Michael Matulewicz-Crowley
mmc@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
400 Madison Avenue, Suite 14D
New York, NY 10017
Telephone: (212) 381-1965

Christine E. Lehman
clehman@reichmanjorgensen.com
Adam Adler
aadler@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
1909 K Street NW, Suite 800
Washington DC, 20006
Telephone: (202) 894-7310

Courtland L. Reichman (CA Bar No. 268873)
creichman@reichmanjorgensen.com
Jennifer Estremera (CA Bar No. 251076)
jestremera@reichmanjorgensen.com
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

Scott Cole
scole@reichmanjorgensen.com
Ariane S. Mann
amann@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
515 Congress Avenue, Suite 1900
Austin, TX 78701
Telephone: (650) 623-1401

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

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

Concurrent Ventures, LLC and XtreamEdge, Inc.,	§	
Plaintiffs,	§	
v.	§	**CIVIL ACTION NO. 4:25-cv-09567-JST**
Advanced Micro Devices, Inc. and Pensando Systems, Inc.,	§	**JURY TRIAL DEMANDED**
Defendants.	§	

I. FIRST AMENDED COMPLAINT FOR PATENT INFRINGEMENT

Plaintiffs Concurrent Ventures, LLC ("Concurrent Ventures") and XtreamEdge, Inc. ("XtreamEdge") (collectively, "Plaintiffs") file this First Amended Complaint against Defendants Advanced Micro Devices, Inc. ("AMD") and Pensando Systems, Inc. ("Pensando") (collectively, "Defendants"), and in support thereof alleges as follows:

II. 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 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, in the form of additional Central Processing Units ("CPUs"). 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. The artificial intelligence ("AI") revolution has compounded the data processing problem. Graphics Processing Units ("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. "Agentic AI is driving higher demands across the stack—requiring not just powerful GPUs, but also high-performance CPUs and secure, efficient networking."[1] GPU sales have correspondingly skyrocketed in recent

[1] https://www.amd.com/en/blogs/2025/amd-delivering-open-rack-scale-ai-infrastructure-to-unlock-agentic-ai.html.

years. Defendants themselves advertise that their GPUs and CPUs are specially designed for AI uses.[2]

4. Adding more CPUs and GPUs is not a sustainable strategy, and is not by itself sufficient. Additional CPUs and GPUs lead to higher capital expenditures and operational expenditures, lower flexibility, and increased latency. Simply using more 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. And 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 CPUs and other network components. Plaintiffs' innovations, among other benefits, solves long-standing problems by offloading data-centric computations, which are estimated to be between 31-83% of the workload,[4] to more efficient units designed to handle data computations. By offloading the networking, communication, and other tasks from the CPU, the CPU is freed up to focus on what it does best.

---

[2] https://www.amd.com/en/products/graphics/radeon-ai.html; https://www.amd.com/en/solutions/ai.html.

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

[4] https://www.sdxcentral.com/articles/interview/amd-shares-vision-for-pensando-smart-switch-dpu/2023/04.

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. Pensando Systems Inc. pursued similar solutions as XtreamEdge—years after XtreamEdge—but called its product a Data Processing Unit ("DPU").[5] AMD acquired Pensando in 2022, leading to the integration of Pensando's DPU-enabled systems into AMD's networking, computing, and security services platforms.

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. *See, e.g.*, https://www.sdxcentral.com/articles/interview/amd-shares-vision-for-pensando-smart-switch-dpu/2023/04 ("'If you look at the journey that AMD is on in the data center, it's really to look at how to go about building the industry's highest performing adaptive computing portfolio,' Jiandani said. 'I think DPUs are the most efficient way to deal with things that need to have data processing at the edge.'"); https://www.techradar.com/pro/future-servers-could-have-a-shared-dpu-could-the-next-decade-see-a-rise-in-socket-heterogeneity (predicting that "there is a world where all hyperscalers will leverage servers with DPUs" and that "AMD is ideally positioned to be the one-stop shop for compute."). Robert Hormuth, AMD Vice President of Architecture & Strategy, stated that "[m]odern infrastructure is moving to offload much of the OS and hypervisor software stack to the DPU/SmartNIC. DPUs like the AMD Pensando™ enable DPU sharing across multiple servers. Single-socket servers are ideal for this new deployment model - connecting multiple

[5] https://techcrunch.com/2022/04/04/amd-to-acquire-data-center-optimization-startup-pensando-for-2b.

single-socket servers to one DPU."[6] AMD has 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, (b) DPUs, CPUs, and GPUs are functionally interrelated, and (c) Defendants leverage that functional relationship to promote sales of CPUs and GPUs. 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 (e.g., DPUs, also referred to as AI Network Interface Cards ("NICS")) as a part of their "End-to-End" AI Solutions.[8] For example, Defendants' accused "AMD Pensando™ Pollara 400 AI NIC is engineered to accelerate applications running across AI nodes in mega-scale and giga-scale data centers, achieving up to 400 Gigabit per second (Gbps) Ethernet speeds."[9] For another example, AMD's accused Helios Rack is a combination of Defendants' GPUs, CPUs, and the Accused Pensando NICs, all working together.[10] Defendants

---

[6] https://www.amd.com/en/solutions/data-center/insights/myths-and-urban-legends-about-dual-socket-servers.html.

[7] https://www.nextplatform.com/2023/05/03/amd-says-ai-is-the-number-one-priority-right-now/?mc_cid=c589551dac&mc_eid=57a4c0e31b; https://www.amd.com/en/newsroom/press-releases/2025-11-11-amd-unveils-strategy-to-lead-the-1-trillion-compu.html (identifying "AMD networking solutions power AI at scale, with Pensando™ Pollara and next-generation 'Vulcano' AI NICs" as a part of Defendants' "leadership across a broad portfolio of hardware, software and solutions to power the full spectrum of high-performance and AI compute.").

[8] https://www.amd.com/en/solutions/ai.html.

[9] https://www.amd.com/en/products/network-interface-cards/pensando.html (discussing "The Critical Role of NIC Programmability in Scaling Out Data Center Networks for AI").

[10] https://www.amd.com/en/blogs/2025/amd-delivering-open-rack-scale-ai-infrastructure-to-unlock-agentic-ai.html.

advertise the Helios rack as the answer to the needs of the AI boom.[11] This integration is not surprising—AMD acquired Pensando explicitly in order to [redacted]

10. Defendants explicitly advertise that the Accused Products work together with CPUs and GPUs to improve those processors.[13] 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' DPU product page states that their DPU allows customers to save 30% on costs by "[o]ffloading" work from CPU cores.[14] Similarly, in a June 3, 2025 blog post, Defendants boasted that their DPU "lowers the workload on CPUs and GPUs, making the whole system more efficient" and that DPUs "lower[] host compute (CPU/GPU) usage."[15] Defendants' promotional materials show that they used these synergies to promote bundled sales of DPUs and CPUs and GPUs — including its Next-Gen Helios rack.[16]

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

[11] https://www.amd.com/en/blogs/2025/amd-delivering-open-rack-scale-ai-infrastructure-to-unlock-agentic-ai.html.
[12] Exhibit 8 (AMD-CV-0054650) at AMD-CV-0054662.
[13] https://www.amd.com/en/products/network-interface-cards/pensando.html ("Achieve up to 25% improvement in RCCL performance, significantly boosting multi-GPU and scale-out network efficiency") ("With up to 400 Gbps GPU-GPU communication speeds . . . .").
[14] https://www.amd.com/en/products/data-processing-units/pensando.html.
[15] https://www.amd.com/en/blogs/2025/accelerating-vtap-performance-in-hyperscale-clouds.html; *see also* https://www.amd.com/en/blogs/2025/amd-and-oracle-cloud-infrastructure-are-powering-the-ne.html ("By integrating the Pensando Pollara into OCI's AMD Instinct MI355X superclusters, data can move at the speed of compute, helping ensure that the full capability of each GPU is realized without bottlenecks.").
[16] https://www.amd.com/en/blogs/2025/amd-helios-ai-rack-built-on-metas-2025-ocp-design.html (boasting that Helios "integrate[s]" GPUs, CPUs, and DPUs).

DPUs and CPUs "complement each other in data-centric tasks. The combination of the two increases system performance and optimizes resource consumption."[17] The point of DPUs is to function with CPUs and GPUs, to make them more efficient.

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

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").

III. PARTIES

[17] https://monovm.com/blog/cpu-vs-dpu; *see also, e.g.*, https://network-switch.com/blogs/networking/smart-nics-and-dpus-explained ("DPUs won't replace CPUs or GPUs, but they'll make both more effective by handling the critical third function – data movement and protection."); https://arxiv.org/abs/2503.22930 (DPUs achieve "substantial reductions in host CPU usage and, in many cases, significant performance improvements"); https://www.techtarget.com/searchdatacenter/tip/How-do-CPU-GPU-and-DPU-differ-from-one-another ("By using multiple types of processing units in your data center, you can have all the units support each other and further speed up large or complex tasks.").

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. Advanced Micro Devices, Inc. is a corporation organized and existing under the laws of Delaware that maintains established places of business at 2485 Augustine Drive, Santa Clara, California 95054 and 7171 Southwest Parkway, Austin, Texas 78735.

17. Pensando Systems, Inc. is a corporation organized and existing under the laws of Delaware that maintains established places of business at 2485 Augustine Drive, Santa Clara, California 95054 and 7171 Southwest Parkway, Austin, Texas 78735.

IV. ASSERTED PATENTS

18. 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.

19. The '596 provides "a system and method for dividing and synchronizing a processing task across multiple processing elements/processors in hardware." '596 pat., 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.

20. 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.

21. 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 pat., 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.

22. 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.

23. 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.

24. 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.

25. 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 pats. 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.

26. 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.

27. 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 pat., 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.

28. 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 pat., 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.

29. 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.").

30. 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 pat., 1:6-29.

31. 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 pat., 13:65-14:13; 8:51-67; 15:40-16:8.

32. 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.

33. The '767 patent claims an invention in "the field of storage devices." '767 pat., 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.

34. 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.

35. 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 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.

36. 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 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.

37. 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, where 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.

V. JURISDICTION AND VENUE

38. Plaintiffs incorporate by reference paragraphs 1-16 herein.

39. 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).

40. 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.

41. 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 Texas.

42. Defendants have admitted that venue is appropriate and convenient in this district. Dkt. 33.

43. 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).

VI. DEFENDANTS ARE INTENTIONALLY AND KNOWINGLY USING PLAINTIFFS' TECHNOLOGY

45. Plaintiffs incorporate by reference paragraphs 1-26 herein.

46. 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.

47. The Accused Products are all products including or related to AMD's DPU technology, including without limitation, AMD's Pensando Data Processing Units including the AMD Pensando Giglio Data Processing Unit, the Elba Data Processing Unit, the Capri Data Processing Unit, Salina Data Processing Unit, Salina 2 Data Processing Unit, Vulcano Data Processing Unit, and future versions, the AMD Pensando Distributed Services Card, the AMD Pensando Distributed Services Platform, the AMD Pensando DPU system, the AMD Pensando SmartNIC, the AMD

Pensando SmartSwitch, AMD Pensando Pollara 400 AI NIC, Helios Rack, and Next Gen Rack as well as any products incorporating those items (collectively, the "Accused Products/Services"). Each of these products and services practice at least one claim of each Asserted Patent.

48. In 2014, AMD held detailed technical discussions with Concurrent Ventures' commercialization agent, HellaStorm, Inc., during which AMD was provided with the specifics of Concurrent Ventures' technology. HellaStorm and AMD entered into a "Mutual Nondisclosure Agreement" for the purpose of "Internal evaluation and/or testing," which was signed by Jesse Beeson, the CEO of Concurrent Ventures, and Harry Wolin, AMD's Senior Vice President and General Counsel. The parties met on-site at AMD in June 2014, and the meeting included a number of AMD executives, including Anil Rao, the then Corporate Vice President of Products who "[d]rove data center platforms, systems and software for AMD's Server Business Unit"[18] and Dhiraj Mallick, the then General Manager and Corporate Vice President responsible "for $20B data center group" and "next-generation data center architectures, solutions, and innovations."[19]

49. After that meeting, AMD was sent a proposed engagement plan, outlining a potential project. The parties did not ultimately agree to this proposed project.

50. The 2014 meeting was not Defendants' only exposure to the patented technology. In 2019, HellaStorm met with Dell Inc., during which the participants were again given the specifics of Concurrent Ventures' patented technology. One of the participants of this meeting was Robert Hormuth, who was then Dell's "CTO, VP/Fellow, Server & Infrastructure Systems."[20] Mr. Hormuth left Dell just over a year after this meeting and joined AMD, becoming its "Corporate Vice President, Architecture and Strategy, Data Center Solutions Group"—a role he is still in.[21]

---

[18] https://www.linkedin.com/in/anilrrao.
[19] https://www.linkedin.com/in/dhirajmallick/details/experience.
[20] https://www.linkedin.com/in/roberthormuth.
[21] *Id.*

On information and belief, Mr. Hormuth is pictured below in the Dell meetings, in the process of being informed of Concurrent Ventures' patented technology.

51. Notably, AMD's acquisition of Pensando (and the beginning of their infringment) began relatively soon after Mr. Hormuth joined AMD, in early 2022. Mr. Hormuth was a [redacted] Mr. Hormuth was one of the AMD employees with [redacted] On information and belief, Mr. Hormuth brought his knowledge of the patented technology with him to AMD and it was that knowledge that, in part, prompted AMD to acquire Pensando and thus to begin infringing. On information and belief, AMD was thus aware (through Mr. Hormuth and its earlier meeting with HellaStorm) of its specific infringement of Concurrent Ventures' patented technology when it began to infringe. And,

[22] Exhibit 9 (AMD-CV-0054619) at -54622.
[23] Exhibit 8 (AMD-CV-0054650) at -54668.

on information and belief, Mr. Hormuth also informed Pensando, both before and after the acquisition of its use of Concurrent Ventures' patented technology. Thus, on information and belief, Pensando was aware (through Mr. Hormuth and its interactions with AMD) of its specific infringement of Concurrent Ventures' patented technology at least as of the beginning of 2022.

52. In sum, on information and belief, Defendants have been on notice of the Asserted Patents and their infringement since at least 2014 and, for later issuing patents, since their issuance. This is evidenced by (1) AMD's direct knowledge of the patented technology from the 2014 meeting; (2) Defendants' knowledge of the patented technology and their infringement of it through their senior executive Mr. Hormuth; (3) the otherwise unexplained similarities between Defendants' Accused Products and their marketing and the Plaintiffs' patented technology and marketing; (4) the small number of players in the DPU/AI NIC market,[24] Plaintiffs' pioneering role in that market, and Defendants' direct competition against Plaintiffs; and (5) Defendants' intense interest in the DPU market, evidenced by its nearly $2 billion acquisition of Pensando and its subsequent focus in marketing and advertising on their DPUs, particularly for AI applications. In particular, given all of these factors, on information and belief, Defendants were aware of Plaintiffs, Plaintiffs' patents, and Defendants' infringement of them and tracked their progress and development such that Defendants were aware of Plaintiffs' patents that issued in 2020 and 2021, and their infringement of those patents.

53. Defendants have also been on notice of the Asserted Patents and their specific infringement of those patents since the filing of the original complaint, filed on March 29, 2024.

[24] The DPU market has a relatively small number of players. *See, e.g.*, https://www.datacenterknowledge.com/data-center-hardware/data-processing-units-what-are-dpus-and-why-do-you-want-them- ("Key vendors in the DPU market include NVIDIA, Marvell, Fungible (acquired by Microsoft), Broadcom, Intel, Resnics, and AMD Pensando.").

54. Defendants' infringement of the Asserted Patents 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.

VII. FIRST CLAIM
Infringement of the '596 Patent

55. Plaintiffs incorporate by reference paragraphs 1-32 herein.

56. 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.

57. 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;

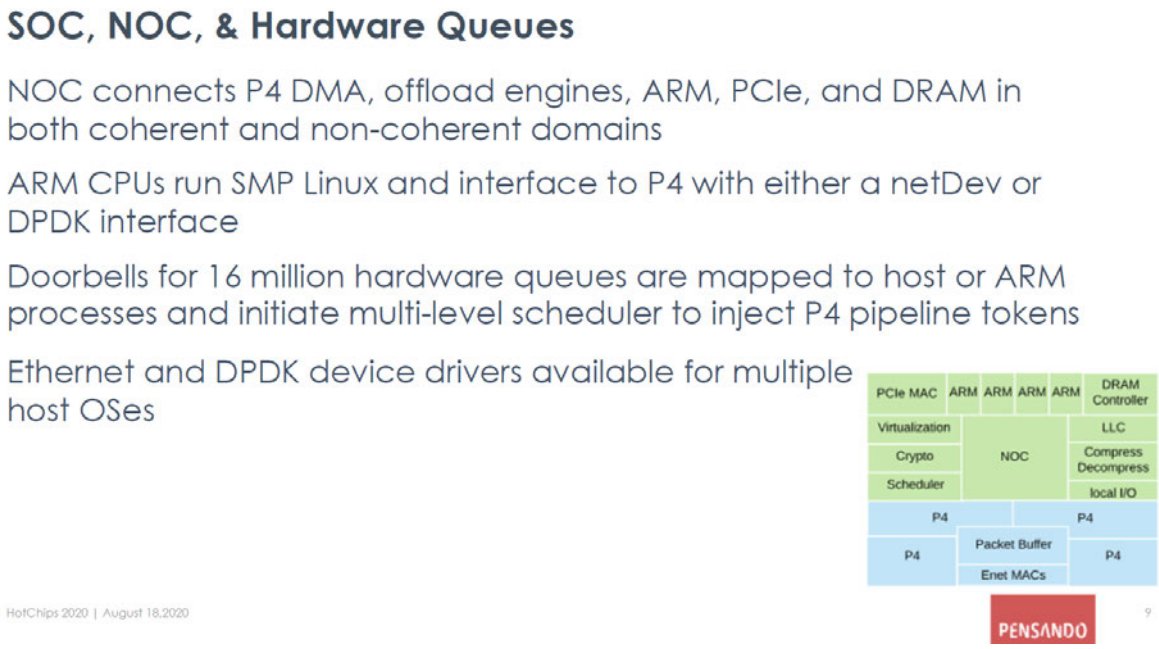
notifies said first processing element to issue said command to said input queue; and

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.

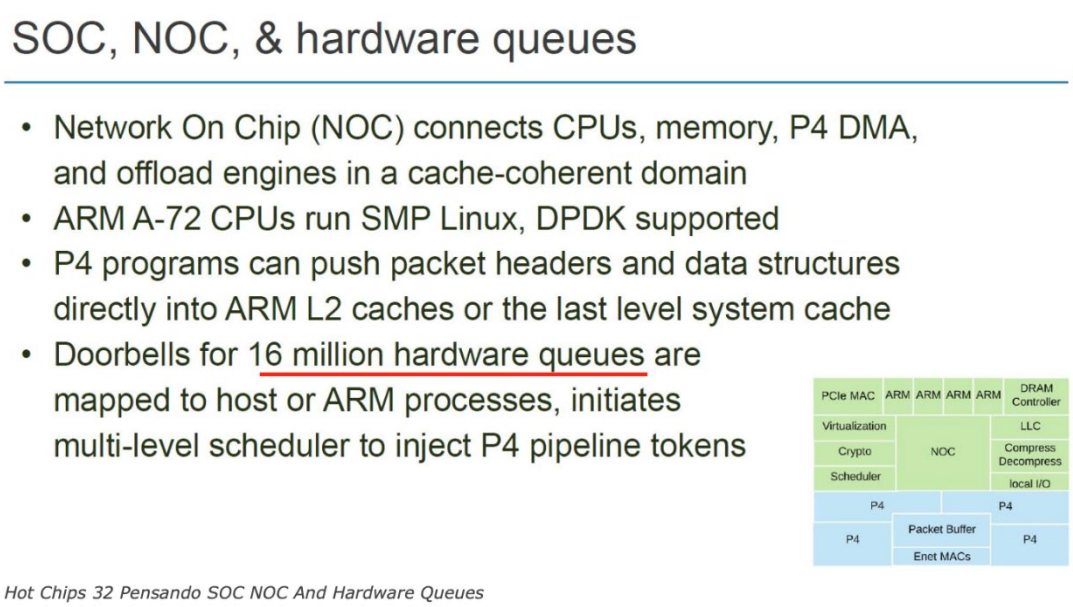
58. The Accused Products, including AMD Pensando's DPU technology and DPU-enabled systems, allow for hardware processing of specific functions on the data path with hardware queues supporting various operations implemented in hardware. The Accused Products include systems for dividing and synchronizing a processing task across a plurality of processing elements. For example, the AMD Pensando DPUs "offload infrastructure services from the computing CPU and make use of hardware accelerators to boost performance of those services" and "improve security by running security softw[a]re like distributed firewalls (on the DPU) on different cores than the workloads (on the x86 CPU)." Exhibit 33, https://www.delltechnologies.com/asset/en-us/products/networking/briefs-summaries/smartdpu-software-solutions-solution-brief.pdf; *see* https://www.amd.com/system/files/documents/pensando-dsc-200-product-brief.pdf; https://www.servethehome.com/pensando-distributed-services-architecture-smartnic; https://pensando.io/project-monterey-early-access-program (accessed May 6, 2023).

59. The Accused Products include an input queue implemented in hardware and an output queue implemented in hardware. For example, the AMD Pensando Products include numerous hardware queues, such as host or ARM processes, which initiate multi-level scheduler to inject P4 pipeline tokens. *See* https://www.servethehome.com/pensando-distributed-services-architecture-smartnic; *see also* https://hc32.hotchips.org/assets/program/conference/day2/HotChips2020_Networking_Pensando_v3.pdf. The Accused DPU products contain input queues (for example, P4 RxDMA) and output queues (for example, P4 TxDMA). *Id.* The below evidence provides exemplary evidence indicating that the input and outputs queues in Defendants' Accused Products are implemented in "hardware."

Exhibit 10 (AMD-CV-0005069) at -5080.

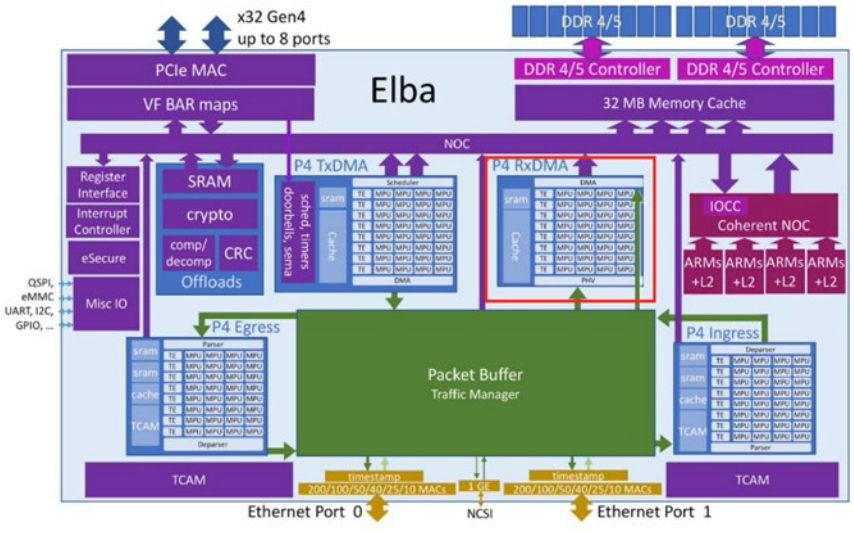


Distributed Services Architecture by Francis Matus, HotChips (2020), Exhibit 11 (CV-AMD0016381).

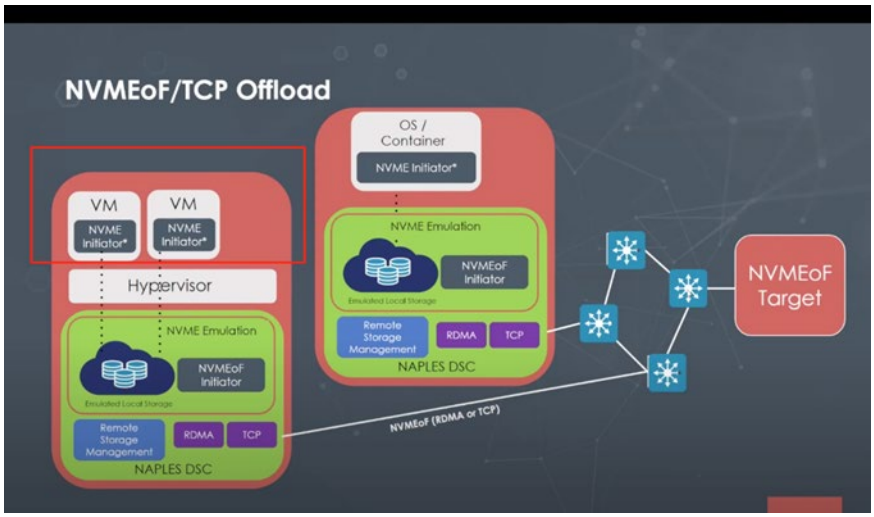


https://www.servethehome.com/pensando-distributed-services-architecture-smartnic, Exhibit 12 (CV-AMD0016001).

60. 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. As examples, hosts within the system and Virtual Machines ("VMs") running on the system containing a DPU, as well as processors within the DPU, have access to the input queue for (as an example) scheduling operations performed by parts of the DPU. *See* https://www.youtube.com/watch?v=in7athW-PaQ at 17:11.

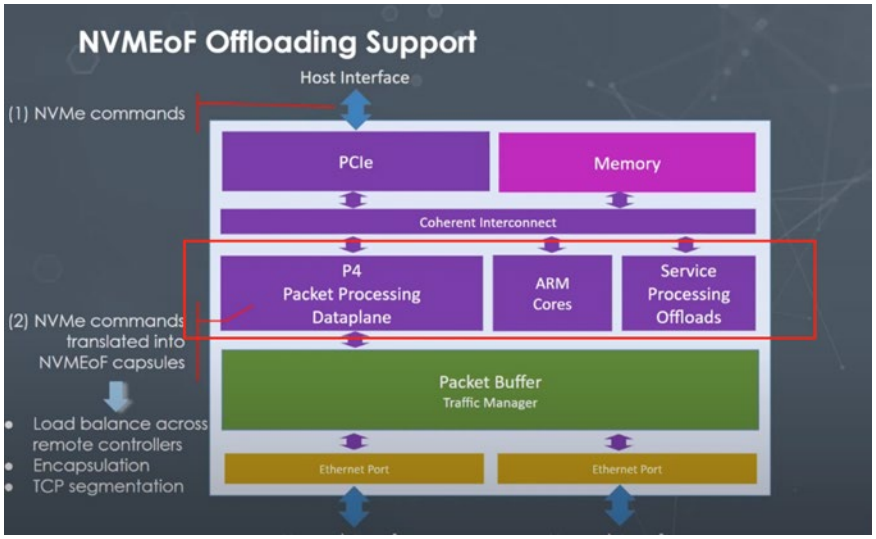


https://www.servethehome.com/pensando-distributed-services-architecture-smartnic.



https://www.youtube.com/watch?v=in7athW-PaQ at 17:11.

61. The Accused Products include at least one second processing element in communication with said first processing element, for example a second processing element connected via internal interconnects and a PCIe interface to the first processing element. *See* https://www.youtube.com/watch?v=in7athW-PaQ at 18:02.

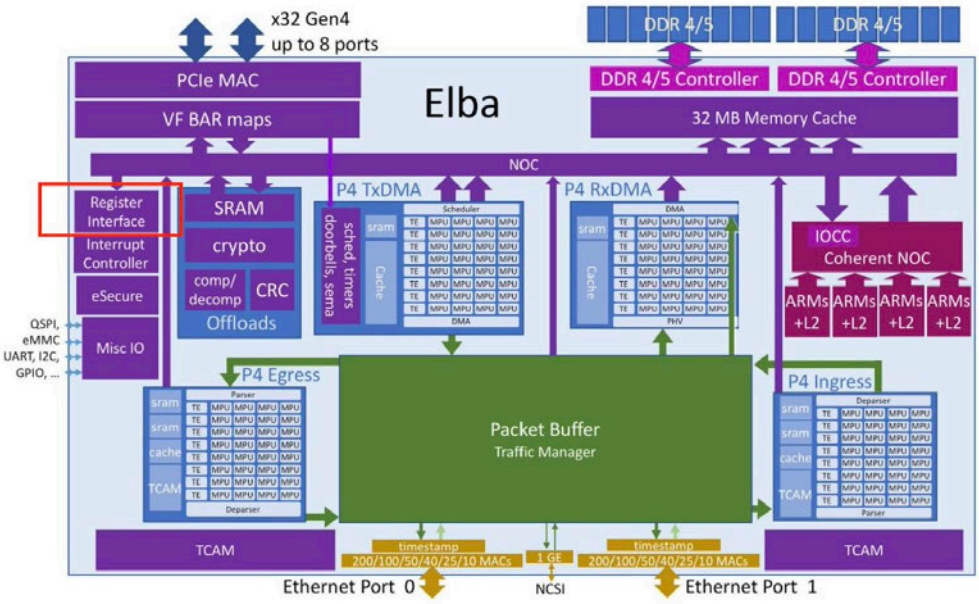


https://www.youtube.com/watch?v=in7athW-PaQ at 18:02.

62. 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. *See* https://www.servethehome.com/pensando-distributed-

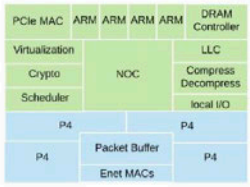
services-architecture-smartnic. For example, the doorbells for the hardware queues initiate the scheduler to inject P4 pipeline tokens and this doorbell mechanism allows objects to be added to any queue. *Id.* For another example, hardware semaphores are claimed reservation registers implemented in hardware. *See, e.g.*, documents produced by Defendants in this litigation, Exhibit 13 (AMD-CV-0001155) at -1194; Exhibit 14 (AMD-CV-0003945) at -4083; Exhibit 15 (AMD-CV-0007010) at -024. As shown in the cited evidence, these examples of the infringing "reservation register[s]" are [redacted] Exhibit 16 (AMD-CV-0031160).

Exhibit 16 (AMD-CV-0031160).



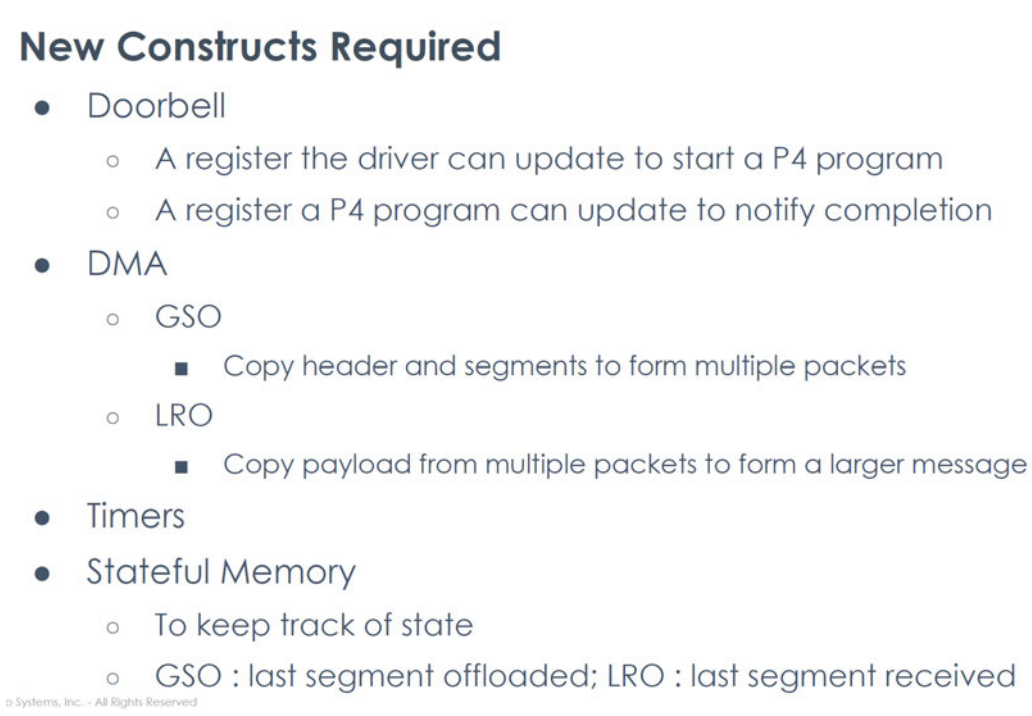
SOC, NOC, & hardware queues

- Network On Chip (NOC) connects CPUs, memory, P4 DMA, and offload engines in a cache-coherent domain
- ARM A-72 CPUs run SMP Linux, DPDK supported
- P4 programs can push packet headers and data structures directly into ARM L2 caches or the last level system cache
- Doorbells for 16 million hardware queues are mapped to host or ARM processes, initiates multi-level scheduler to inject P4 pipeline tokens



Hot Chips 32 Pensando SOC NOC And Hardware Queues

https://www.servethehome.com/pensando-distributed-services-architecture-smartnic.



https://opennetworking.org/wp-content/uploads/2020/04/Plenary-4-Slide-Deck.pdf.

Hardware Queue Management

Packets which enter the P4 pipeline form the wire or which originate from internal events are placed in hardware queues. Hardware queues are used to manage interfaces, flows, connections, and any other objects which require ordered tracking and scheduling. Software can configure up to 16 million hardware queues (Figure 5). Each queue stores its current state in a DRAM-based qstate record, which includes a count of enqueued objects, pointers to arrays or linked lists of enqueued objects, connection state and peer information, and a process ID used for memory protection.

http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf, p. 46.

Key Features

Feature	Description
Network Interfaces	56 Gbps PAM4 SerDes supporting Dual 200 GE, Quad-100/50/25/10GE 1GE management port
PCIe Interface	32 Lanes of PCIe Ge n4, configurable as root complex or end-point mode 2x16 / 4x8 / 8x4 with QOS support in multi-host applications
Data Pipeline	P4 data pipeline comprising 144 match processing units (MPUs) @2 GHz Provides high performance capabilities in packet and message processing, at line rate
SOC	16x Arm A72 CPU cores @3 GHz with I-Cache, D-Cache and LLC Cache QSPI Flash, EMMC storage for embedded OS Secure Boot with hardware root of trust
Memory	Dual DDR4-3200 interfaces supporting 8 GB - 64 GB system memory
Offloads	Inline IPsec and DTLS, Bulk Crypto, PKE, Compression, Decompression, Checksums, Deduplication, Erasure Coding
Scale	2K VNICs, 16 M hardware queues, highly scalable P4 tables (stateful and stateless) accessible at every stage of the pipeline
Scheduling	Queue Group scheduling with Min/Max rate High Priority Option per queue group P4 meters and QoS priorities
ROCEv2	Memory based scatter/gather Lists (SGL) in DMA commands Latency optimized hardware data path with all context on DPU
Single Wire Management	Connects 1GE BMC controller and Uplink ports in Standby Power mode

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/product-briefs/pensando-elba-product-brief.pdf; Exhibit 27 (AMD-CV-0004687), AMD Pensando Giglio Product Brief.

63. 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 31, https://www.amd.com/system/files/documents/pensando-dsc-200-product-brief.pdf; http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf. The Accused Products determine when said read value indicates available space in said input queue for said first processing element to issue a command. https://www.servethehome.com/pensando-distributed-services-architecture-smartnic; http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf. With the Accused Products, the second processing element notifies said first processing element to issue said command to said input

queue. https://www.youtube.com/watch?v=in7athW-PaQ; https://www.principledtechnologies.com/Pensando/DSC-200-performance-0122.pdf. For example, once the DPU determines available space, the first processing element issues a command utilizing the scheduler and the reservation register.

64. 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. https://www.servethehome.com/pensando-distributed-services-architecture-smartnic; https://www.principledtechnologies.com/Pensando/DSC-200-performance-0122.pdf; http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf; https://opennetworking.org/wp-content/uploads/2020/04/Mario-Baldi-Slide-Deck.pdf.

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

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

67. 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.

68. Defendants 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.

69. 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. SECOND CLAIM
(Infringement of the '753 Patent)

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

71. 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.

72. 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;

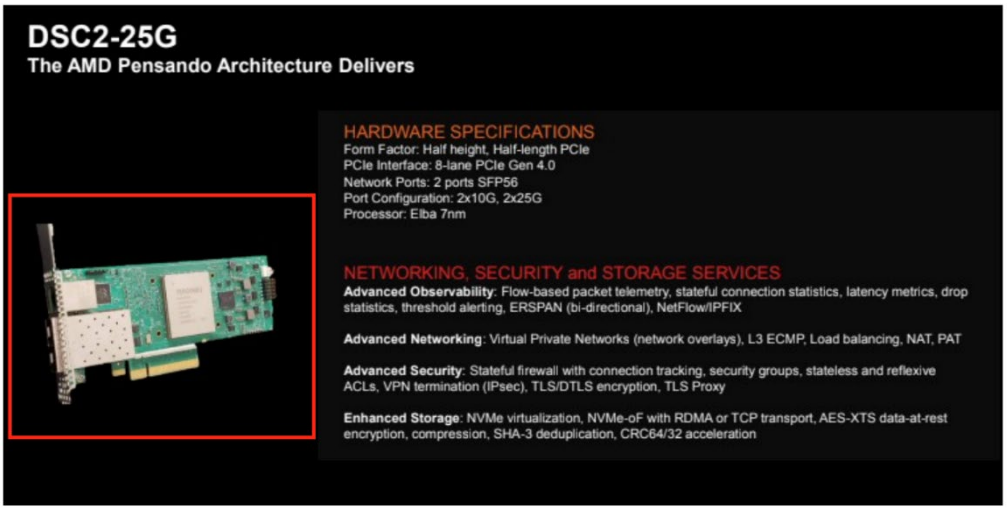
one or more of the plurality of swappable pods or cards having a portion for user-definable hardware modules or user-definable software modules;

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.

73. 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 by, for example, supporting the P4 programming language.



AMD Pensando DSC2 25G

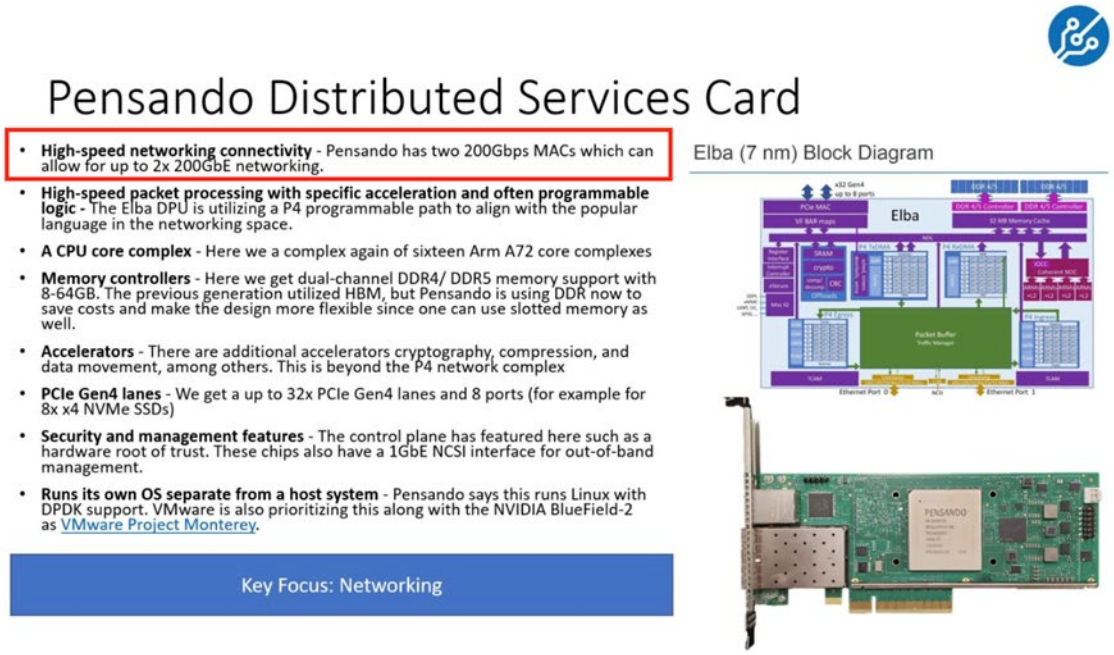
https://www.servethehome.com/amd-pensando-giglio-dpu-for-2023-salina-dpu-in-2024-and-amd-epyc; *see* http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf, p. 44 (stating that "Pensando ASICs include multiple P4 pipelines."); https://opennetworking.org/news-and-events/blog/pensando-announces-p4-programmable-platform-and-joins-p4-community.

74. The Accused Products include swappable pods or cards in, for example, the Pensando DSC which has hardware modules and software modules and, for example, the Pensando DSC contains user-definable hardware (via P4) modules and user-definable software modules (via ARM cores). https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible; Exhibit 31, https://www.amd.com/system/files/documents/pensando-dsc-200-product-brief.pdf.

75. The Accused Products are designed to have portions that are user-configurable to implement data flow processing architectures such as, for example, a SmartSwitch.

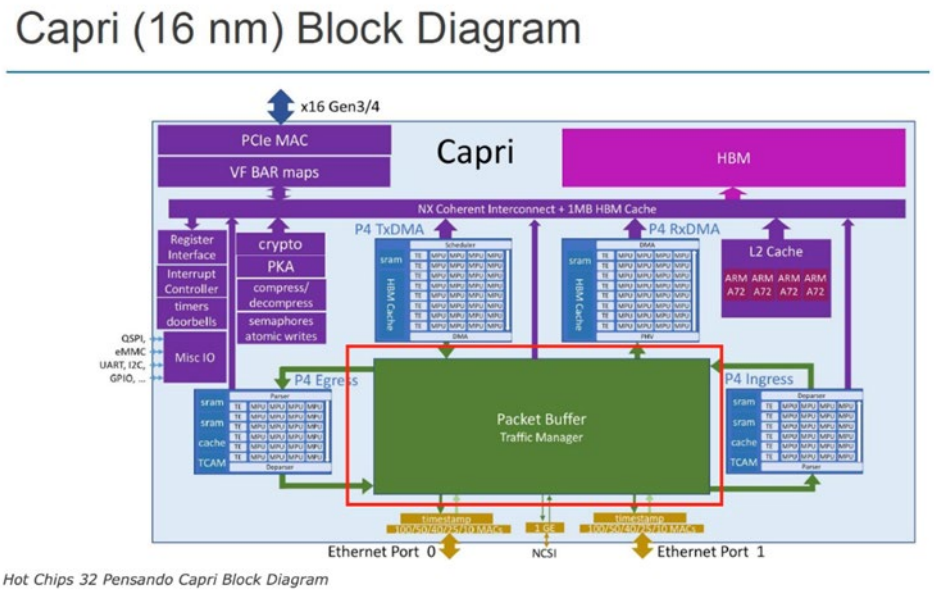
https://www.theregister.com/2022/04/07/amd_pensando_aws; https://www.networkworld.com/article/3690334/aruba-to-prioritize-sase-private-5g-data-center-networking.html.

76. The Accused Products include a network coupled to the one or more swappable pods or cards.



Pensando DSC DPU Example Q2 2021

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible.



Hot Chips 32 Pensando Capri Block Diagram

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible; *see* Exhibit 31, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf.

77. The Accused Products support 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, specifically, the Pensando DPUs are programmable using P4 which allows for the customization of how a network device processes packets. https://packetpushers.net/aruba-puts-dpus-into-new-top-of-rack-switch-5-questions; Exhibit 32, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf; Exhibit 29, https://www.amd.com/system/files/documents/pensando-project-monterey.pdf. Defendants' MAC address of DSS (distributed services switch), Unit ID for DSM (distributed services module), DSM Internal IP, VLAN/VNIC ID, the "devicename," "serialnumber," and "unitid" are examples of chassis and board identifiers. Further, the headers that allow packets to be directed from P4 pipelines and the packet buffer to other parts of the P4 pipelines, ARM Cores, and accelerators are the claimed instance, type, and module identifiers (e.g., tm_iq, tm_oq, qid, qtype, tm_span_session, app_type). *See, e.g.*, Exhibit 19 (AMD-CV-0000935) (listing exemplary claimed headers); Exhibit 13 (AMD-CV-0001155) at -[redacted]); Exhibit 10 (AMD-CV-0005069); Exhibit 20 (AMD-CV-0000608) at -620-21.

The Key Difference in Architecture: The Pensando Flow-Based Engine

We have previously mentioned that:

SmartSwitches have a different forwarding mechanism that is stateful. Each packet between host A and host B is processed as a part of a session between A and B. The session comprises two unidirectional flows, one from A to B and one from B to A.

Sessions and flows are identified using an extended five-tuple that includes the basic five fields (IP source and destination addresses, protocol type, TCP/UDP source destination ports) and other information like input interface, MAC addresses, TCP flags, etc.

When you build hardware capable of implementing this new forwarding mechanism at wire speed, many new features become possible.

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/white-papers/pensando-smartswitches.pdf.

multicast or span replication; then enter the P4 egress pipeline to apply, for example, network address translation, policers, and telemetry. Having completed network processing tasks, the packet can be sent back out an Ethernet MAC port or it can be passed up toward the ARM subsystem or local host.

If the packet is sent to the embedded ARM or local host, it must be processed according to the needs of the kernel or user level driver running on that CPU. The packet passes through a third P4 pipeline, where it runs another program in accordance with the protocol type of the packet, i.e., Ethernet, RDMA, NVMe, or a new protocol. A significant advantage of running protocol-specific P4 programs associated with a targeted kernel or user driver is the ability to customize descriptor formats, completion events, interrupt schemes, and data formats to the needs of the targeted driver.

In summary, network packets enter the ASIC, pass through P4 ingress, and P4 egress pipelines to apply security and network services, are associated with a target interface and device type, and pass through P4 driver interface programs. The final packet or data buffer are delivered to the attached host, embedded ARM, or integrated offload engine. Packets sent by the host enter through the PCIe MAC to be processed by the P4 driver interface programs, then by the P4 ingress and egress pipelines to apply security and network services before being sent out through the Ethernet MAC. The following section takes a closer look at the P4 pipeline internals.

P4 Pipelines

Pensando ASICs include multiple P4[3] pipelines (Figure 2) as the centerpiece of a domain-specific processing architecture. Each P4 pipeline starts with a programmable parser. At its core, the programmable parser

FIGURE 2. Single P4 pipeline with 8 stages.

includes multiple engines designed to extract header fields and populate a packet header vector (PHV) at rates of 100 million packets per second or more. Following the parser are six to eight match action stages, each stage able to load multiple tables per packet and perform complex actions which modify the PHV and update memory data structures. A single P4 pipeline includes local TCAM and SRAM resources where high bandwidth tables can be stored. Large scale tables as well as tables shared between multiple pipelines are stored in attached DRAM or host memory. The table base address steers table requests to SRAM or to DRAM, with per-pipeline caches reducing DRAM traffic for commonly accessed table entries. The P4 pipeline concludes with a deparser to reassemble the packet from the updated PHV fields. Alternatively, the pipeline can conclude with a DMA engine to directly copy packet header, payload, and meta data structures to and from local or host memory for CPU or offload engine processing.

P4 Stages

Each P4 stage (Figure 3) within a pipeline begins with a table engine (TE). The TE extracts bits and bytes from the PHV in any combination to build a table key. Table keys constructed from multiple header fields can be up to 512 bits in width, or keys can be chained together to create extra wide keys up to 2048 bits in width. The resulting key may be hashed or used as a direct index, then matched against any data structure in TCAM, SRAM, DRAM, or attached host memory.

The table data and lookup result (match or no match) are then forwarded to a match processing unit

http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf, p. 44.

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

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

80. 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.

81. 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.

82. 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. THIRD CLAIM
(Infringement of the '943 Patent)

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

84. 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.

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

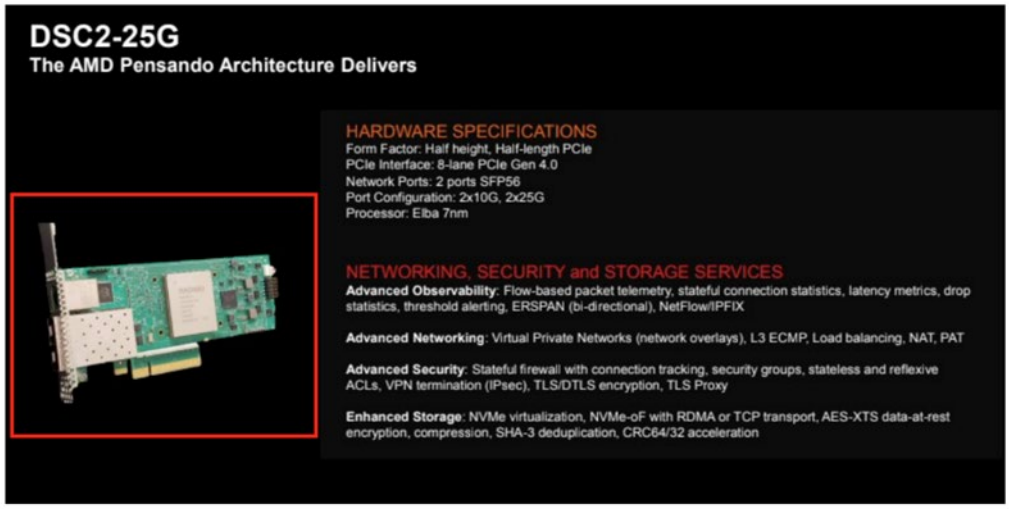
1. A programmable device for data flow processing in a user-configurable server with swappable pods or cards, comprising:

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.

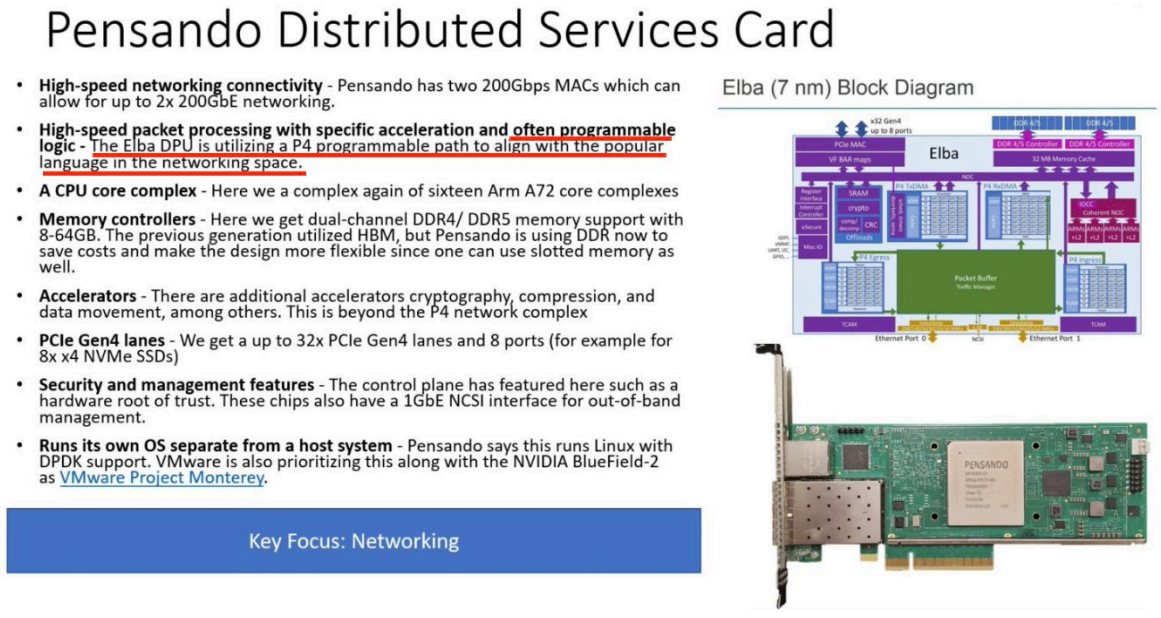
86. The Accused Products include a programmable device for data flow processing in a user-configurable server with swappable pods or cards, for example, supporting the P4 programming language.



AMD Pensando DSC2 25G

https://www.servethehome.com/amd-pensando-giglio-dpu-for-2023-salina-dpu-in-2024-and-amd-epyc; http://gibsonnet.net/aix/ibm/mmi202102.issue.pdf (stating that "Pensando ASICs include multiple P4 pipelines.").

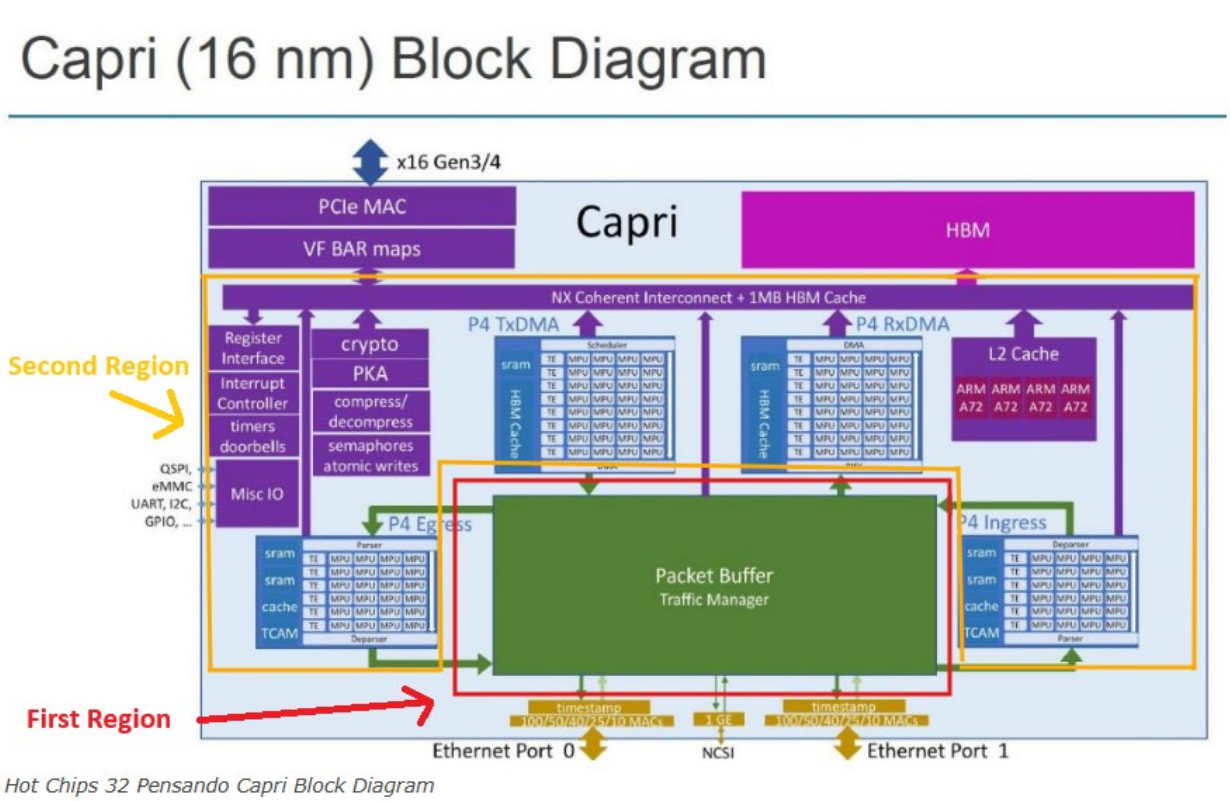
87. The Accused Products contain a programmable logic device ("PLD") (e.g., the device made up of and including at least the P4 blocks, packet buffer, ARM cores, and related portions) having a first region and a second region. For example, the AMD Pensando DSC is a programmable device. https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible; Exhibit 28, https://www.amd.com/system/files/documents/aruba-cx-10000-series-datasheet.pdf; https://www.amd.com/en/solutions/infrastructure-acceleration. The AMD Pensando DSCs contain programmable logic. In particular, the AMD Pensando DSC includes portions that are programmable, including the P4 blocks, using P4 pipelines to accelerate and customize cloud infrastructure services.



Pensando DSC DPU Example Q2 2021

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible/, Exhibit 21 (CV-AMD0016146) at -16147.

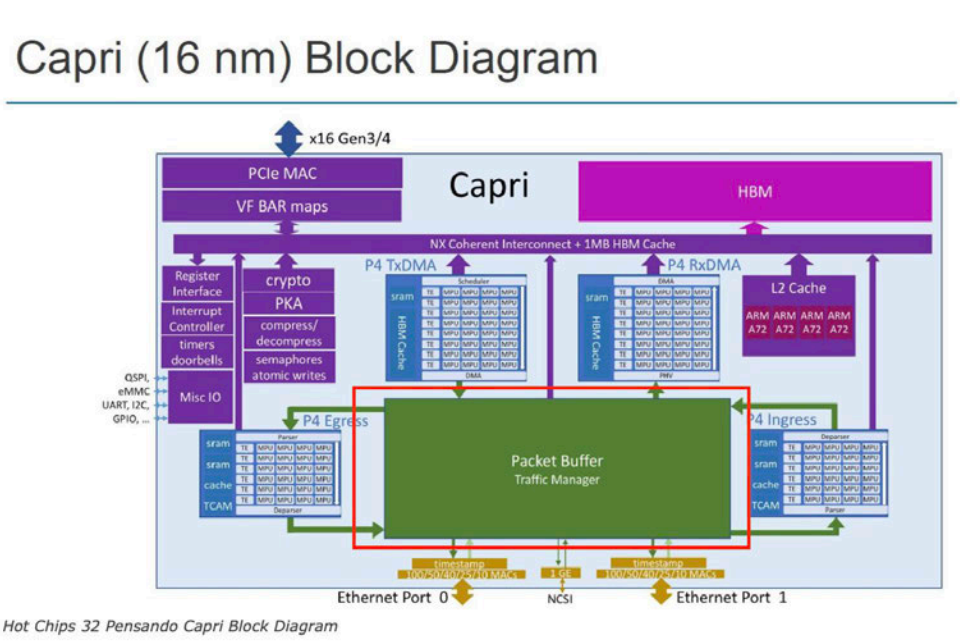
88. For example, the Pensando DSC contains the following exemplary first and second regions.



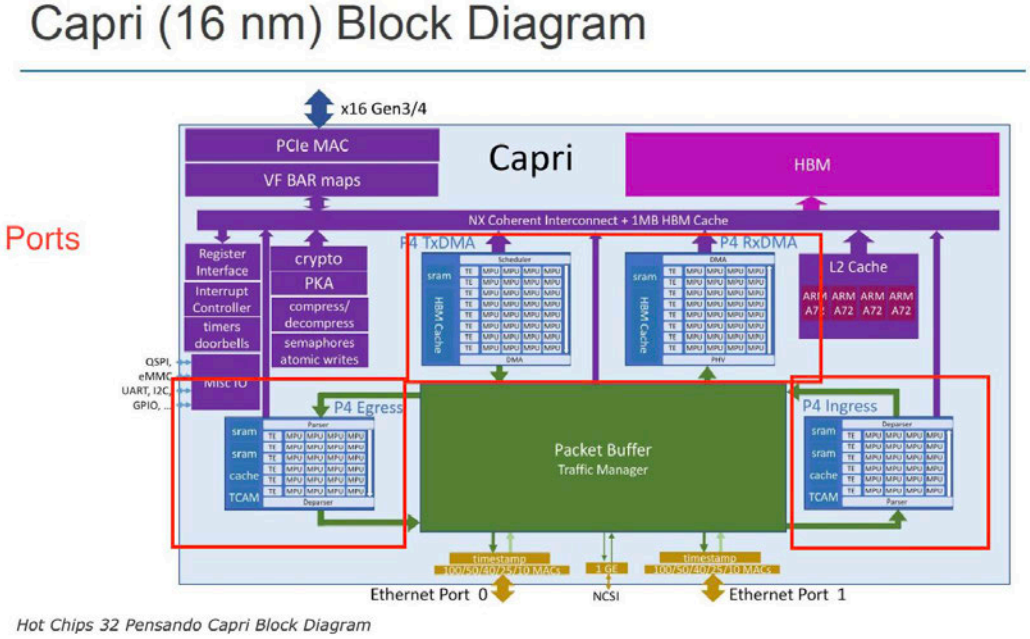
Hot Chips 32 Pensando Capri Block Diagram

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible.

89. The 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 includes a packet buffer traffic manager for routing data between the network and the second region.



Hot Chips 32 Pensando Capri Block Diagram

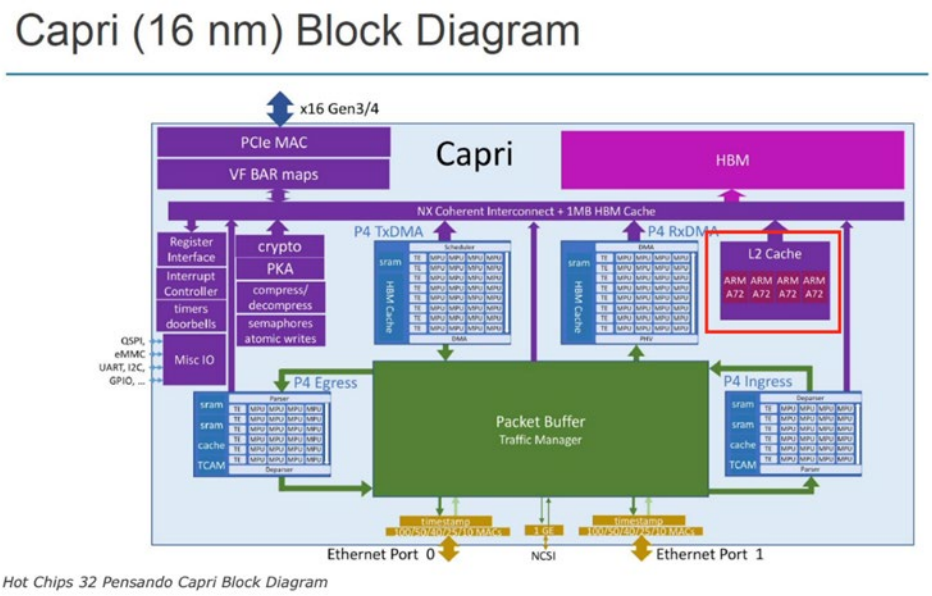


Hot Chips 32 Pensando Capri Block Diagram

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible.

90. The second region comprises one or more sandboxes with user-definable programmable electronic circuits of the PLD. For example, the Accused Produces have one or more sandboxes with user-definable programmable electronic circuits of the PLD. For example, the second region of the Accused DSC contains the P4 programmable components and the ARM CPU cores.

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible.



https://www.servethehome.com/amd-pensando-giglio-dpu-for-2023-salina-dpu-in-2024-and-amd-epyc.

91. 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. In particular, the Accused DSC has a port that 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, each port between the packet buffer and the P4 blocks are bridges that can be locked or unlocked to prevent or enable access to the P4 sandboxes. For another example, in the Accused DSC, AXI filters guard paths from Arm control-plane to P4 pipeline stages (sandboxes) and Ethernet/PCIe ports wherein the locked rules prevent tenant/user access to port registers or sandbox and similarly, unlocked state enables programming via SSDK APIs. In general, AXI filters are hardware blocks inserted inline on AXI bus paths to monitor, validate, and selectively block transactions from masters to slaves based on programmable rules, enforcing security and access control in SoC designs like Pensando DPUs.

Exhibit 14 (AMD-CV-0003945) at -3979.

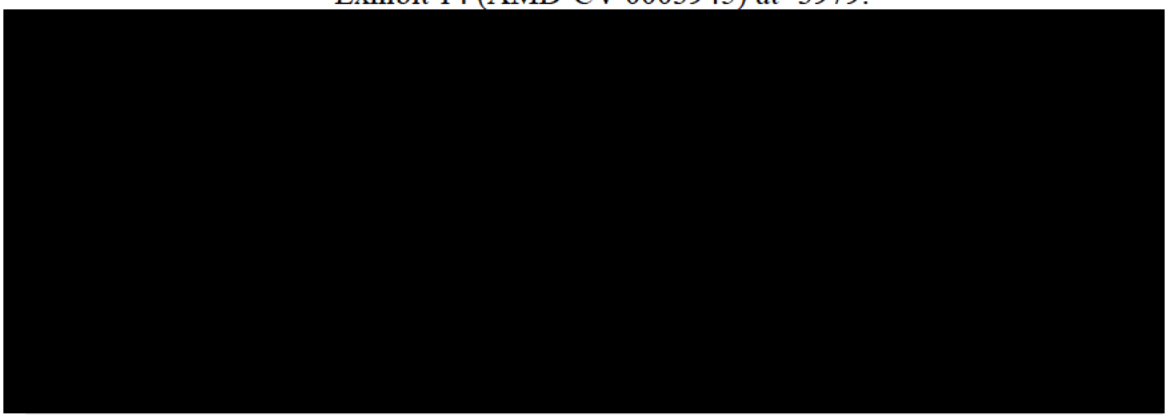


Exhibit 14 (AMD-CV-0003945) at -3979.

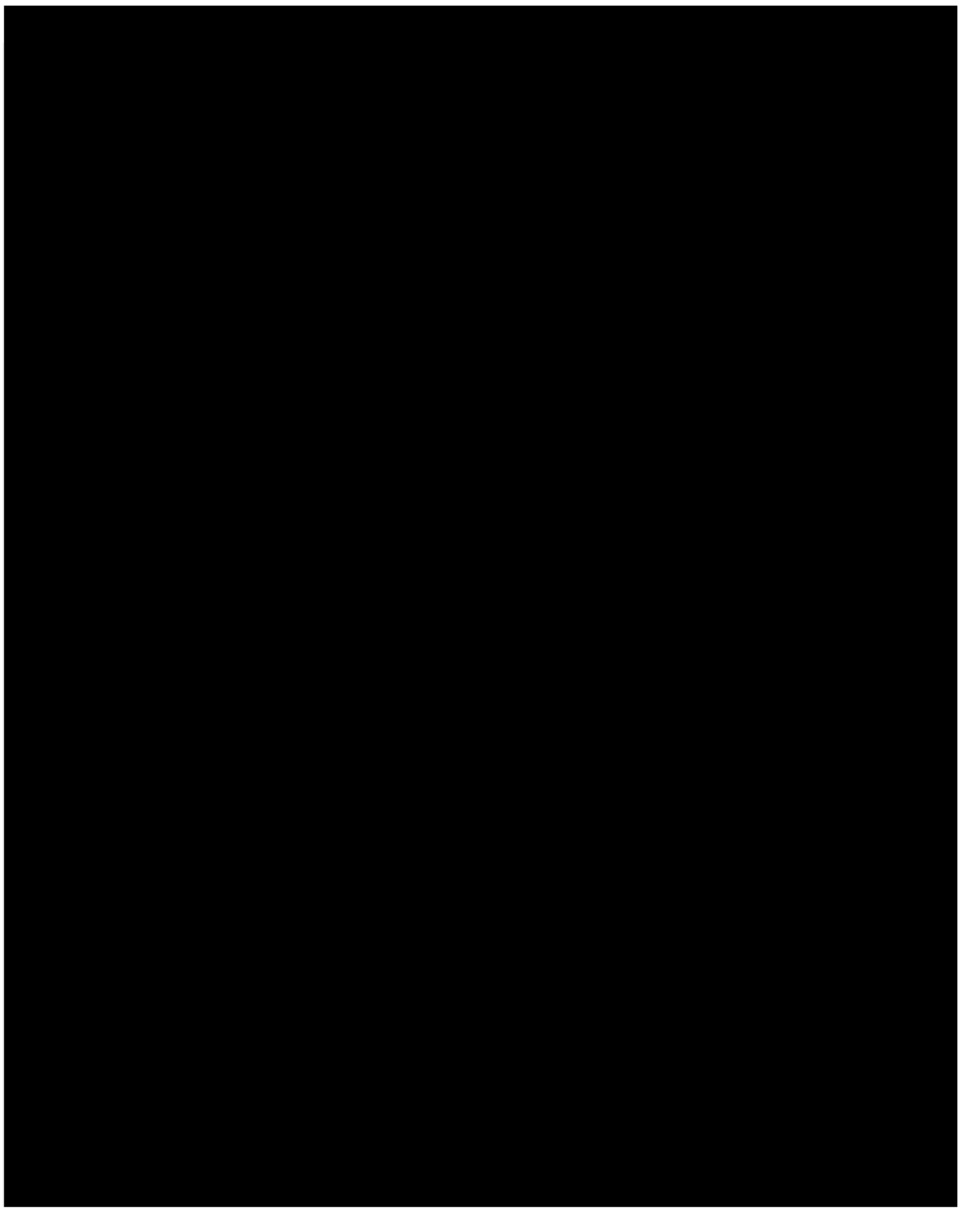
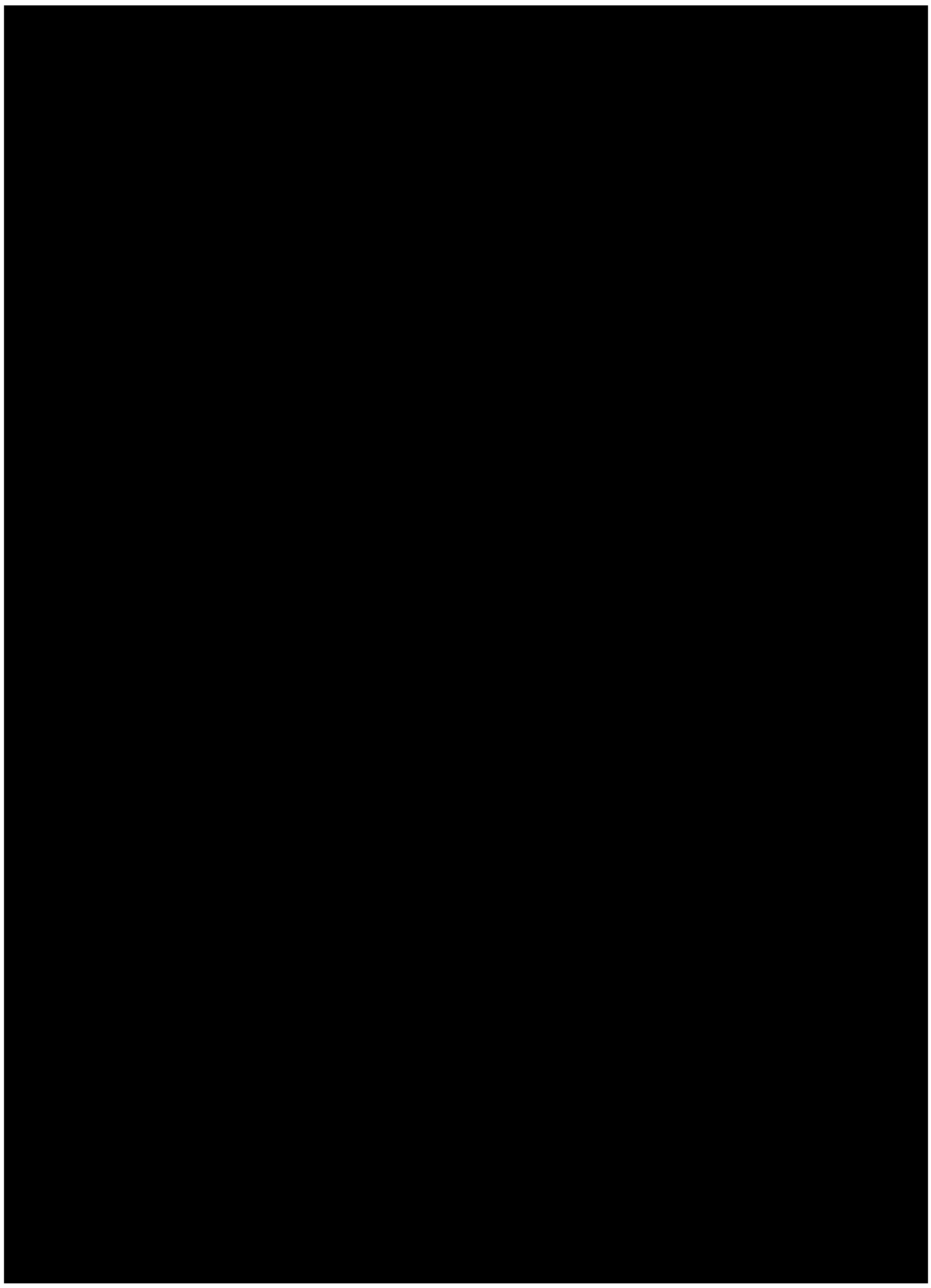
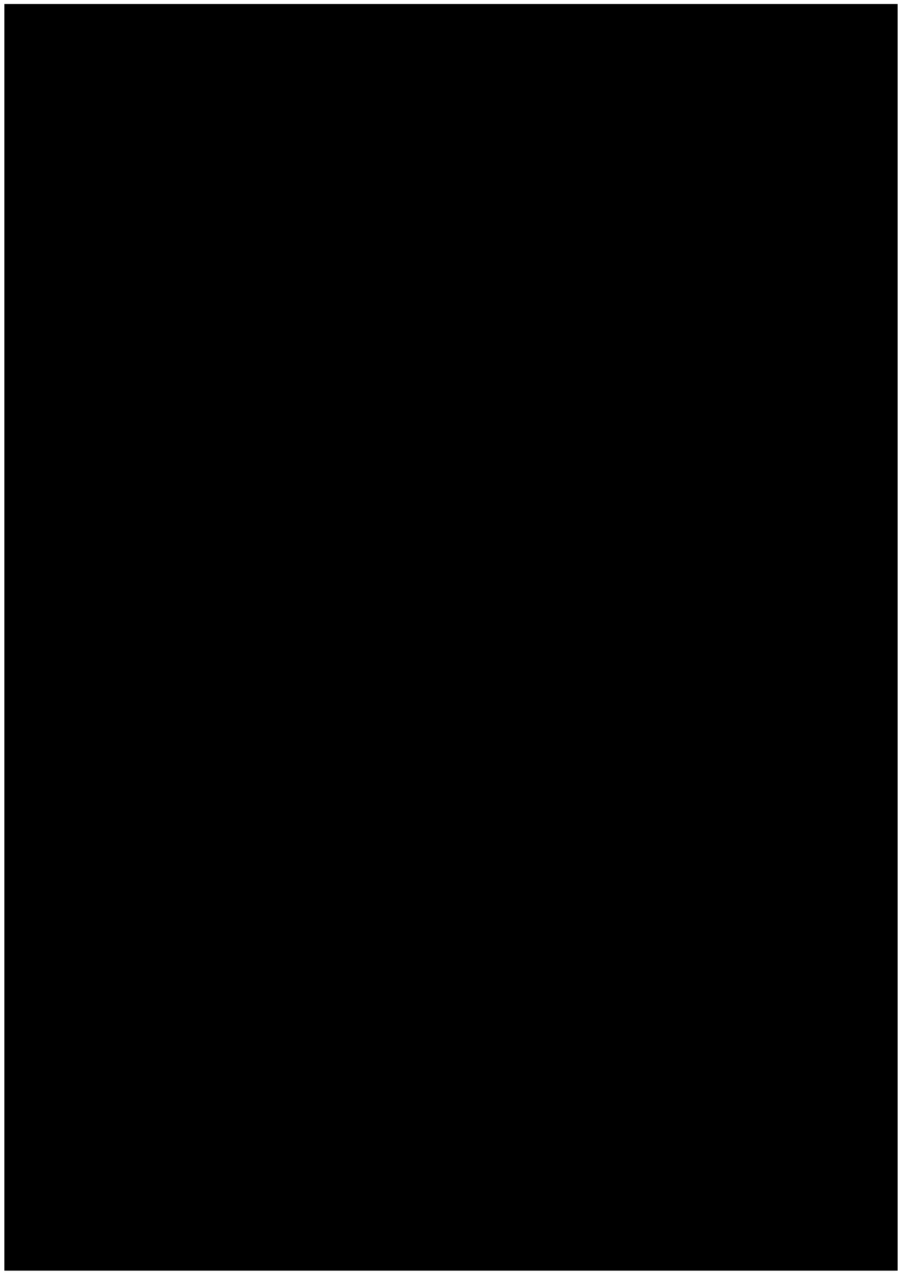
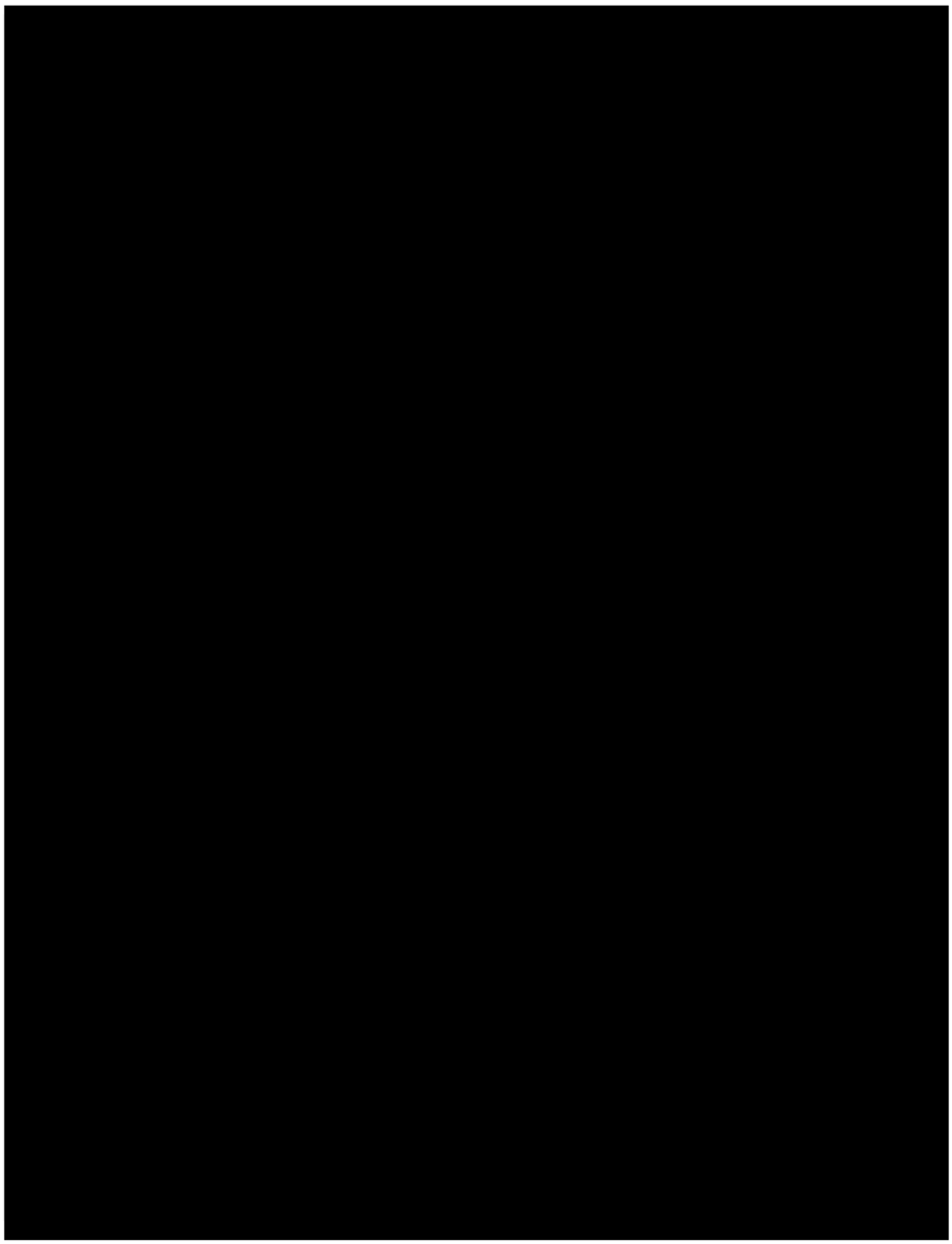
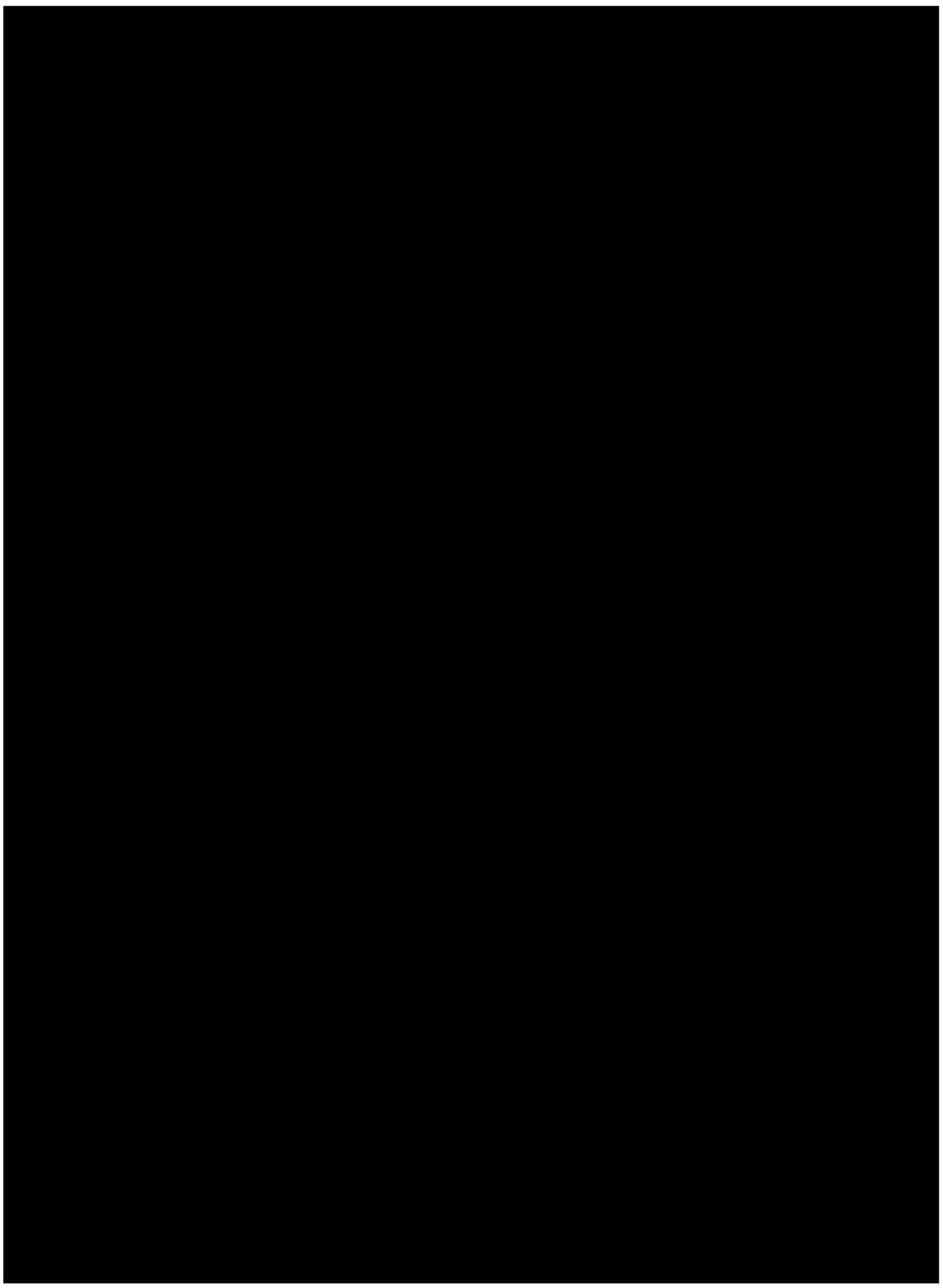


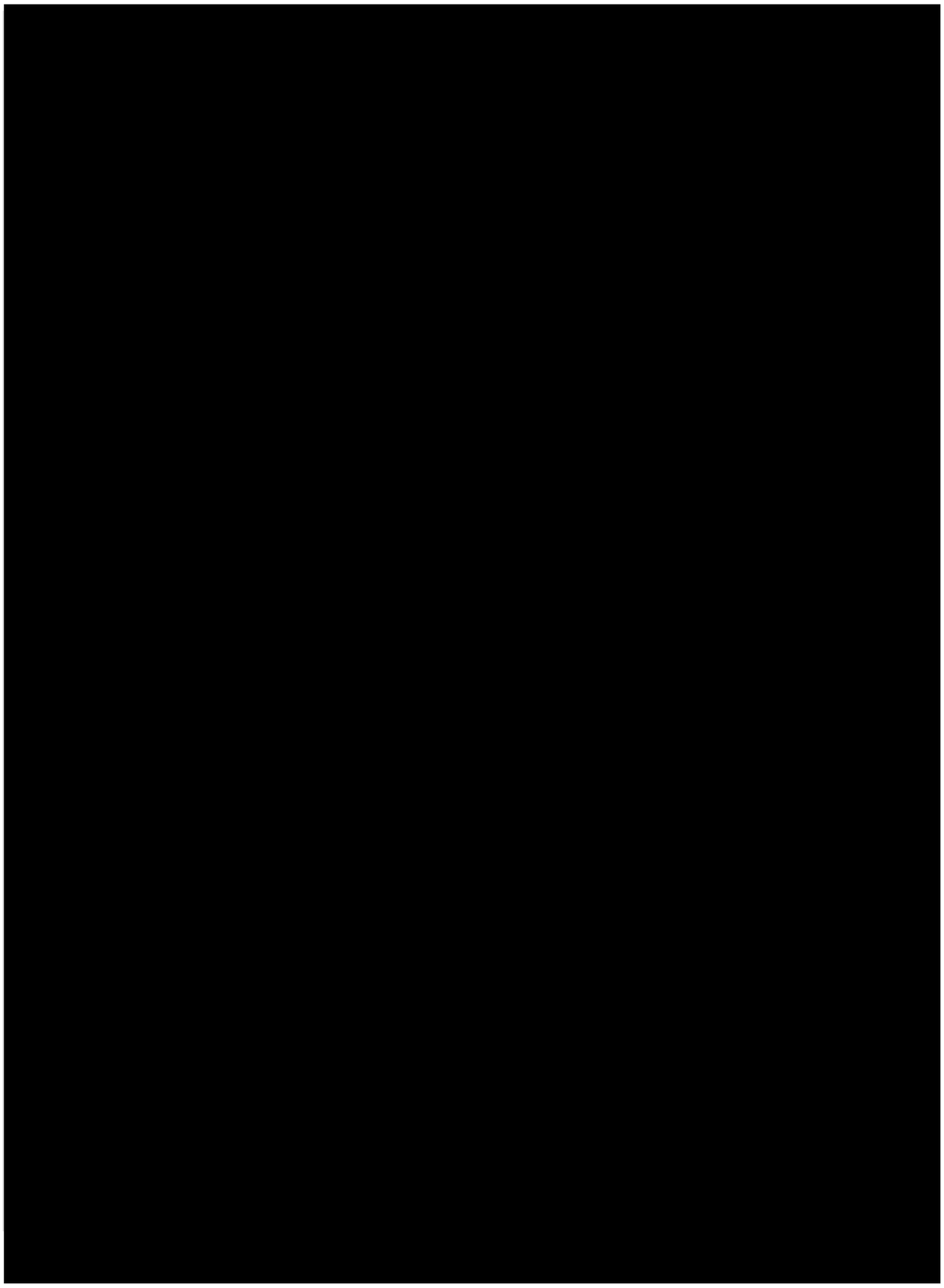
Exhibit 14 (AMD-CV-0003945) at -3981.











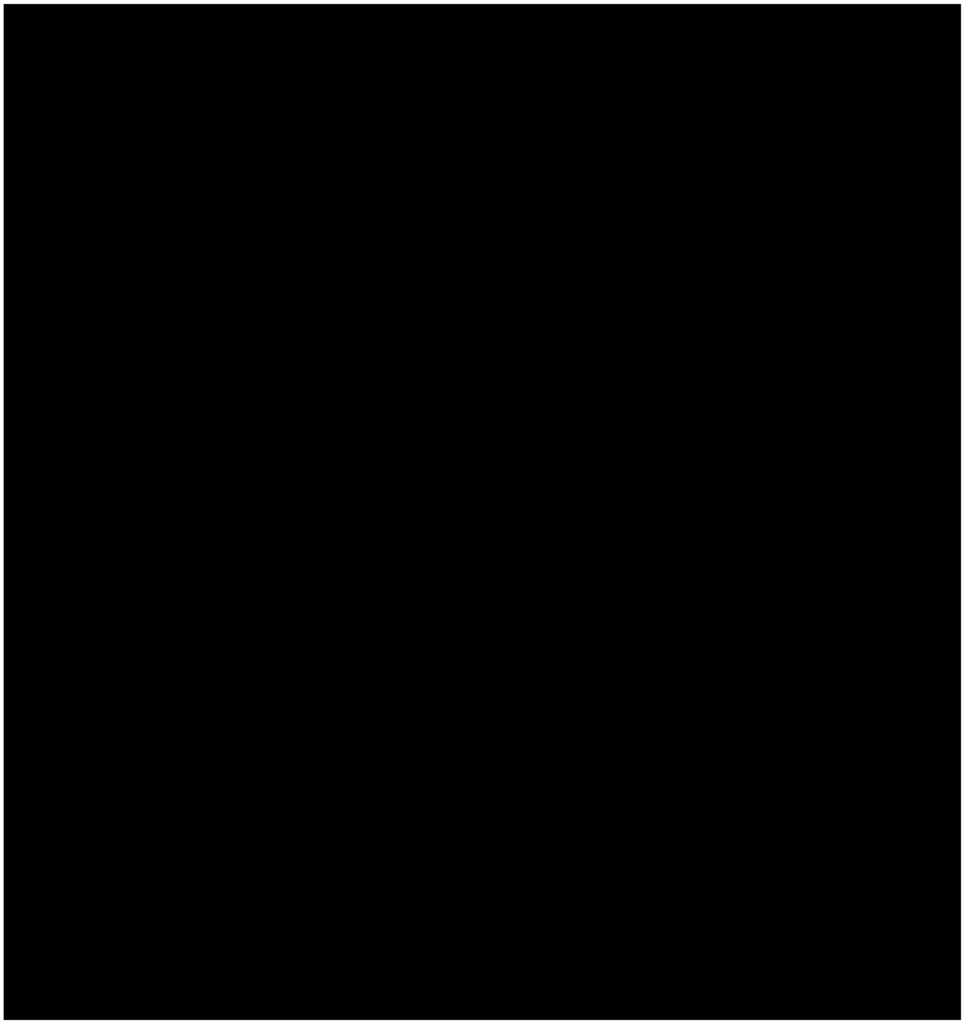
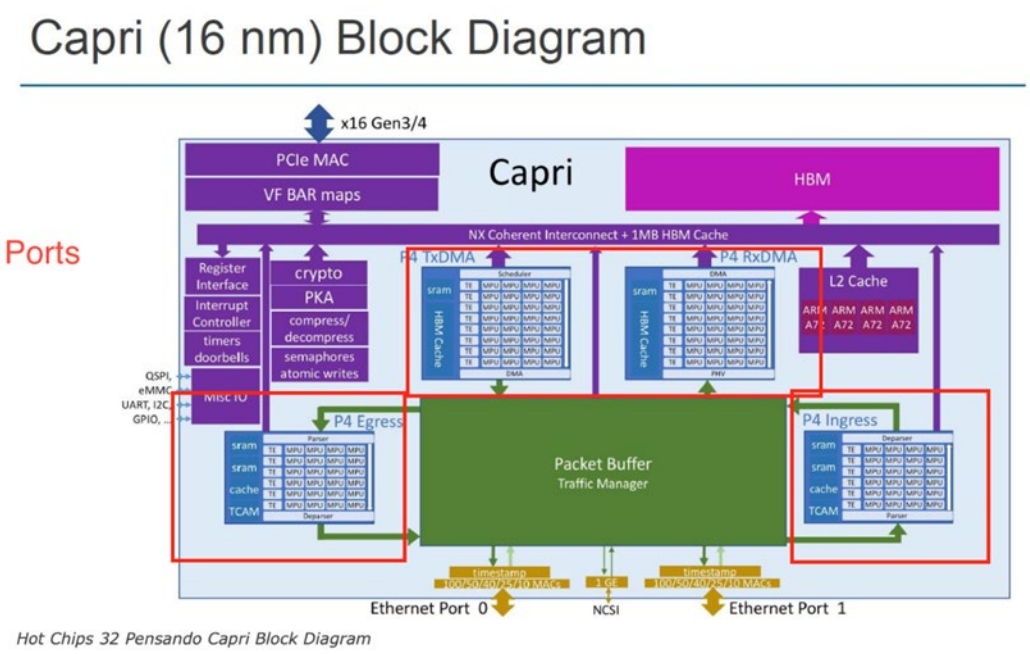


Exhibit 14 (AMD-CV-0003945) at -4085-91.

Exhibit 22 (AMD-CV-0012020) at -12023.

Exhibit 22 (AMD-CV-0012020) at -12027.

Exhibit 22 (AMD-CV-0012020) at -12028.



Hot Chips 32 Pensando Capri Block Diagram

https://www.servethehome.com/amd-acquires-pensando-for-its-dpu-future-intel-ipu-nvidia-bluefield-marvell-octeon-fungible.

Encryption Services Architecture

The IPsec data path as well as handshake and key management elements are implemented within the DSC, assuring an extremely high level of security, as both keys and sensitive encryption policies are maintained inside the secure perimeter of the DSC ASIC. The PenAccel™ subsystem implements high-performance hardware-based crypto acceleration in the DSC data path to provide IPsec encryption services at line speed (25Gbps or 100Gbps). The on-chip ARM subsystem runs the crypto handshake and key management control plane for the associated crypto protocol (e.g. TLS handshake or IKE for IPsec).

The default ciphersuite for IPsec ESP operation is AES-GCM-256, providing very high security, CNSA-compliant encryption and payload integrity and authentication.

In order to support very high crypto connection rates, hardware acceleration of both public key and symmetric encryption algorithms is available to the ARM subsystem, including RSA, Elliptic Curve, and Diffie Hellman. In addition, a high-grade entropy-based random number generator, compliant with NIST SP 800-90A/B, is used.

Platform Security

The *PenTrust* subsystem is the Root of Trust (RoT) for the Pensando Programmable Services Processor. PenTrust is the first subsystem to boot at power-on or hardware reset from an immutable embedded ROM, making it the first link in the Secure Boot Chain of Trust. The firmware for the PenTrust subsystem is burned in ROM during ASIC manufacturing and cannot be modified or tampered with once it is programmed. Together with a chip-unique key and identity, the PenTrust subsystem validates all code that runs on the DSC with a digital signature.

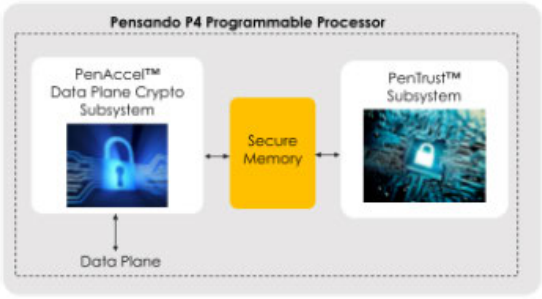


Exhibit 6, Pensando IPsec Solutions, Pensando White Paper (2021); *see* https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/user-guides/pensando-ssdk-ipsec-gw-user-guide.pdf.

Segmentation is key to preventing unwanted lateral movement, by inspecting all East-West traffic in the data center and applying policies that stop bad actors from moving through the internal network. Several approaches have been tried to achieve segmentation in the past, but with very limited success:

- Hardware **next-generation firewall** appliance-based segmentation
- **Virtualized firewall appliance**-based segmentation
- **Software agent**-based segmentation
- **Network switch** (stateless ACL)-based segmentation

Exhibit 30 (AMD-CV-0004848), https://www.amd.com/system/files/documents/zero-trust-building-secure-data-centers-cx-10000.pdf.

PenTrust Subsystem

Internal to Pensando's Capri silicon at the heart of each Distributed Services Card (DSC) is a security element called the PenTrust Subsystem that is responsible for the fundamental protection mechanisms in the device. The module is isolated from the rest of the chip and has its own CPU, ROM, RAM, and cryptographic engines, similar to Trusted Platform Modules (TPM) used in server systems. *PenTrust* accesses chip resources outside of its secure perimeter via its own bus mastering DMA engine, and it can receive requests (e.g. to create keys or sign certificates) via a narrow, secure-access interface. Other modules can deposit requests and asynchronously pick up responses as they become available.

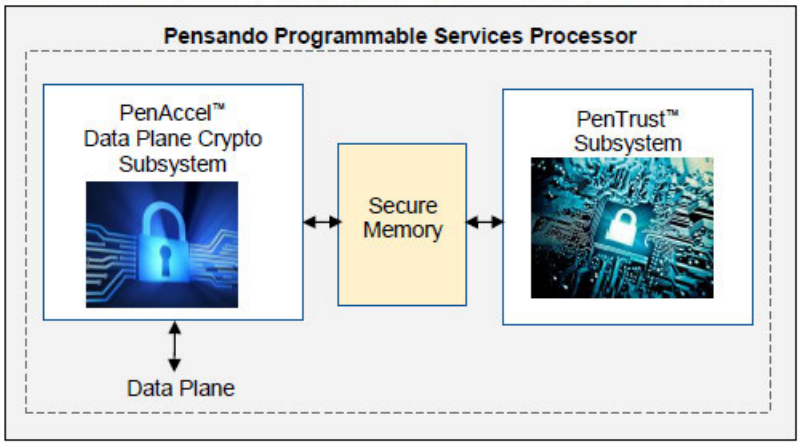


Fig 2. Capri P4 Programmable Processor

The *PenTrust* subsystem provides trusted crypto services such as asymmetric (public key) sign and verify operations, as well as symmetric (secret key) encryption/decryption and secure hashing operations. It also enables secure key store and management via primitives to generate keys, wrap/unwrap keys, and import/export them in encrypted form.

Root of Trust

The *PenTrust* subsystem is the Root-of-Trust (RoT) for the Pensando Programmable Services Processor. As discussed in the Secure Boot section below, *PenTrust* is the first subsystem to boot at power-on, and it does so from its immutable embedded ROM; making it the first link in the secure boot chain. At the heart of the RoT is a Physically Unclonable Function (PUF). This is a specialized silicon element, tied to specific physical properties of each silicon die, that provides a device-specific seed key. The result is that each chip produced by Pensando will have a unique PUF seed key that cannot be read outside of the device nor tampered with. The seed key is exclusively used to derive a 256-bit AES Storage Root Key (SRK) and an ECDSA-P384 Endorsement Key (EK).

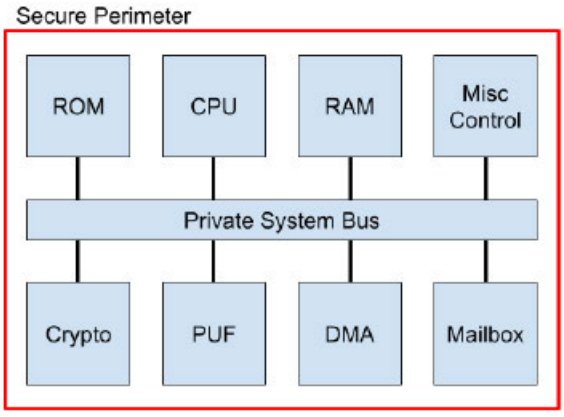


Fig 3. PenTrust Subsystem

Exhibit 7, Security Architecture of Pensando's Distributed Services Platform, Pensando White Paper (2020).

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

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

94. 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.

95. 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.

96. 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. FOURTH CLAIM
(Infringement of the '634 Patent)

97. Plaintiffs incorporate by reference paragraphs 1-73 herein.

98. 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.

99. 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.

100. 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, the SmartSwitch includes a PCIe card that is installed in any server to implement distributed services and the SmartSwitch is a top-of-rack switch that also supports distributed services for all the hosts in the rack. https://www.amd.com/system/files/documents/pensando-smartswitches.pdf.

101. 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, the SmartSwitch collects parameter values for determination of bandwidth of network connections to further endpoint devices via the servers which collect network telemetry. https://www.amd.com/system/files/documents/pensando-smartswitches.pdf. For example, the

Pensando DPUs can determine the transmission bandwidth and other telemetry related to a network flow.

> The Pensando flow-based engine stores each flow in a flow table. Each flow table entry includes counters and timers. For the first time in the history of networking, it is possible to provide precise information on how each flow behaves, including parameters like bytes and packet exchange, delays, jitter, and bandwidth. Previously this information was available only at the aggregate level, for example, at the interface level. This new granularity is what we call "Telemetry," It allows us to pinpoint performance issues at the network and application level.

Exhibit 32, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf.

102. 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 SmartSwitch running a Pensando DPU (an example of a tuning server) determines that a next network connection from the endpoint (e.g., a server) to a second endpoint device (e.g., another server) matches a geographical area of past network connection (e.g., flow state table).

> The stateful forwarding mechanism is also called *cache-based forwarding*. It relies on a flow cache, a binary data structure capable of "exact" matching packets belonging to a particular flow. The word "exact" implies a binary match easier to implement both in hardware or software, unlike a ternary match, such as LPM. The flow cache contains an entry for each flow, i.e., two entries per session. The flow can be defined with an arbitrary number of fields, thus supporting IPv4 and IPv6 addresses, different encapsulations, policy routing, and firewalling.
>
> Cache entry contains information needed to forward the packet (e.g., layer 2 and/or layer 3 address of the next hop, type of encapsulation, etc.).
>
> The separate initial processing that leads to the forwarding decision may create new cache entries. When a packet is received by the switch, if it does not match any entry in the flow cache (a "miss") it is processed according to the initial processing. Otherwise ("hit"), it is forwarded according to the information in the matching flow cache entry.

> This hardware is the *Pensando flow-based engine*, which:
>
> - identifies and tracks all the flows for all the sessions.
> - uses this information to forward packets and to apply services
> - operates at wire rate
> - tracks million of flows simultaneously

Exhibit 32, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf.

103. 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, Pensando DPUs are able to determine the transmission bandwidth and other telemetry related to a network flow.

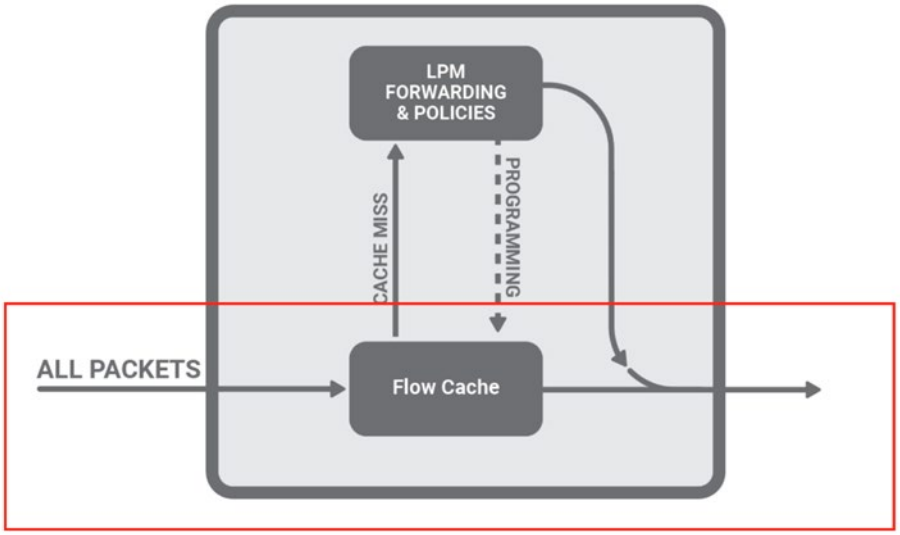


Exhibit 32, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf.

> The Pensando flow-based engine stores each flow in a flow table. Each flow table entry includes counters and timers. For the first time in the history of networking, it is possible to provide precise information on how each flow behaves, including parameters like bytes and packet exchange, delays, jitter, and bandwidth. Previously this information was available only at the aggregate level, for example, at the interface level. This new granularity is what we call "Telemetry," It allows us to pinpoint performance issues at the network and application level.

Exhibit 32, https://www.amd.com/system/files/documents/pensando-smartswitches.pdf; *see* https://www.varindia.com/news/pensando-brings-hyperscale-dpu-technology-to-amds-data-center-capabilities ("SmartSwitch technology, as exemplified by the Aruba CX 10000 with AMD Pensando, allows for 'East-West' traffic within the data center itself to be gathered and analyzed, either via intrinsic capabilities of the AMD Pensando Distributed Services Platform, or by providing the data to 3rd party Extended Detection and Response (XDR) vendor products.").

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

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

106. 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.

107. 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.

108. 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.

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

109. Plaintiffs incorporate by reference paragraphs 1-85 herein.

110. 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.

111. 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.

112. The Accused Products, including AMD Pensando's DPU-enabled systems, make up a routable packet-switched network supported by an abstraction protocol.

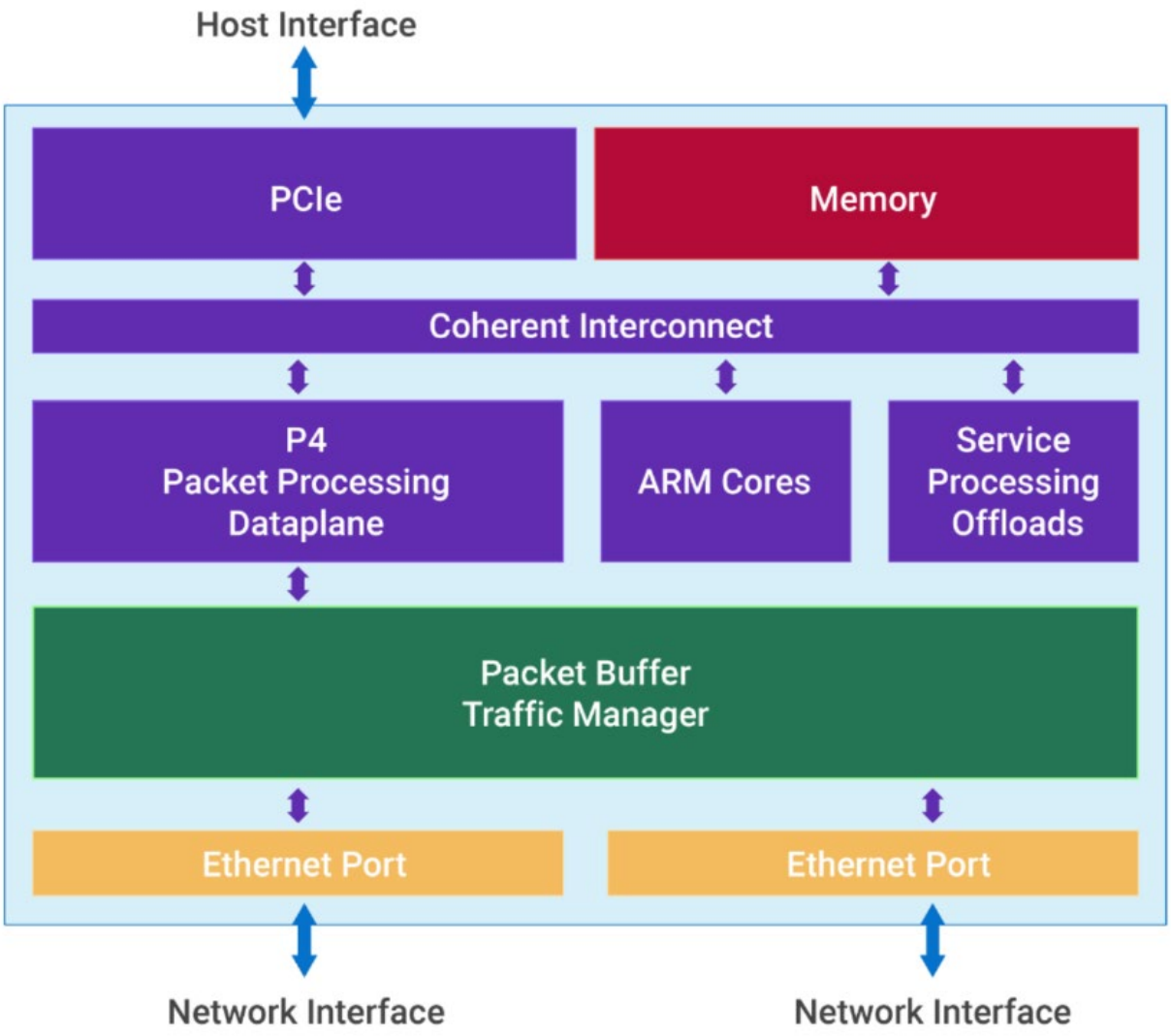
> AMD Pensando DPUs are programmed using the industry-standard P4 language, facilitating the implementation of a wide variety of system solutions. It supports software-defined networking and storage protocols, including NVMe virtualization and transport, and is designed to give developers the agility to develop and deploy new features and modifications throughout the product lifecycle.
>
> The Elba DPU form factor and power profile are designed to support multiple system level implementations ranging from a half-height, half-length PCIe card that can fit into the power and cooling profiles of any standard server to network and security appliances and SmartSwitches. The P4-programmable design enables these applications to dynamically re-configure the data processing inside the DPU.

Key Features

Feature	Description
Network Interfaces	56 Gbps PAM4 SerDes supporting Dual 200 GE, Quad-100/50/25/10GE 1GE management port
PCIe Interface	32 Lanes of PCIe Ge n4, configurable as root complex or end-point mode 2x16 / 4x8 / 8x4 with QOS support in multi-host applications
Data Pipeline	P4 data pipeline comprising 144 match processing units (MPUs) @2 GHz Provides high performance capabilities in packet and message processing, at line rate

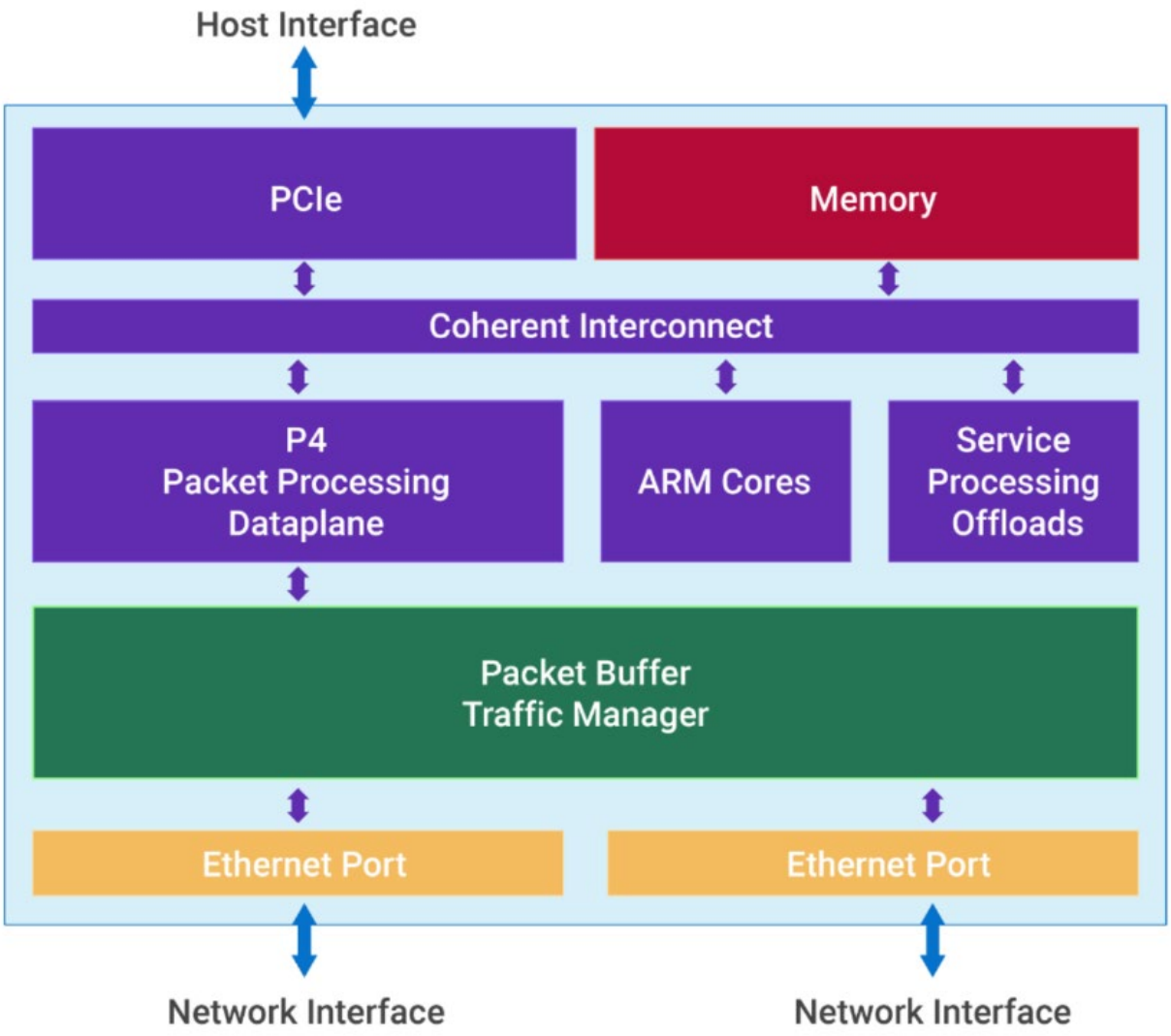
https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/product-briefs/pensando-elba-product-brief.pdf; Exhibit 27 (AMD-CV-0004687), AMD Pensando Giglio Product Brief.

113. The Accused Products, including AMD Pensando's DPU-enabled systems, include at least one host.



https://infohub.delltechnologies.com/en-US/l/dpus-in-the-new-vsphere-8-0-and-16th-generation-dell-poweredge-servers/amd-pensando-dpu.

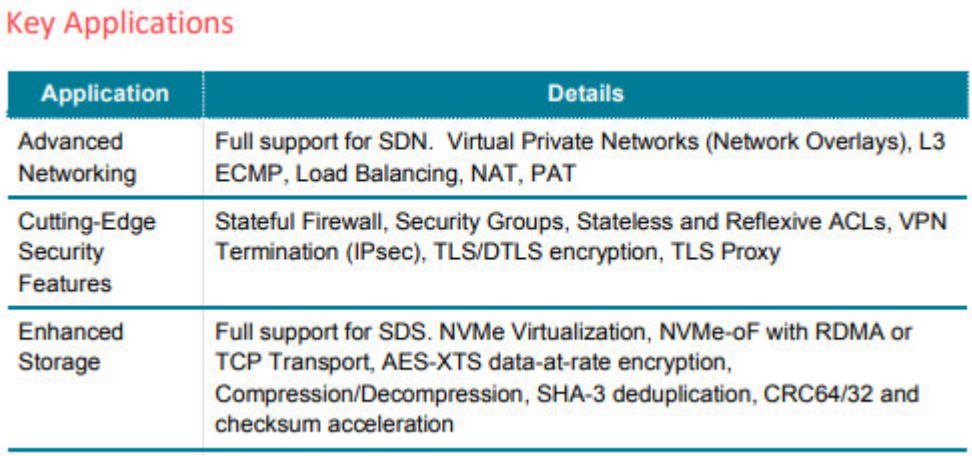
114. The Accused Products, including AMD Pensando's DPU-enabled systems, include a master controller, the master controller comprising a master processing element, a master controller interface to communicate with the host, and at least one master controller control link interface. For example, the DPU itself can serve as a master controller, and includes a master processing element (e.g., a processor).



https://infohub.delltechnologies.com/en-US/l/dpus-in-the-new-vsphere-8-0-and-16th-generation-dell-poweredge-servers/amd-pensando-dpu.

115. The Accused Products, including AMD Pensando'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, the "edge controllers" can be other DPUs, which the master controller can communicate with. In turn, the "edge processing element" can be the processer (e.g., DPU chip) within that edge DPU and the "edge controller control link interface" is the interface

for the network connection that attaches the edge controller to the master controller, with a logical separation for control functions.



Key Applications

Application	Details
Advanced Networking	Full support for SDN. Virtual Private Networks (Network Overlays), L3 ECMP, Load Balancing, NAT, PAT
Cutting-Edge Security Features	Stateful Firewall, Security Groups, Stateless and Reflexive ACLs, VPN Termination (IPsec), TLS/DTLS encryption, TLS Proxy
Enhanced Storage	Full support for SDS. NVMe Virtualization, NVMe-oF with RDMA or TCP Transport, AES-XTS data-at-rate encryption, Compression/Decompression, SHA-3 deduplication, CRC64/32 and checksum acceleration

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/product-briefs/pensando-elba-product-brief.pdf; Exhibit 27 (AMD-CV-0004687), AMD Pensando Giglio Product Brief

116. The Accused Products, including AMD Pensando'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. In particular, the master controller and the edge controller communicate via NVMe over Fabrics, which is an abstraction protocol, and relies on the underlying NVMe Transport to ensure reliable delivery of NVMe commands and data. NVMe over Fabrics also relies on RDMA. Defendants' implementation and use of NVMe over Fabrics by itself and as implementing and using NVMe Transport and RDMA is a claimed abstraction protocol. The underlying NVMe Transport (as well as RDMA) serves as a part of the NVME over Fabrics abstraction protocol, independent of any specific physical interconnect characteristics. *See, e.g.*, Exhibit 23 (CV-AMD0015909); Exhibit 24 (AMD-CV-0000001) at -10–11; Exhibit 25

(AMD-CV-0011209) at -11226; *see also infra*. NVMe over Fabrics is a full-duplex protocol and supports full command queuing for the at least one storage media device.[25]

AMD Pensando DPUs are programmed using the industry-standard P4 language, facilitating the implementation of a wide variety of system solutions. It supports software-defined networking and storage protocols, including NVMe virtualization and transport, and is designed to give developers the agility to develop and deploy new features and modifications throughout the product lifecycle.

The Elba DPU form factor and power profile are designed to support multiple system level implementations ranging from a half-height, half-length PCIe card that can fit into the power and cooling profiles of any standard server to network and security appliances and SmartSwitches. The P4-programmable design enables these applications to dynamically re-configure the data processing inside the DPU.

Key Features

Feature	Description
Network Interfaces	56 Gbps PAM4 SerDes supporting Dual 200 GE, Quad-100/50/25/10GE 1GE management port
PCIe Interface	32 Lanes of PCIe Ge n4, configurable as root complex or end-point mode 2x16 / 4x8 / 8x4 with QOS support in multi-host applications
Data Pipeline	P4 data pipeline comprising 144 match processing units (MPUs) @2 GHz Provides high performance capabilities in packet and message processing, at line rate

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/product-briefs/pensando-elba-product-brief.pdf; Exhibit 27 (AMD-CV-0004687), AMD Pensando Giglio Product Brief.

Key Applications

Application	Details
Advanced Networking	Full support for SDN. Virtual Private Networks (Network Overlays), L3 ECMP, Load Balancing, NAT, PAT
Cutting-Edge Security Features	Stateful Firewall, Security Groups, Stateless and Reflexive ACLs, VPN Termination (IPsec), TLS/DTLS encryption, TLS Proxy
Enhanced Storage	Full support for SDS. NVMe Virtualization, NVMe-oF with RDMA or TCP Transport, AES-XTS data-at-rate encryption, Compression/Decompression, SHA-3 deduplication, CRC64/32 and checksum acceleration

[25] https://nvmexpress.org/wp-content/uploads/NVMe-over-Fabrics-1.1a-2021.07.12-Ratified.pdf; https://en.wikipedia.org/wiki/NVM_Express; https://docs.broadcom.com/doc/nvme-over-fibre-channel-for-dummies-book.

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/product-briefs/pensando-elba-product-brief.pdf; Exhibit 27 (AMD-CV-0004687), AMD Pensando Giglio Product Brief.

117. The Accused Products, including AMD Pensando'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, on information and belief, AMD's DPU edge controllers can contain an element acting as a master controller for communication with other edge controllers. *See, e.g.*, https://www.datanami.com/this-just-in/amd-showcases-continued-enterprise-data-center-momentum-with-epyc-cpus-and-pensando-dpus ("Additionally, as data center applications continue to grow in scale and complexity, AMD Pensando DPUs are the ideal solution to help offload infrastructure services from the CPU to free up valuable work cycles. VMware vSphere 8-enabled systems, powered by AMD EPYC CPUs and Pensando DPUs, deliver the performance, efficiency and flexibility IT leaders need to run a broad set of business-critical workloads."); https://www.techinsights.com/blog/amds-new-pensando-dpu-cuts-power; https://www.amd.com/content/dam/amd/en/documents/pensando-business-docs/solution-brief/csp-case-study.pdf.

118. A particular example is shown in the figures below. Each of the host servers, storage nodes, and Just a Bunch of Flash ("JBOF") include accused and infringing DSC cards with DPUs or DPUs which support NVMEof. In this example, the host server operates as the top level master controller, and the SAN cluster sitting in the middle of the figures acts as a composite controller and includes both an edge controller (edge to the DPU connected to the host) and a master controller (master to the DPU connected to the JBOF). As a result, the SAN cluster node simultaneously behaves as an edge relative to the host master controller and as a master relative to

the JBOF side edge controller, creating a cascaded master–edge–master–edge control hierarchy over the abstraction protocol.

Exhibit 25 (AMD-CV-00112209) at -11226.

Exhibit 25 (AMD-CV-00112209) at -11227.

119. In this example, the SAN Cluster DPUs, through their role as master controllers, are in communication with the edge controllers of the JBOF DPU (the other edge controllers). The SAN Cluster DPUs' edge processing elements (for example, their P4 ingress, P4 egress, and Arm Cores by themselves or in combination) authors abstract protocol messages intended for the other edge controller, the JBOF DPUs.

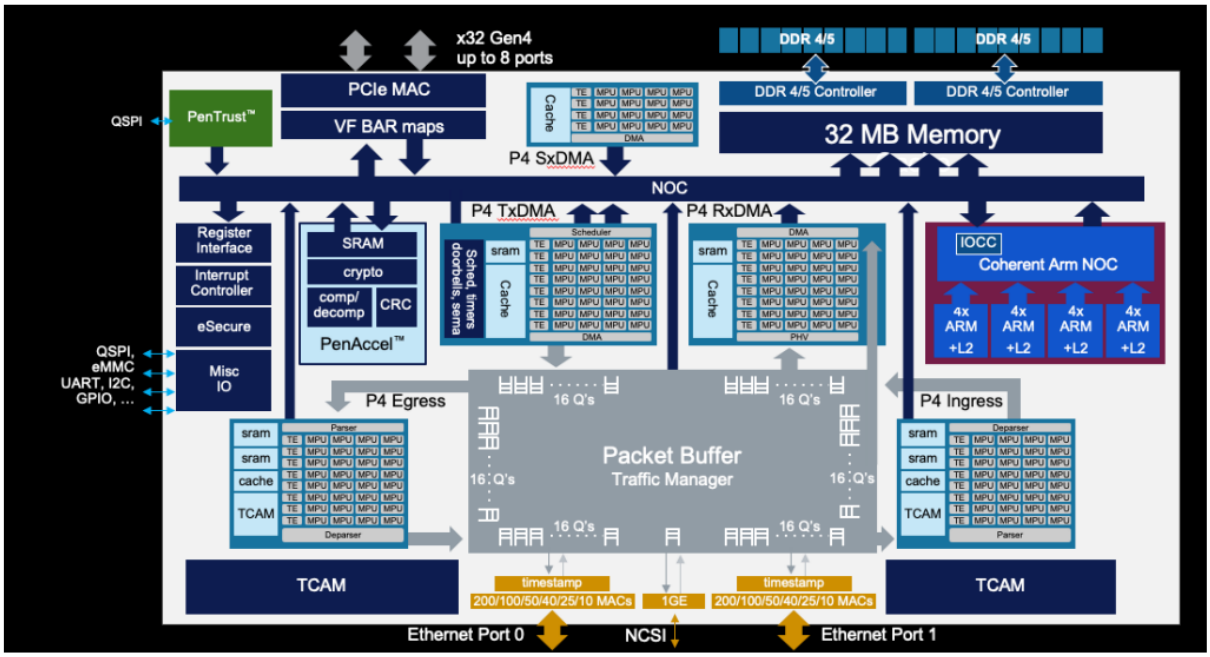


Figure 13. AMD Pensando Architecture

https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/white-papers/pensando-comparison-of-dpu-hardware-strategies.pdf, Exhibit 26 (CV-AMD0015795) at -15807.

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

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

122. 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.

123. 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.

124. 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.

XII. 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.

XIII. 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;

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: December 30, 2025

Respectfully submitted,

/s/ Michael Matulewicz-Crowley

Michael Matulewicz-Crowley
mmc@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
400 Madison Avenue, Suite 14D
New York, NY 10017
Telephone: (212) 381-1965

Christine E. Lehman
clehman@reichmanjorgensen.com
Adam Adler
aadler@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
1909 K Street NW, Suite 800
Washington DC, 20006
Telephone: (202) 894-7310

Courtland L. Reichman (CA Bar No. 268873)
creichman@reichmanjorgensen.com
Jennifer Estremera (CA Bar No. 251076)
jestremera@reichmanjorgensen.com
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

Scott Cole
scole@reichmanjorgensen.com
Ariane S. Mann
amann@reichmanjorgensen.com
REICHMAN JORGENSEN
LEHMAN & FELDBERG LLP
515 Congress Avenue, Suite 1900
Austin, TX 78701
Telephone: (650) 623-1401

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