration

Exhibit 45

139. The Accused Products, including NVIDIA's DPU-enabled systems, include at least one host. For example, NVIDIA's server platforms, e.g., NVIDIA's DGX GB200 NVL72 server system, include multiple compute trays in a rack, with each tray containing BlueField-3 DPUs. Furthermore, the NVIDIA BlueField DPU is connected to the network via Ethernet ports or NDR400 InfiniBand for network connectivity.



https://www.youtube.com/watch?v=Y2F8yisiS6E

The NVIDIA® BlueField®-3 networking platform is designed to accelerate data center infrastructure workloads and usher in the era of accelerated computing and AI. Supporting both Ethernet and InfiniBand connectivity, BlueField-3 offers speeds up to 400 gigabits per second (Gb/s). It combines powerful computing with software-defined hardware accelerators for networking, storage, and cybersecurity—all fully programmable through the NVIDIA DOCA™ software framework. Drawing on the platform's robust capabilities, BlueField data processing units (DPUs) and BlueField SuperNICs revolutionize traditional computing environments, transforming them into secure, high-performance, efficient, and sustainable data centers suitable for any workload at any scale.

Exhibit 39

Network and Host Interfaces

Network Interfaces

> 1 or 2 ports with up to 400Gb/s Ethernet or NDR InfiniBand connectivity

PCI Express Interface

> 32 lanes of PCIe Gen 5.0

> Flexible PCIe switch supporting self-hosting and server-hosting

Exhibit 39

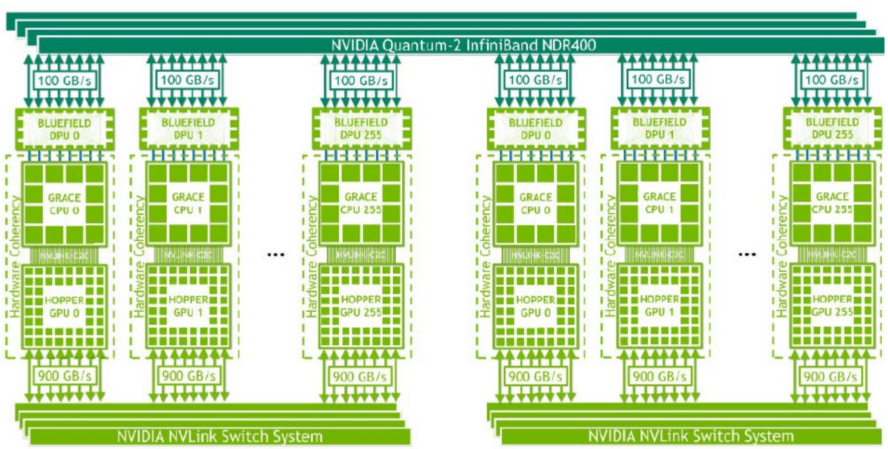


Figure 14. NVIDIA HGX Grace Hopper with NVLink Switch System for strong-scaling giant ML and HPC workloads

Exhibit 40

140. The Accused Products, including NVIDIA's DPU-enabled systems, include a master controller, the master controller comprising a master processing element, a master controller interface to communicate with the host, and at least one master controller control link interface. For example, each of the one or more Nvidia Bluefield DPUs comprises a master processing element such as ARM cores, Data

path accelerators, and ASAP2. Further, the Nvidia Bluefield DPU includes various types of master controller interfaces and control link interface, such as PCIe, ethernet, or InfiniBand to communicate with the host. One of the one or more Nvidia Bluefield DPUs can act as the master controller.

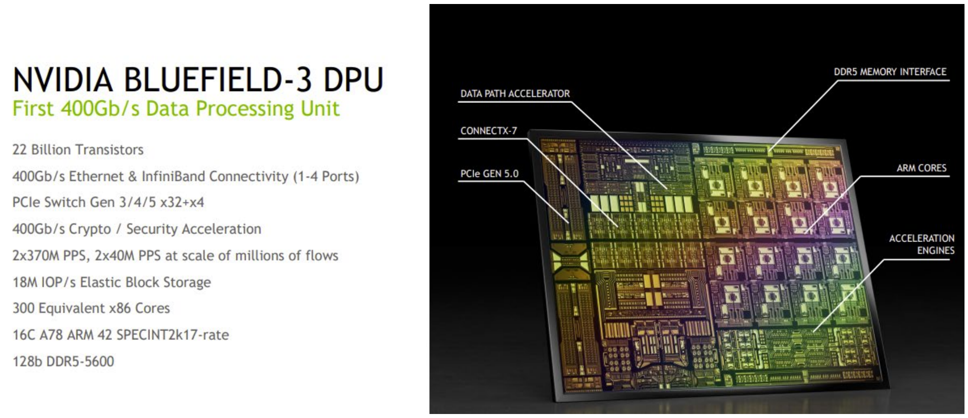
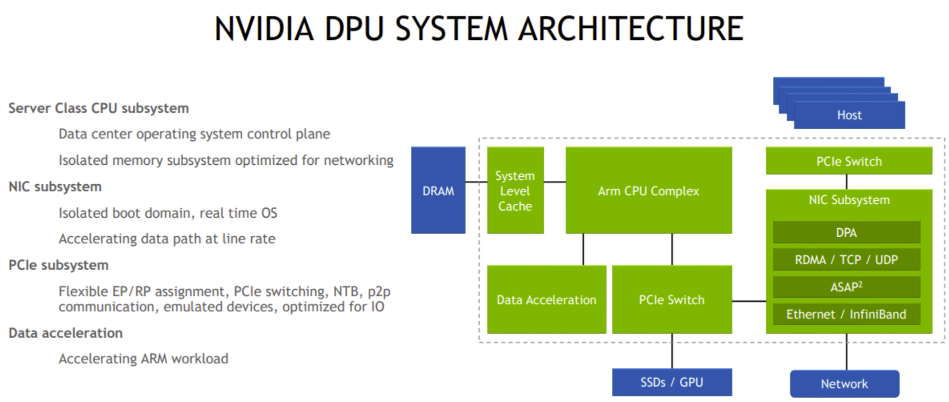


Exhibit 41



Id.

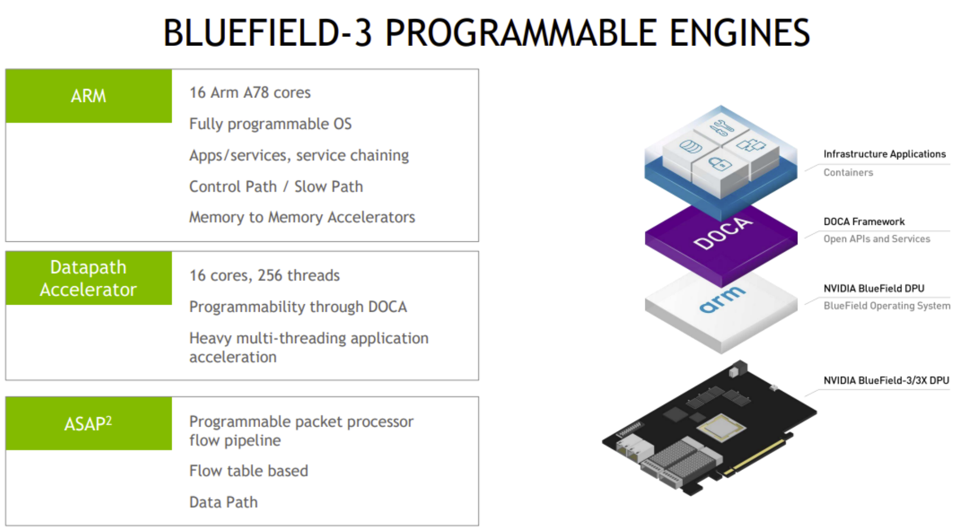


Exhibit 41

141. The Accused Products, including NVIDIA's DPU-enabled systems, include at least one edge controller, the edge controller comprising an edge processing element, an edge controller control link interface to communicate with the master controller via the master controller control link interface, and at least one storage media device interface to communicate with at least one storage media device. For example, NVIDIA servers contain multiple Bluefield DPUs connected with each other to create a cloud fabric. Further, in a datacenter for example, Bluefield DPUs are deployed in other servers as well as storage controllers that can act as an edge controller. These DPUs have similar capabilities to the master controller. Furthermore, these edge controllers are connected to storage mediums. These Bluefield DPUs are examples of and operate as the claimed edge controllers, and thus contain an edge processing element, an edge controller control link interface to communicate with the master controller via the master controller control link interface, and at least one storage media device interface to communicate with at least one storage media device.

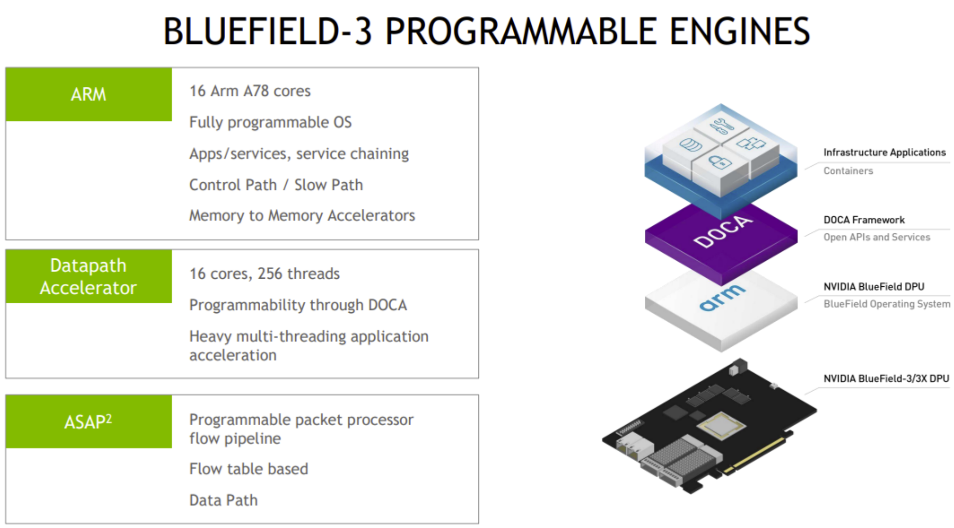
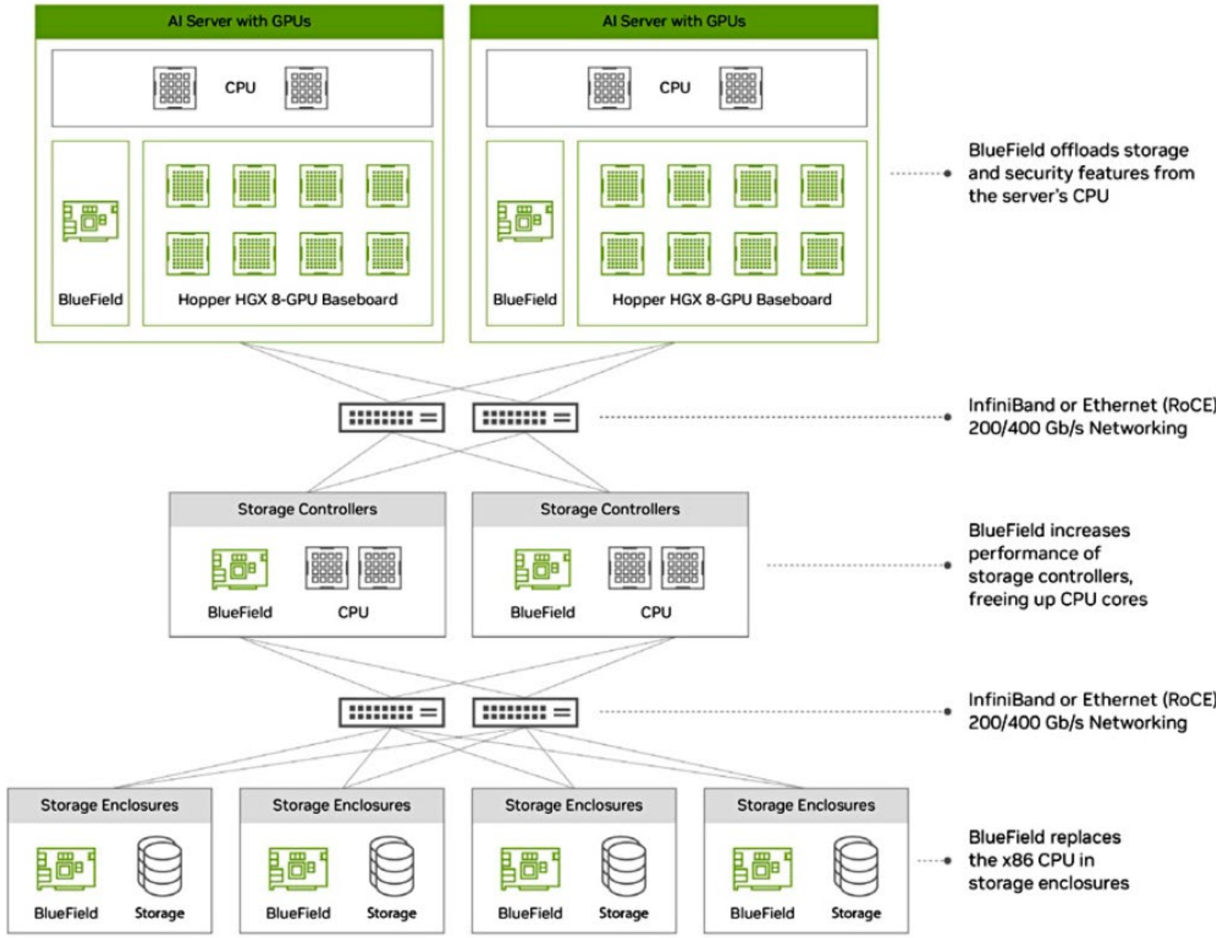


Exhibit 41



BlueField enhances storage performance, security, and efficiency in the server, storage controller, and storage enclosure.

Exhibit 56, https://blocksandfiles.com/2024/03/12/vast-data-turns-bluefield3-dpus-into-storage-controllers-for-nvidia-gpu-servers/

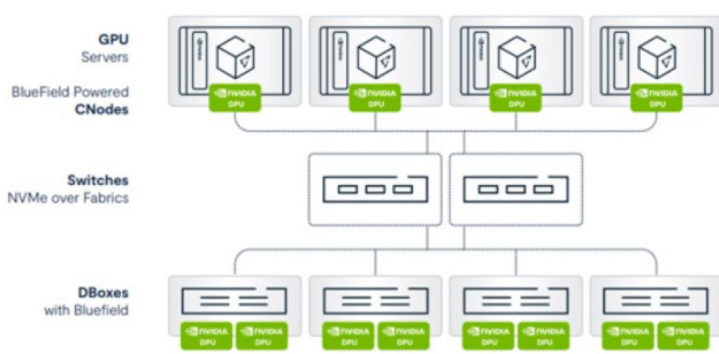


Exhibit 57, https://developer.nvidia.com/blog/spotlight-nvidia-bluefield-dpus-power-the-vast-data-platform-for-ai-workload-optimization/

142. The Accused Products, including NVIDIA's DPU-enabled systems, include the master controller and the edge controller communicating via the abstraction protocol, the abstraction protocol comprising a full-duplex protocol supporting full command queuing for the at least one storage media device, as shown by the following diagrams:

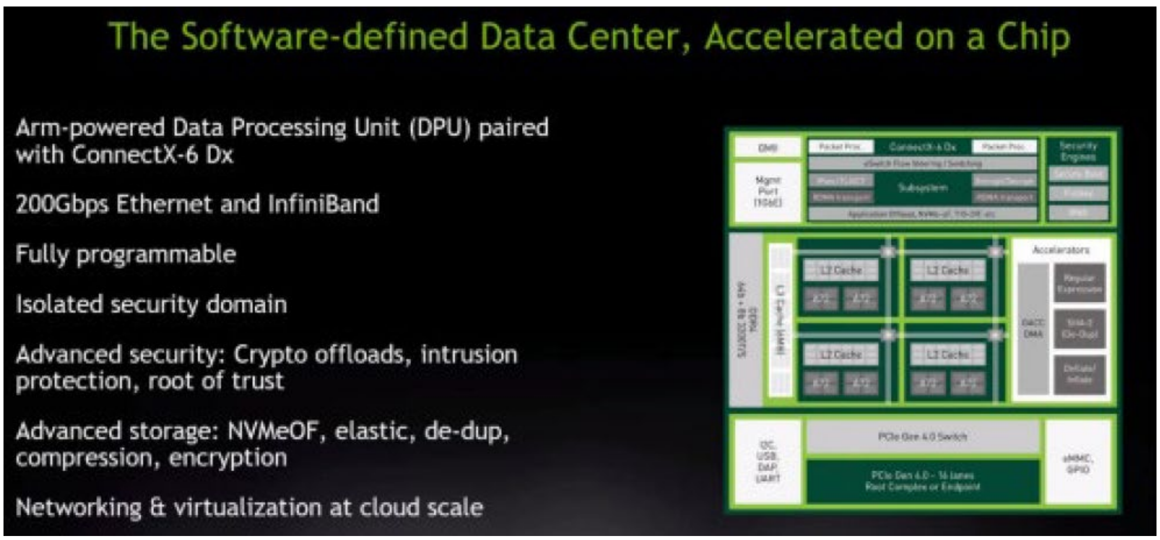
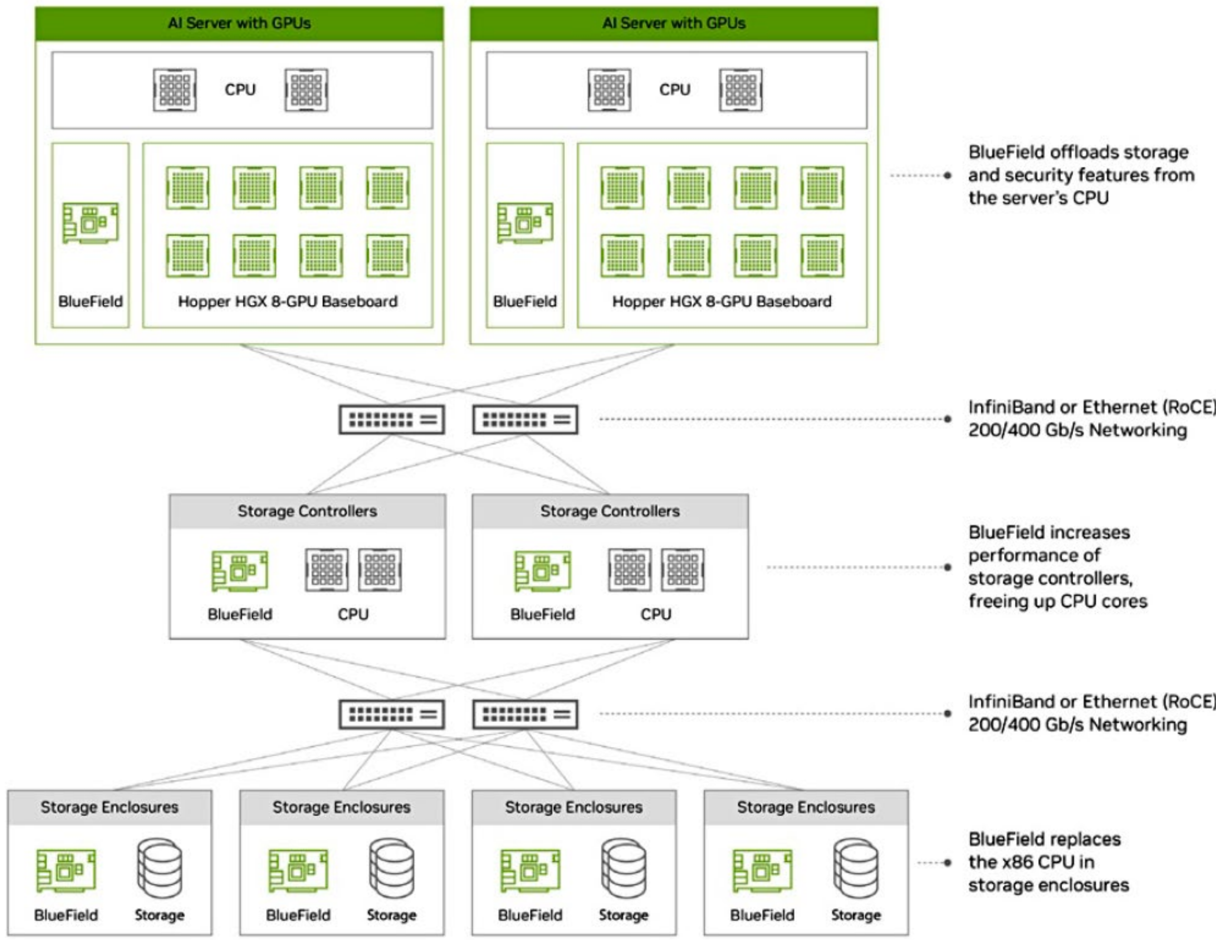


Exhibit 42



BlueField enhances storage performance, security, and efficiency in the server, storage controller, and storage enclosure.

Exhibit 56

143. The Accused Products, including NVIDIA's DPU-enabled systems, include the edge controller further comprising at least one other master controller further in communication with at least one other edge controller, when the edge processing element associated with the edge controller authors abstract protocol messages intended for the other edge controller. For example, NVIDIA servers containing Bluefield DPUs (e.g., master controller) can communicate with different NVIDIA servers with different Bluefield DPUs (e.g., edge controller) via InfiniBand or RoCE. Both RoCE and InfiniBand are abstraction technologies at least because they provide abstraction through their layered architecture, verbs interface, and support for upper layer protocols. *See* Exhibit 42; Exhibit 56. Furthermore, the SCSI RDMA Protocol (SRP) is supported by InfiniBand and RDMA. It is a network protocol that enables systems to use Remote Direct Memory Access (RDMA) to access SCSI devices attached to remote systems.

Storage Protocols

There are several storage protocols that use the advantage of InfiniBand and RDMA for performance reasons (high throughput, low latency and low CPU utilization). In this chapter we will discuss the following protocols:

- **SCSI RDMA Protocol (SRP)** is designed to take full advantage of the protocol off-load and RDMA features provided by the InfiniBand architecture.
- **iSCSI Extensions for RDMA (iSER)** is an extension of the data transfer model of iSCSI, a storage networking standard for TCP/IP. It uses the iSCSI components while taking the advantage of the RDMA protocol suite. ISER is implemented on various stor- age targets such as TGT, LIO, SCST and out of scope of this manual.
  For various ISER targets configuration steps, troubleshooting and debugging, as well as other implementation of storage protocols over RDMA (such as Ceph over RDMA, nbdX and more) refer to Storage Solutions on the Community website.

Exhibit 58, https://docs.nvidia.com/doca/sdk/storage+protocols/index.html

iSCSI Extensions for RDMA (iSER)

iSCSI Extensions for RDMA (iSER) extends the iSCSI protocol to RDMA. It permits data to be transferred directly into and out of SCSI buffers without intermediate data copies.

iSER uses the RDMA protocol suite to supply higher bandwidth for block storage transfers (zero time copy behavior). To that fact, it eliminates the TCP/IP processing overhead while preserving the compatibility with iSCSI protocol.

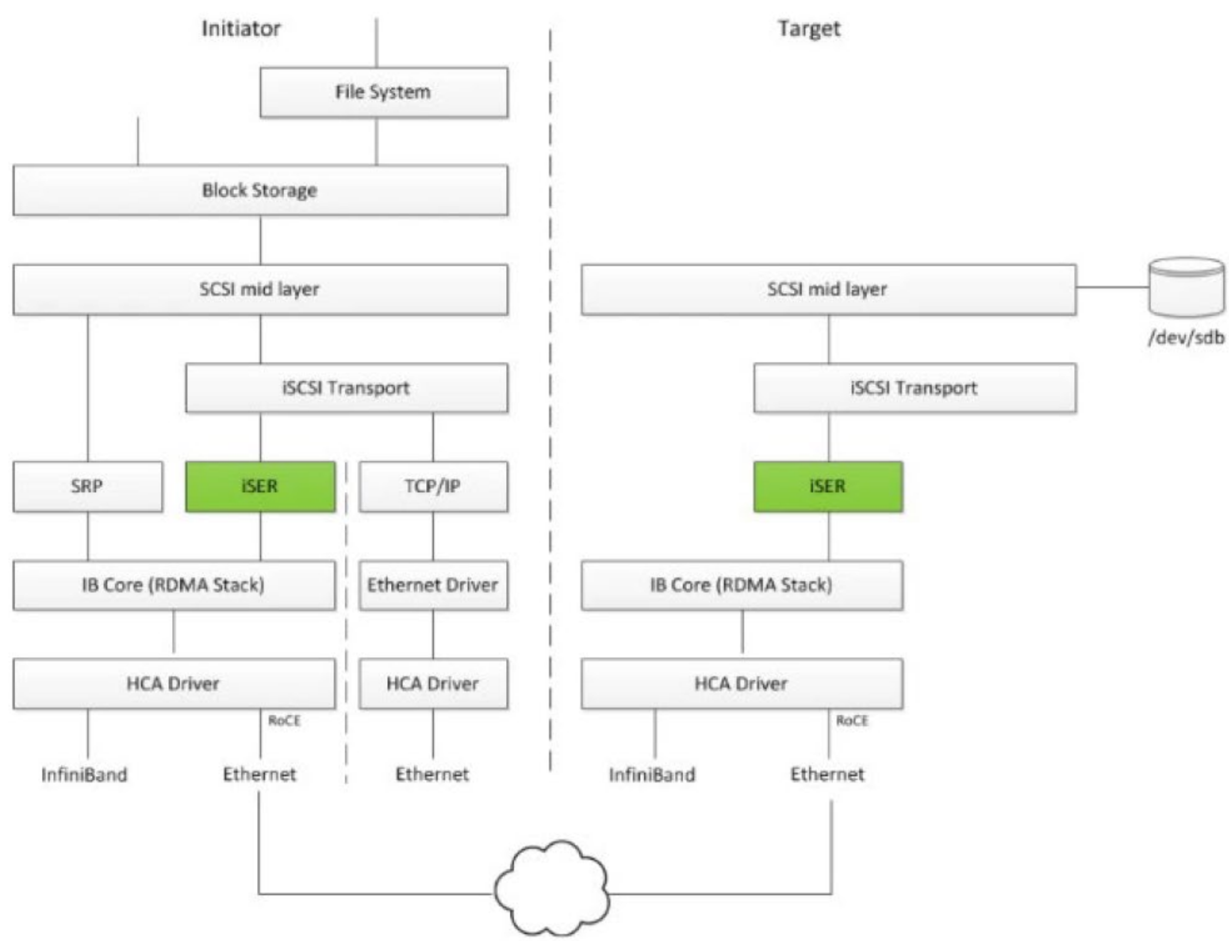


Exhibit 59, https://networking-docs.nvidia.com/doca/archive/2-5-4/iscsi-extensions-for-rdma-iser

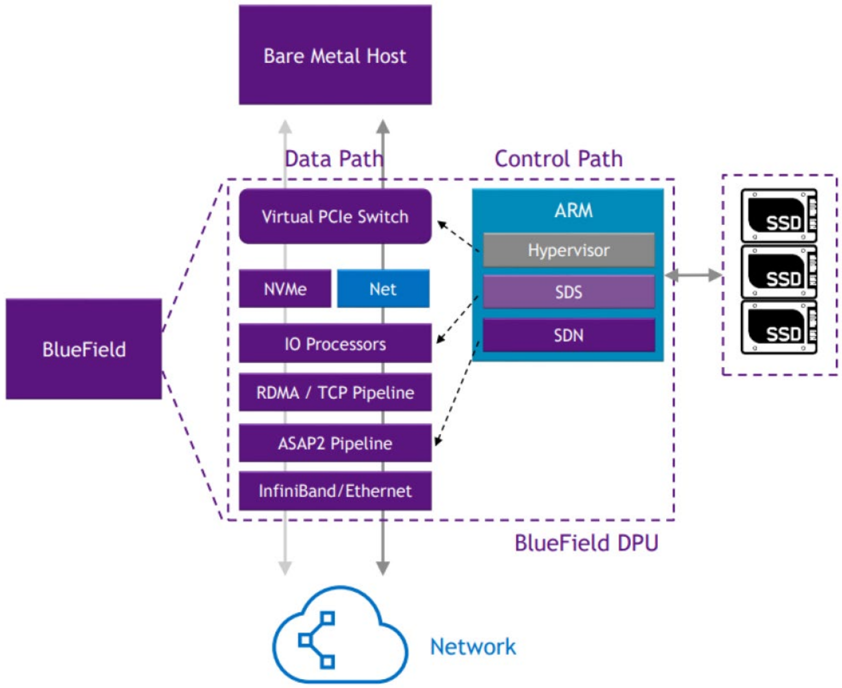


Exhibit 41

144. Furthermore, other Bluefield DPUs (e.g., edge controller, other master controller, other edge controller) has its own respective ARM cores, DPA, and ASAP2 processors that communicate with at least one other DPU deployed in another server in the data center or other data centers, and these DPUs (and others) are in communication with each other in various ways, including when the abstract protocol messages (e.g., InfiniBand, RoCE, SCSI RDMA Protocol) are authored. In at least this way, the edge controller further comprises at least one other master controller in communication with at least one other edge controller (e.g., another DPU) and authors abstract protocol messages intended for the other edge controller.

145. Plaintiffs incorporate by reference Exhibit 36, which further demonstrates Defendants' Accused Products' infringement of claim 1 of the '767 patent.

146. On information and belief, Defendants' infringement through its use of its DPU technology and DPU-enabled systems, described above, is exemplary of its direct infringement with respect to all the Accused Products.

147. Defendants have also indirectly infringed, and continue to indirectly infringe, the ’767 Patent under 35 U.S.C. § 271(b) and (c).

148. Defendants knowingly and intentionally actively aided, abetted, and induced others to directly infringe at least claim 1 of the ’767 Patent (such as its customers in this District and throughout the United States), and continue to do so, by, for example, selling and offering access to and encouraging and supporting use of the Accused Products. Defendants have had knowledge of the ’767 Patent and of the infringing nature of the Accused Products since no later than September 9, 2024, as alleged above, and, at the latest, since the filing and service of this Complaint. In the alternative, before September 9, 2024, Defendants were willfully blind to the existence of the ’767 Patent and to the infringing nature of the Accused Products, as alleged above.

149. Defendants’ customers and end users—including cloud service providers, data center operators, storage system vendors, server OEMs, systems integrators, and enterprise customers—directly infringe at least claim 1 of the ’767 Patent, literally or under the doctrine of equivalents, when they deploy multiple BlueField DPUs across servers and storage systems in a data center, configure them as controllers for storage media devices, and use RDMA-based storage protocols described in Defendants’ documentation to communicate among those controllers and with storage media devices in the manner described above with respect to claim 1 of the ’767 Patent.

150. Defendants take affirmative steps, directed at their customers, to cause and encourage the infringing use, including by authoring and publishing the DOCA storage protocol documentation cited above, which instructs customers how to configure RDMA-based storage communications among controllers and storage media devices. Defendants also cause and encourage the infringing use by publishing materials promoting the deployment of BlueField DPUs as storage controllers, developing and distributing the DOCA software and firmware that implement and enable that functionality, and by marketing multi-DPU storage fabrics and DPU-based storage acceleration as an intended use of the Accused Products. On information and belief, Defendants engage in these affirmative acts with the specific intent that their customers use the Accused Products in the infringing manner, and with knowledge or willful blindness to the fact that the induced acts constitute infringement of the ’767 Patent, and that

the induced use is an intended, advertised, and promoted use of the Accused Products, and customers in fact deploy and use the Accused Products as Defendants direct.

151. Defendants contributed to the direct infringement of at least claim 1 of the ’767 Patent under 35 U.S.C. § 271(c), and continue to do so, by, for example, supplying, with knowledge of the ’596 Patent, a material part of a claimed invention, where the material part is not a staple article of commerce and is incapable of substantial noninfringing use. For example, Defendants have provided, owned, operated, sold, offered to sell, leased, licensed, used, and/or imported, and continue to do so, various hardware and/or software that make up and enable the Accused Products, including as used in third-party (including customer) systems (including as discussed above), are a material part of the claimed invention, are not a staple article of commerce, and are incapable of substantial non-infringing uses. Defendants sell, offer to sell, and import into the United States components that constitute a material part of the invention claimed in the ’767 Patent, including BlueField DPUs and the associated DOCA storage protocol software whose function is to implement the storage communications described above. On information and belief, Defendants know that these components are especially made and especially adapted for use in an infringement of the ’767 Patent, and when deployed as Defendants instruct, there is no substantial use of that storage protocol software that does not practice the claimed invention.

152. Despite their knowledge of the ’767 Patent and of their customers’ infringement, Defendants have continued—including after September 9, 2024, and after the filing and service of this Complaint—to make, sell, offer for sale, and import the Accused Products, to distribute the software and documentation described above, and to instruct and encourage the infringing use, without redesigning the Accused Products, modifying their instructions, or taking any other action to avoid or mitigate infringement. Defendants’ induced infringement and contributory infringement of the ’767 Patent are both pre-suit and post-suit, and ongoing.

153. As explained above, Defendants’ infringement has been and continues to be willful in view of the facts asserted above and its failure to take any action, even after being put on notice, to stop its infringement or inducement of, or contribution to, infringement by others.

DEMAND FOR JURY TRIAL

Pursuant to Federal Rule of Civil Procedure 38(b), Plaintiffs hereby demand a trial by jury on all issues triable to a jury.

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs pray for judgment against Defendants as follows:

A. That Defendants have infringed each of the Asserted Patents, and unless enjoined, will continue to infringe one or more of the applicable Asserted Patents;

B. That Defendants' infringement of one or more of the applicable Asserted Patents has been willful;

C. That Defendants pay Plaintiffs damages adequate to compensate Plaintiffs for Defendants' past infringement of each of the Asserted Patents, and present and future infringement of the applicable Asserted Patents, together with interest and costs under 35 U.S.C. § 284, including supplemental damages for any continuing post-verdict infringement up until entry of judgment and beyond, with accounting, as needed;

D. That Defendants pay prejudgment and post-judgment interest on the damages assessed;

E. That Defendants pay Plaintiffs enhanced damages pursuant to 35 U.S.C. § 284;

F. That Defendants be enjoined from infringing the applicable Asserted Patents, or if its infringement is not enjoined, that Defendants be ordered to pay ongoing royalties to Plaintiffs for any post-judgment infringement of the applicable Asserted Patents;

G. That this is an exceptional case under 35 U.S.C. § 285; and that Defendants pay Plaintiffs' attorneys' fees and costs in this action; and

H. That Plaintiffs be awarded such other and further relief, including equitable relief, as this Court deems just and proper.

Dated: July 17, 2026

Respectfully Submitted,

/s/ *Laura Carwile*

Michael Matulewicz-Crowley*
mmc@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
650 Fifth Avenue, Suite 2320
New York, NY 10019
Telephone: (212) 381-1965

Christine E. Lehman*
clehman@reichmanjorgensen.com
Zachary L. Jacobs*
zjacobs@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
1909 K Street, NW, Suite 800
Washington, DC 20006
Telephone: (202) 894-7311

Laura Carwile (CA Bar No. 291906)
lcarwile@reichmanjorgensen.com
REICHMAN JORGENSEN LEHMAN & FELDBERG LLP
100 Marine Parkway, Suite 300
Redwood Shores, CA 94065
Telephone: (650) 623-1401

**Pro hac vice* forthcoming

Attorneys for Plaintiffs
Concurrent Ventures, LLC and XtreamEdge, Inc.