IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

REX COMPUTING, INC.,	)	
	)	
Plaintiff,	)	
	)	C.A. No. 21-525 (MN)
v.	)	
	)	JURY TRIAL DEMANDED
CEREBRAS SYSTEMS INC.,	)	
	)	
Defendant.	)	

FIRST AMENDED COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff Rex Computing, Inc. ("Rex"), as and for its complaint against Defendant Cerebras Systems Inc. ("Cerebras") alleges as follows:

NATURE OF THE ACTION

1. In this action, Rex alleges that Cerebras infringes U.S. Patent Nos. 10,355,975 ("the '975 patent"), 10,700,968 ("the '968 patent"), and 10,127,043 ("the '043 patent") (collectively "the Asserted Patents").

2. Rex brings this lawsuit to end Cerebras' unauthorized, willful, and infringing manufacture, use, sale, offer for sale, and/or importation into the United States of Cerebras products that incorporate Rex's patented inventions. Rex seeks to recover damages adequate to compensate Rex for Cerebras' unlawful actions.

THE PARTIES

3. Rex Computing, Inc. is a Delaware corporation with a principal place of business at 10300 Old Victory Highway, Lovelock, NV 89419.

4. On information and belief, Cerebras Systems Inc. is a Delaware corporation with a principal place of business at 175 S. San Antonio Road, Los Altos, CA 94022. Cerebras' registered agent, Incorporating Services, Ltd., is located at 3500 S. DuPont Hwy, Dover, DE 19901.

5. On information and belief, Cerebras, founded in 2016, is a computer systems company that makes, uses, sells, and offers to sell computer systems including the Cerebras CS-1 deep learning system ("Cerebras CS-1") and the Cerebras CS-2 deep learning system ("Cerebras CS-2"), which are powered by versions of Cerebras' Wafer Scale Engine (WSE and WSE-2, respectively), chips that Cerebras claims are the largest ever built.[1] The Cerebras CS-1, Cerebras CS-2, and all other Cerebras computer systems having similar functionality are collectively referred to herein as the "Accused Products."

JURISDICTION AND VENUE

6. This action for patent infringement arises under the patent laws of the United States, 35 U.S.C. § 271, *et seq*.

7. This Court has subject matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338(a).

8. This Court has personal jurisdiction over Cerebras. On information and belief, Cerebras is a Delaware corporation, and conducts business in this judicial district.

9. Venue is proper in this Court pursuant to 28 U.S.C. § 1400(b). On information and belief, Cerebras is subject to personal jurisdiction in this judicial district, conducts business in this judicial district, and resides in this judicial district.

[1] *See e.g.,* Homepage | Cerebras https://www.cerebras.net/ (last visited May 4, 2021); Cerebras Systems: Achieving Industry Best AI Performance Through A Systems Approach, White Paper 02 ("CS-1 Overview") at 2, *available at* https://secureservercdn.net/198.12.145.239/a7b.fcb.myftpupload.com/wp-content/uploads/2020/03/Cerebras-Systems-Overview.pdf (last visited May 4, 2021); Cerebras Systems: Achieving Industry Best AI Performance Through A Systems Approach, White Paper 03 ("CS-2 Overview") at 2, *available at* https://cerebras.net/wp-content/uploads/2021/04/Cerebras-CS-2-Whitepaper.pdf (last visited May 4, 2021).

BACKGROUND AND FACTS

10. Rex is a fabless semiconductor company that has developed solutions to achieve significantly reduced power consumption and total chip area, as compared with other chips on the market, by removing unnecessary complexity from hardware.

11. Starting in 2013, Rex worked to develop its Neo Architecture to eliminate the feature creep and bloat of processors that had been produced over the preceding thirty years. In doing so, Rex was able to deliver a 10 to 25x increase in energy efficiency for the same performance level compared to then-existing Graphics Processing Units ("GPU") and Central Processing Units ("CPU") systems.

12. In conjunction with its Neo Architecture, Rex developed its Neo Chip to provide users with power, speed, and efficiency performance gains unseen in the industry.

13. Rex applied for and obtained several patents related to the development of its Neo Architecture and Neo Chip, including the Asserted Patents in this litigation.

14. The '975 patent, titled "Latency Guaranteed Network on Chip," was issued on July 16, 2019. A true and correct copy of the '975 patent is filed herewith as **Exhibit 1**.

15. The '975 patent relates to a system and method that includes a network-on-chip microprocessor having a set of processor cores that are communicatively coupled via a set of routers, thereby creating a "network" of tiles, each tile including a core and router pair. The routers send data packets to other routers in the network based on physical destination addresses of the data packets.

16. The '968 patent, titled "Optimized Function Assignment in a Multi-Core Processor," was issued on June 30, 2020. A true and correct copy of the '968 patent is filed herewith as **Exhibit 2**.

17. The '968 patent relates to a system and method that includes receiving an application having a set of functions to be executed by a multi-core multiprocessor chip, the chip including a network of tiles, each tile including a core and router pair. An optimal configuration for execution of the set of functions is selected from different configurations, where the different configurations include execution of the functions by different groups of tiles.

18. The '043 patent, titled "Implementing Conflict-Free Instructions for Concurrent Operation on a Processor," was issued on November 13, 2018. A true and correct copy of the '043 patent is filed herewith as **Exhibit 3**.

19. The '043 patent relates to a system and method for implementing very long instruction words (VLIW) having slot instructions that correspond to a set of functional units. Slot instructions include opcodes and value fields having bits that may be allocated to other slot instructions.

20. Cerebras makes, uses, sells, and offers to sell the Accused Products.[2]

21. The Accused Products include the Cerebras Wafer Scale Engine, Cerebras Swarm communication fabric, and Cerebras Graph Compiler.[3]

22. The Accused Products are powered by the Cerebras Wafer Scale Engine.[4]

[2] *See e.g.*, Homepage | Cerebras, https://www.cerebras.net/ (last visited May 4, 2021); CS-1 Overview; CS-2 Overview.
[3] *E.g.*, Product | Cerebras, https://cerebras.net/product/ (last visited May 4, 2021); CS-1 Overview at 4; CS-2 Overview at 4.
[4] *E.g.*, Product | Cerebras, https://cerebras.net/product/ (last visited May 4, 2021); CS-1 Overview at 7−8; CS-2 Overview at 7−8.

23. As described by Cerebras, the Cerebras Wafer Scale Engine "provides more compute cores, tightly coupled memory for efficient data access, and an extensive high bandwidth communication fabric for groups of cores to work together."[5]

24. As described by Cerebras, the Cerebras Swarm communication fabric "is a massive on-chip communication fabric that delivers breakthrough bandwidth and low latency at a fraction of the power draw of traditional techniques used to cluster graphics processing units. It is fully configurable; software configures all the cores on the WSE to support the precise communication required for training the user-specified model. For each neural network, Swarm provides a unique and optimized communication path."[6]

25. As described by Cerebras, the Cerebras Graph Compiler "generates a placement and routing, unique for each neural network, to minimize communication latency between the layers."[7]

---

[5] Product - Cerebras, https://cerebras.net/product/ (visited April 13, 2021); *see also* CS-1 Overview at 3 ("the WSE contains 78 times more compute cores, 3,000 times more high-speed on-chip memory, 10,000 times more memory bandwidth and 33,000 times more fabric bandwidth than its graphics processing competitor"); CS-2 Overview at 3 ("the WSE-2 contains 123 times more compute cores, 1,000 times more high-speed on-chip memory, 12,862 times more memory bandwidth and 45,833 times more fabric bandwidth than its graphics processing competitor").

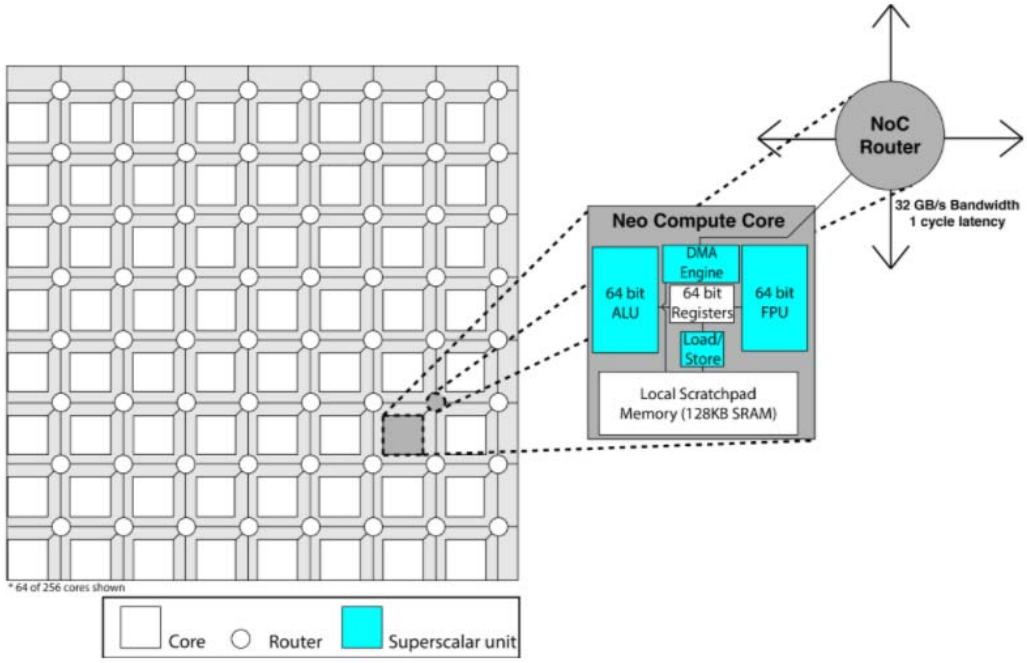
[6] Product - Cerebras, https://cerebras.net/product/ (visited April 13, 2021) *see also* CS-1 Overview at 4 ("The Cerebras Swarm communication fabric creates a massive on-chip network that delivers breakthrough bandwidth and low latency, at a fraction of the power draw of traditional communication techniques that are used to aggregate servers of graphics processing units into large clusters. . . . Swarm is also fully configurable. Cerebras' software can configure all the cores and routers on the WSE to support a unique, optimized communication pattern for each particular neural network."); CS-2 Overview at 4 ("The Cerebras Swarm communication fabric creates a massive on-chip network that delivers breakthrough bandwidth and low latency, at a fraction of the power draw of traditional communication techniques that are used to aggregate servers of graphics processing units into large clusters. . . . Swarm is also fully configurable. Cerebras' software configures all the cores and routers on the WSE to support a unique, optimized communication pattern for each particular neural network.").

[7] Product - Cerebras, https://cerebras.net/product/ (visited April 13, 2021); *see also* CS-1 Overview at 8 ("During this compilation process, kernel placement is formulated as a multi-constraint problem on 1) memory capacity and bandwidth, 2) computation requirements, and 3)

26. The Accused Products utilize the same systems and methods taught by the Asserted Patents.

27. Cerebras has demonstrated a pattern of illegally using Rex's intellectual property.

28. At least as early as March 2015, Rex's website portrayed the arrangement of its Neo Chip with the following diagram.



The Rex Neo Chip[8]

communication costs. The placement engine then takes into account both algorithmic efficiency and compute core utilization to generate a result that maximizes locality, minimizes routing distances, and avoids hotspots."); CS-2 Overview at 8 ("During this compilation process, kernel placement is formulated as a multi-constraint problem on 1) memory capacity and bandwidth, 2) computation requirements, and 3) communication costs. The placement engine then takes into account both algorithmic efficiency and compute core utilization to generate a result that maximizes locality, minimizes routing distances, and avoids hotspots.").

[8] REX Computing - Energy Efficient HPC, https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).

29. Rex created the above diagram to illustrate, *inter alia*, how routers and processor cores are arranged into tiles on Rex's Neo Chip, and each tile's ability to communicate with other tiles on the Neo Chip.

30. On information and belief, Cerebras did not exist as a company at the time Rex first provided the above diagram of the Neo Chip arrangement on Rex's website in early 2015.

31. On information and belief, Cerebras did not establish its own website until the middle of 2016.

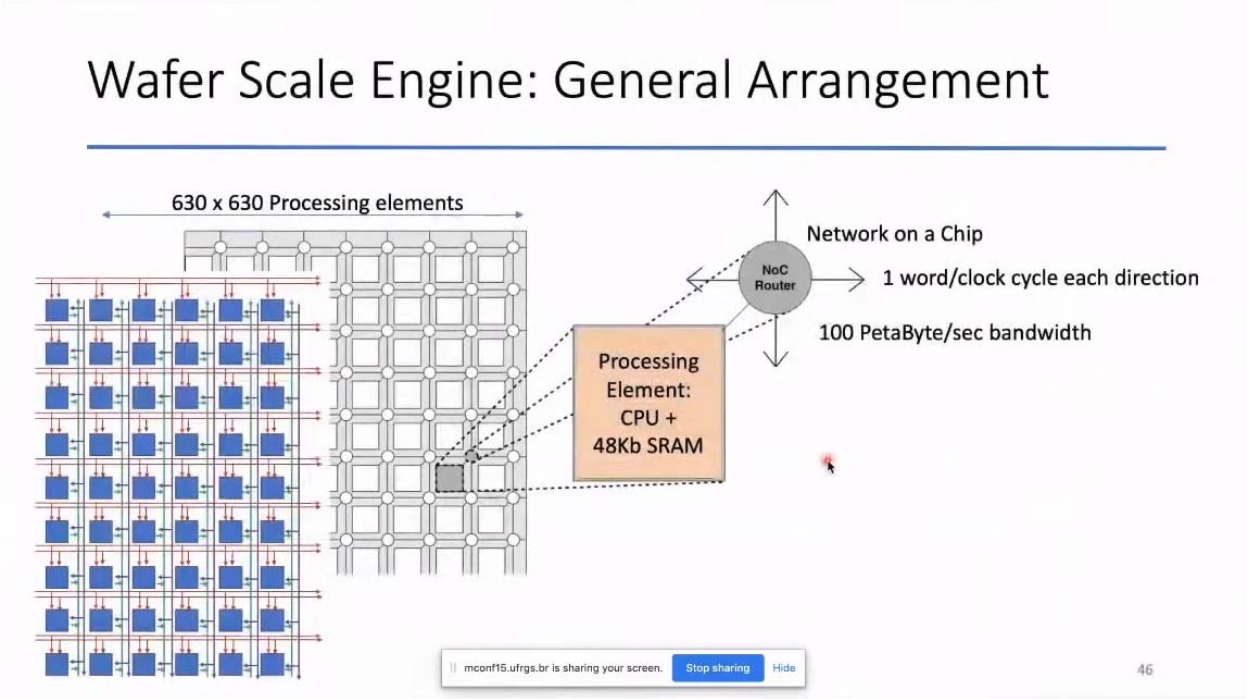
32. The Cerebras Wafer Scale engine has the same arrangement of routers and processor cores as Rex's Neo Chip.

33. Rex has never authorized Cerebras to use Rex's illustrations of the Neo Chip.

34. Nonetheless, in depicting the Cerebras Wafer Scale Engine, Cerebras has blatantly copied Rex's diagram of the Neo Chip.

35. For example, as shown below, in a presentation given at a webinar hosted by the Rio Grande do Sul Chapter of the IEEE CASS (Circuits and Systems Society), which took place on June 26, 2020,[9] Cerebras intentionally, and without authorization, used Rex's diagram when depicting the arrangement of the Cerebras Wafer Scale Engine.

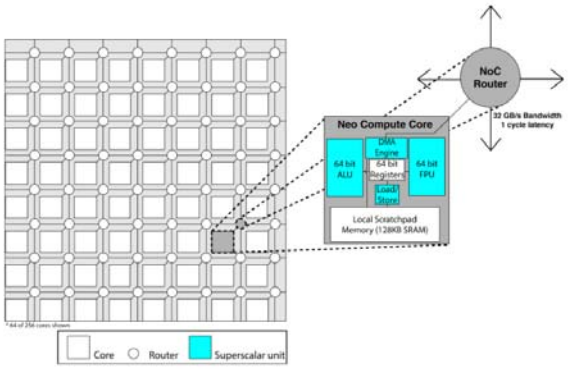
[9] *See* IEEE CASS Webinars hosted by the Rio Grande do Sul Chapter, https://ieee-cas.org/ieee-cass-webinars-hosted-rio-grande-do-sul-chapter (last visited May 4, 2021).



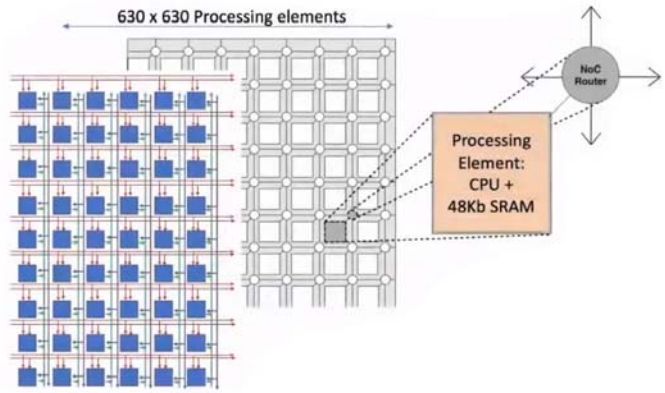
Depiction of Cerebras' Wafer Scale Engine Arrangement in Cerebras' CASS Presentation[10]

36. The side-by-side comparison below demonstrates Cerebras' unauthorized use of Rex's diagram (left) in Cerebras' CASS presentation (right).

[10] *See* CASS Talks 2020 - Patrick Groeneveld, Cerebras Systems, USA - June 26, 2020 ("Groeneveld CASS Pres.") at 43:28, https://www.youtube.com/watch?v=P4YHmwzlkic (last visited April 12, 2021).



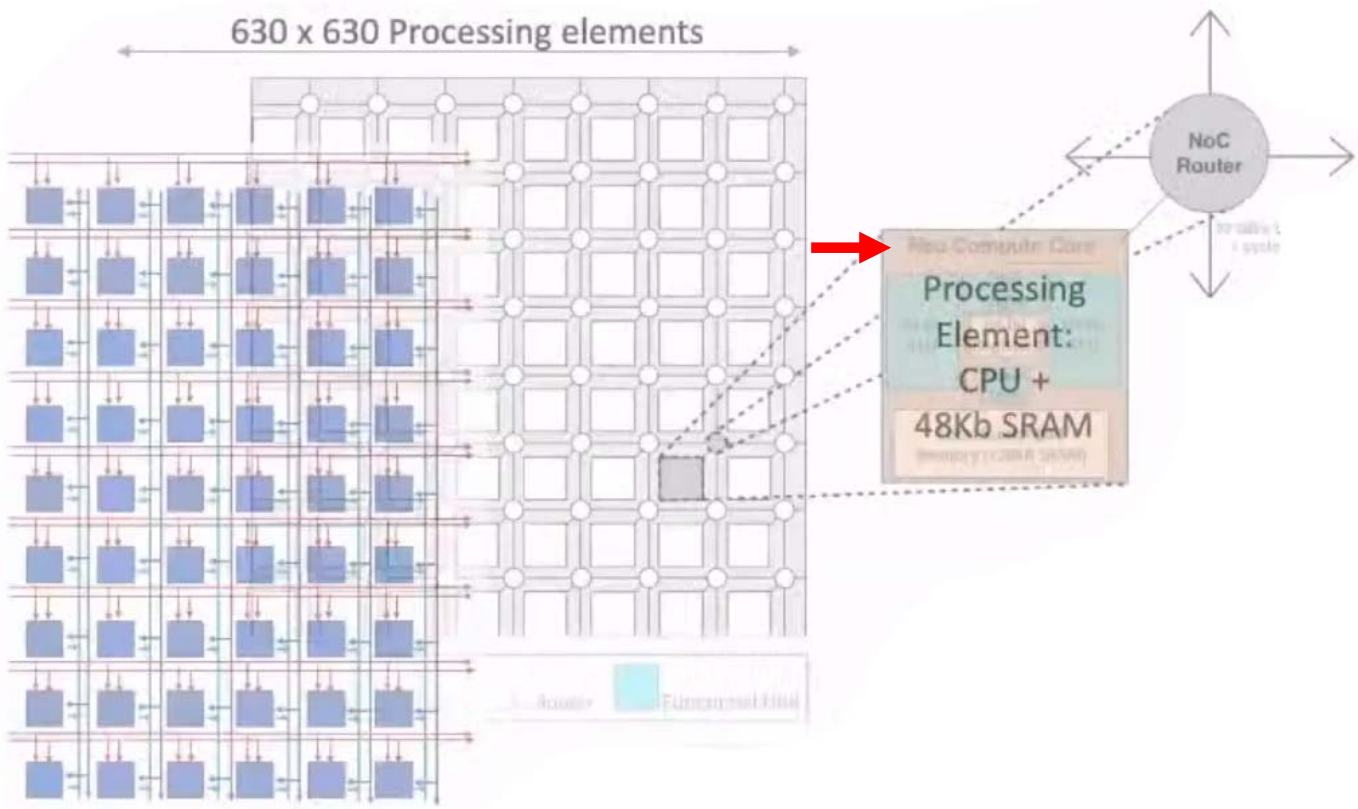
The Rex Neo Chip[11]



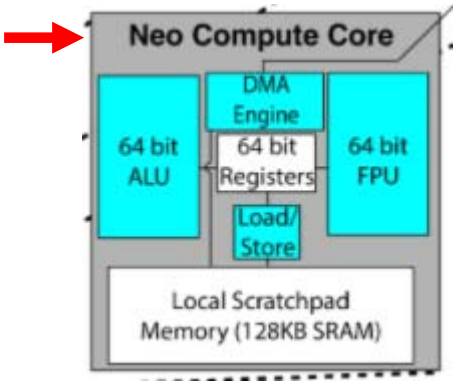
Depiction of Cerebras' Wafer Scale Engine Arrangement in Cerebras' CASS Presentation[12]

37. The following images include a screen-capture image obtained from the recorded video of Cerebras' CASS presentation. The screen-capture image, which is visible for a fraction of a second during a transition from a previous slide in Cerebras' presentation to the slide depicted above in paragraph 35, clearly shows that the words "Neo Compute Core" as well as additional details from Rex's diagram also appear in Cerebras' presentation. The screen-capture image provides further evidence that Cerebras copied Rex's intellectual property, and made a deliberate attempt to conceal its copying by pasting over portions of Rex's diagram.

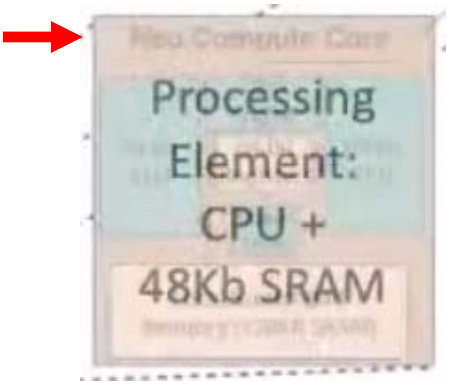
[11] REX Computing - Energy Efficient HPC, https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).
[12] Groeneveld CASS Pres. at 43:28.



Screen-Capture Image from Cerebras' 2020 CASS Presentation During Slide Transition[13]



From Rex's Diagram of the Neo Compute Core[14]

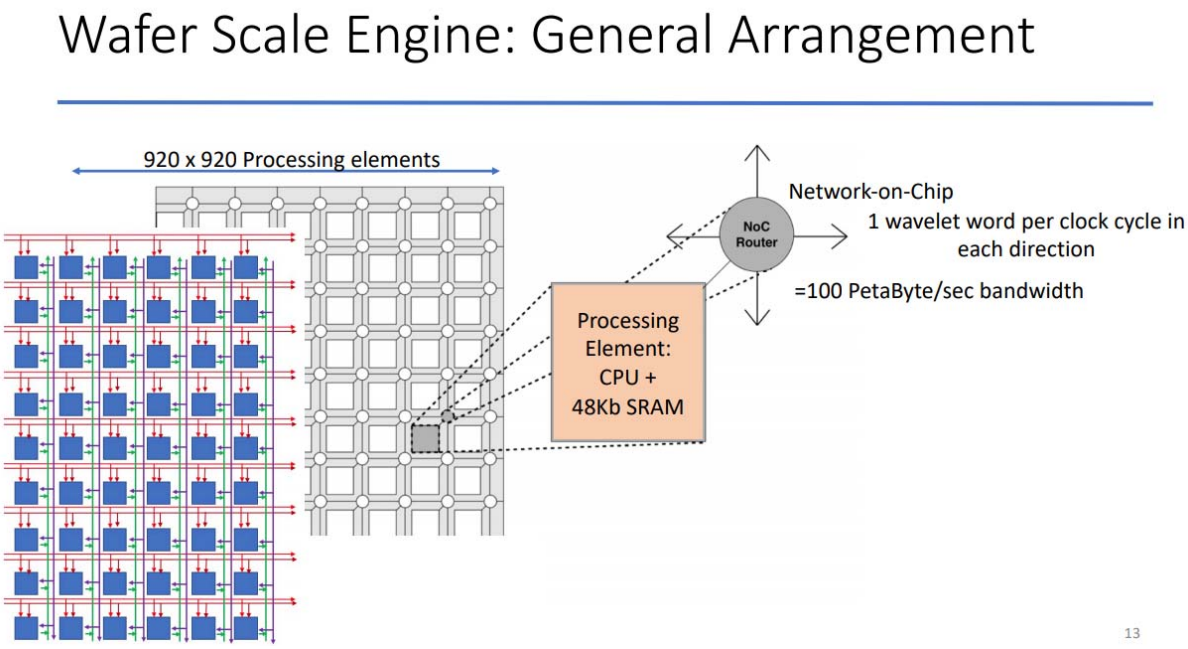


Screen-Capture Image from Cerebras' 2020 CASS Presentation During Slide Transition[15]

38. In another example, as shown below, in a presentation given at the Electronic Design Process Symposium (EDPS), which took place from September 30 to October 1, 2020,[16]

[13] Groeneveld CASS Pres. at 43:28.
[14] REX Computing - Energy Efficient HPC, https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).

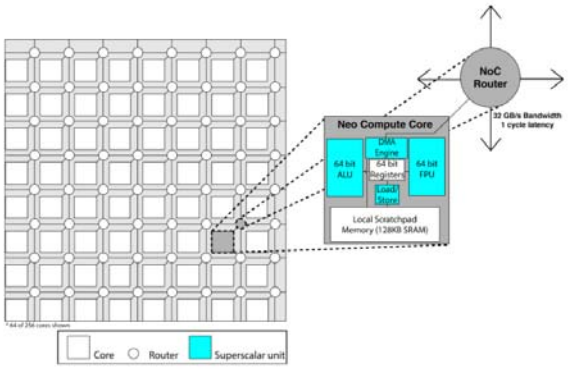
Cerebras intentionally, and without authorization, used Rex's diagram when depicting the arrangement of the Cerebras Wafer Scale Engine.



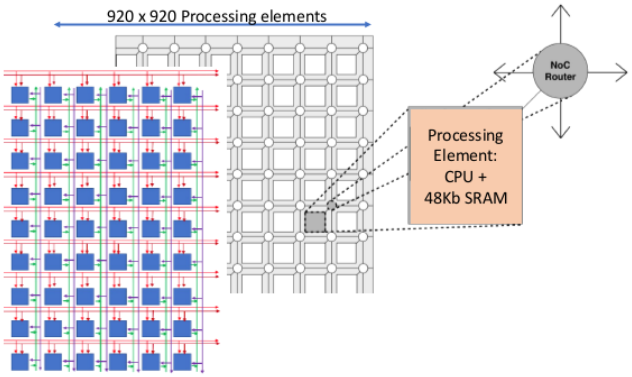
Depiction of Cerebras' Wafer Scale Engine Arrangement in Cerebras' EDPS Presentation[17]

39. The side-by-side comparison below demonstrates Cerebras' unauthorized use of Rex's diagram (left) in Cerebras' EDPS presentation (right).

[15] Groeneveld CASS Pres. at 43:28.
[16] *See* EDPS Archive for 2020, https://ieee-edps.com/archives/2020/program1.html (last visited May 4, 2021).
[17] *See* Patrick Groeneveld, Presentation at the Electronic Design Process Symposium (EDPS) 2020 ("Groeneveld EDPS Pres.") at 14:53, *available at* https://ieee-edps.com/archives/2020/c/1007groeneveld.mp4 (last visited April 12, 2021); *see also* Patrick Groeneveld, Implementing Machine Learning on Massively Parallel Hardware 13 (2020) ("Groeneveld"), *available at* https://ieee-edps.com/archives/2020/c/1000groeneveld.pdf (last visited April 12, 2021).



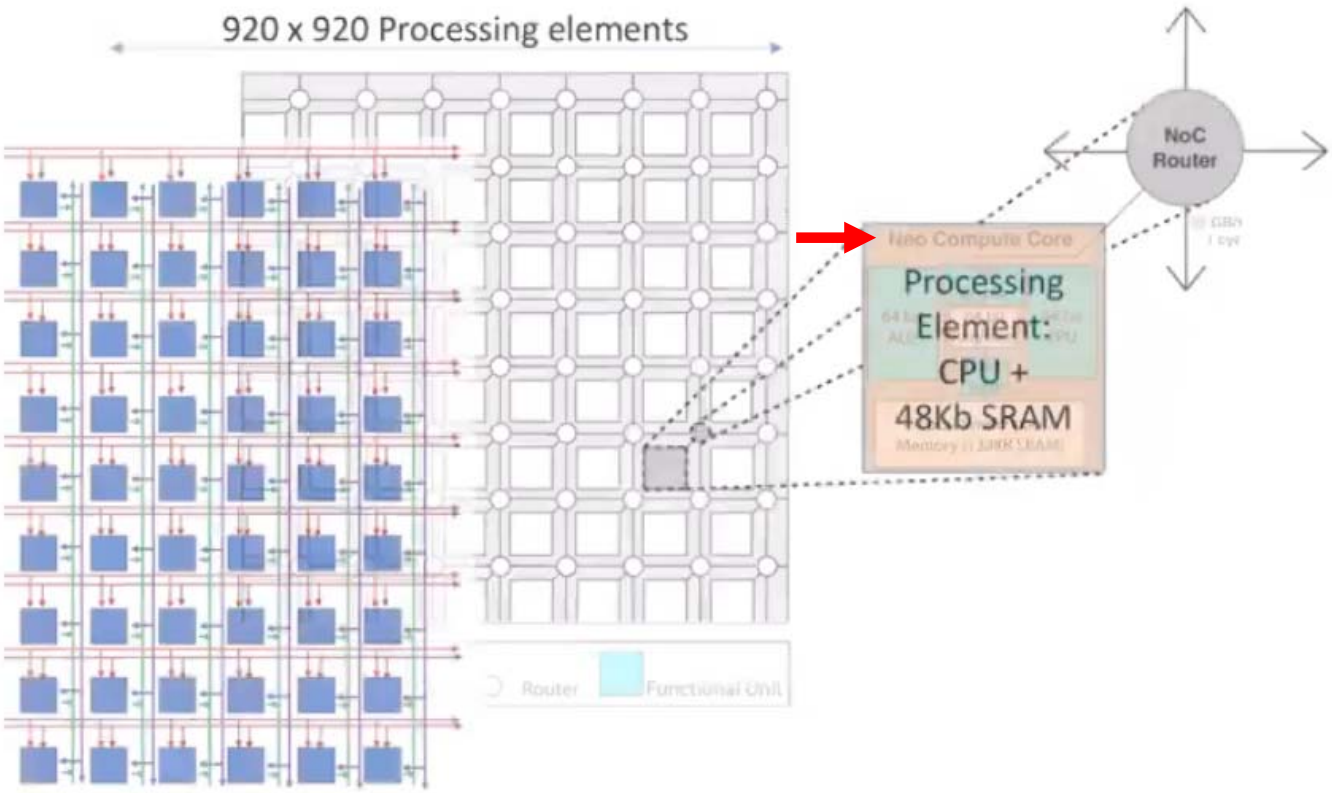
The Rex Neo Chip[18]



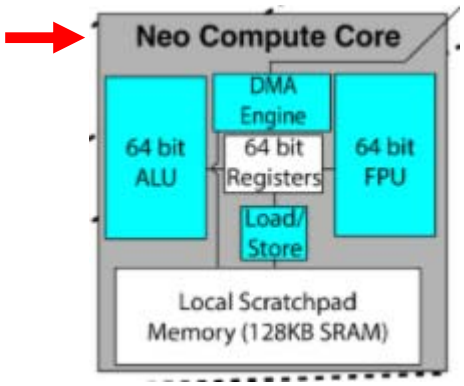
Depiction of Cerebras' Wafer Scale Engine Arrangement in Cerebras' EDPS Presentation[19]

40. The following images include a screen-capture image obtained from the recorded video of Cerebras' EDPS presentation. The screen-capture image, which is visible for a fraction of a second during a transition from a previous slide in Cerebras' presentation to the slide depicted above in paragraph 38, clearly shows that the words "Neo Compute Core" as well as additional details from Rex's diagram also appear in Cerebras' presentation. The screen-capture image provides further evidence that Cerebras copied Rex's intellectual property, and made a deliberate attempt to conceal its copying by pasting over portions of Rex's diagram.

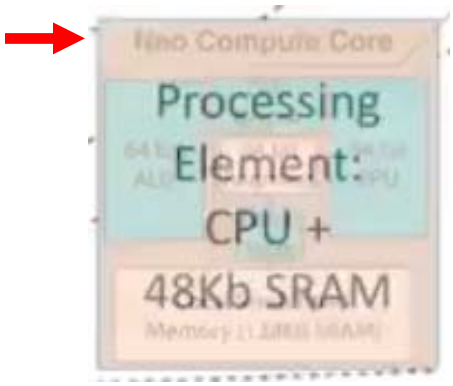
[18] REX Computing - Energy Efficient HPC, https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).
[19] Groeneveld EDPS Pres. at 14:53; Groeneveld at 13.



Screen-Capture Image from Cerebras' 2020 EDPS Presentation During Slide Transition[20]



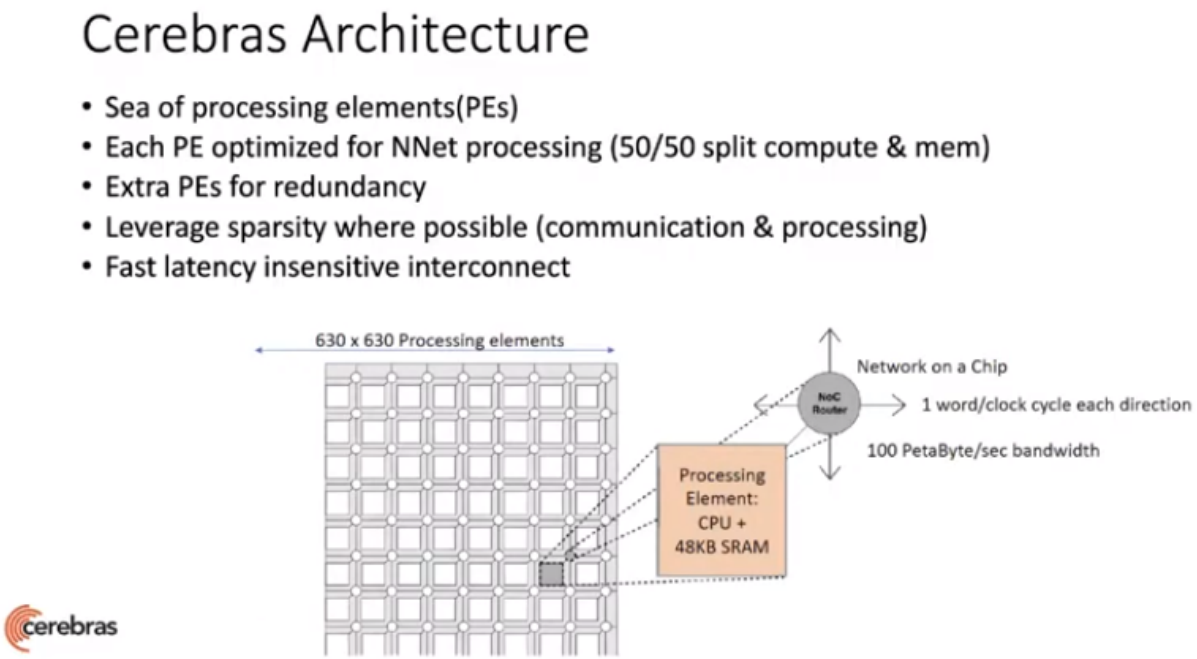
From Rex's Diagram of the Neo Compute Core[21]



Screen-Capture Image from Cerebras' 2020 EDPS Presentation During Slide Transition[22]

41. In yet another example, as shown below, in a Virtual Kanata 2020 FPGA Seminar on October 8, 2020, Cerebras intentionally, and without authorization, used Rex's diagram when depicting the Cerebras Wafer Scale Engine architecture.

[20] Groeneveld EDPS Pres. at 14:53.
[21] REX Computing - Energy Efficient HPC,



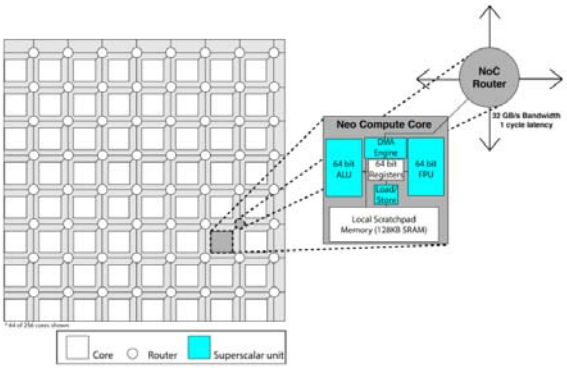
Depiction of Cerebras' Wafer Scale Engine in Cerebras' Virtual Kanata 2020 FPGA Seminar Presentation[23]

42. The side-by-side comparison below demonstrates Cerebras' unauthorized use of Rex's diagram (left) in Cerebras' Virtual Kanata 2020 FPGA Seminar presentation (right).

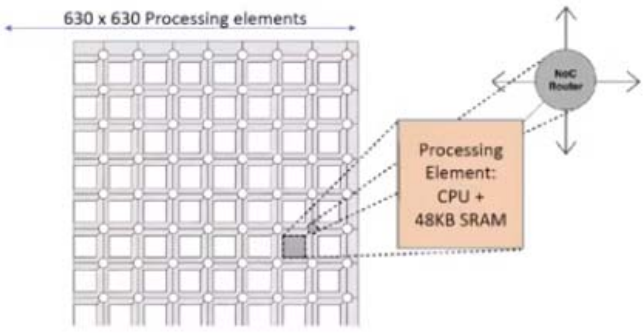
https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).

[22] Groeneveld EDPS Pres. at 14:53.

[23] *See* Cerebras -- Valavan Manohararajah -- Virtual Kanata 2020 FPGA Seminar ("Manohararajah Pres.") at 9:55, 50:46, https://www.youtube.com/watch?v=31CkHmYX9Ho (last visited April 12, 2021).



The Rex Neo Chip[24]



Depiction of Cerebras' Wafer Scale Engine in Cerebras' Virtual Kanata 2020 FPGA Seminar Presentation[25]

43. Cerebras' blatant use and modification of Rex's diagram in multiple Cerebras presentations, and the descriptions of the Accused Products' Wafer Scale Engine in those presentations and other publicly available presentations and documents, clearly demonstrates Cerebras' intentional disregard for Rex's intellectual property, including, but not limited to, Cerebras' continued infringement of the Asserted Patents.

44. On information and belief, Cerebras has had knowledge of Rex since at least February 2017, when Rex gave a presentation at Stanford University regarding the Rex Neo Architecture, including the technology disclosed in the Asserted Patents, which was attended or viewed by Cerebras personnel.[26]

45. On information and belief, Cerebras has had knowledge of the Rex patent applications that matured into the Asserted Patents since at least early 2017, when Rex discussed the Neo Architecture technology and the intellectual property such technology embodies with a

[24] REX Computing - Energy Efficient HPC, https://web.archive.org/web/20150311052700/http:/www.rexcomputing.com/ (last visited May 4, 2021).
[25] Manohararajah Pres. at 9:55, 50:46.
[26] *See* Stanford Seminar: The REX Neo Architecture: An energy efficient new processor architecture, https://www.youtube.com/watch?v=ki6jVXZM2XU (last visited May 4, 2021).

paid Cerebras advisor. In March 2017, Rex submitted a paper on the Neo Architecture to be considered for presentation at a technical conference for which Cerebras' advisor served as program co-chair, and for which Cerebras' co-founder and CTO served as program committee member. Rex continued its discussions with Cerebras' advisor between March and July 2017. Rex further discussed the Neo Architecture technology and the intellectual property such technology embodies with several individuals who are currently Cerebras investors.[27] On information and belief, through at least Cerebras' paid advisor, Cerebras' co-founder and CTO, and Cerebras' investors, Cerebras became aware of the Rex patent applications and, by monitoring the statuses of those applications, became aware, or should have been aware, of the Asserted Patents upon their issuance.

46. Cerebras has had knowledge of Rex and its Neo technology since at least July 26, 2020, when it used Rex's diagram, without authorization, in describing the design and operation of the Accused Products' Wafer Scale Engine.

47. On information and belief, Cerebras has been on notice of the Asserted Patents and its infringement thereof since at least at or around the date each patent issued.

48. Prior to the filing this lawsuit, Rex sent a notice letter to Cerebras on April 12, 2021, attaching the Asserted Patents and alleging Cerebras' infringement thereof. Accordingly, Cerebras has been on notice of the Asserted Patents and its infringement thereof at least since it received the notice letter from Rex on April 12, 2021.

COUNT I

(INFRINGEMENT OF U.S. PATENT NO. 10,355,975)

49. Rex realleges paragraphs 1-48 above as if fully set forth herein.

[27] *See* About Us | Cerebras, https://cerebras.net/about/ (last visited May 4, 2021).

50. The ’975 patent, titled “Latency Guaranteed Network on Chip,” is a valid, enforceable patent that was duly issued by the United States Patent and Trademark Office on July 16, 2019 in full compliance with Title 35 of the United States Code.

51. Rex is the assignee of the ’975 patent with ownership of all substantial rights in the ’975 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

52. The ’975 patent relates to a system and method that includes a network-on-chip microprocessor having a set of processor cores that are communicatively coupled via a set of routers, thereby creating a “network” of tiles, each tile including a core and router pair. The routers send data packets to other routers in the network based on physical destination addresses of the data packets.

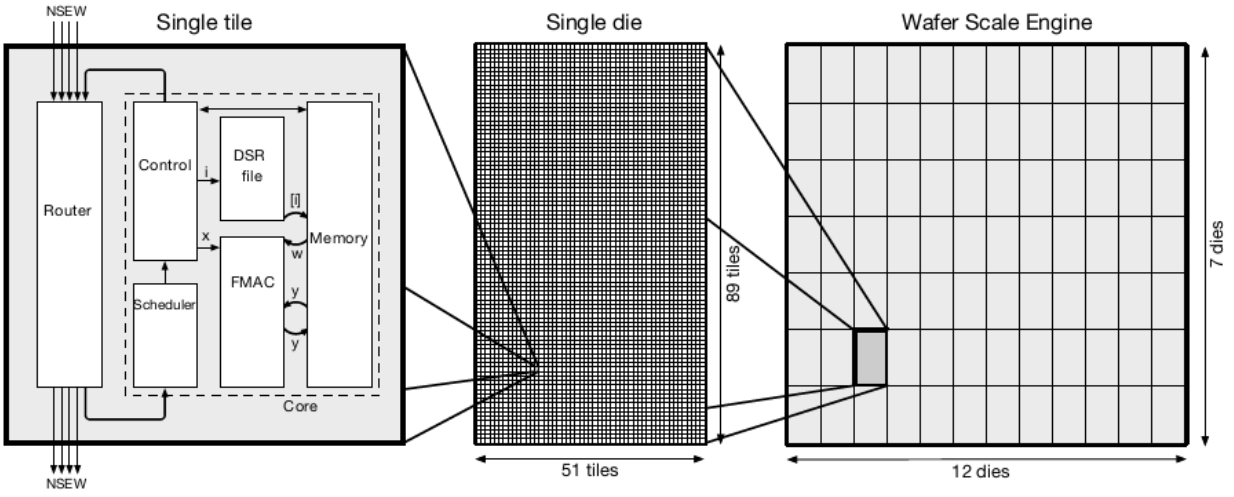
53. Cerebras has directly infringed, and continues to directly infringe, at least claim 1 of the ’975 patent in violation of 35 U.S.C. § 271(a) by, for example and without limitation, making, using, offering to sell, selling, and/or importing in and into the United States certain computer systems, including the Accused Products.

54. Claim 1 of the ’975 patent recites a system comprising: a set of processor cores; a set of routers each including a set of input ports and a set of output ports, wherein: each processor core of the set of processor cores corresponds to a different router of the set of routers, wherein each processor core and corresponding router form a tile, each processor core is communicatively coupled with a corresponding router via the router's set of input ports and set of output ports, each router is communicatively coupled with one or more adjacent routers via the set of input ports and the set of output ports, and wherein each processor core is communicatively coupled with the other processor cores via the set of routers, each router is operable to receive one or more data packets

from the one or more adjacent routers or the processor core corresponding to the router, based on a physical destination address of a data packet, each router is operable to send one or more data packets to the one or more adjacent routers or the processor core corresponding to the router, wherein each router is operable to retain a data packet in the event of a traffic condition, and each router implements a static priority routing policy; and an optimization module configured to: determine optimal function assignment configurations for groups of tiles, and assign two or more functions, which communicate at least unilaterally more frequently with one another than with other functions, to groups of adjacent tiles based on an optimal function assignment configuration determination, wherein the two or more functions are assigned to groups of tiles communicatively coupled in square configurations when the function executes optimally when executed by the groups of tiles communicatively coupled in the square configurations and are assigned to groups of tiles communicatively coupled in linear configurations when the function executes optimally when executed by the groups of tiles communicatively coupled in the linear configurations.

55. The Cerebras CS-1, for example, includes a set of processor cores and a set of routers, each router including a set of input ports and a set of output ports.

56. The leftmost portion of the figure below depicts one of numerous tiles on the Cerebras CS-1 Wafer Scale Engine, the tile including a processor core ("Core") and a router ("Router") with a set of input ports ("NSEW" and a direct input from the Core) and a set of output ports ("NSEW" and a direct output to the Core).



Cerebras CS-1 Wafer Scale Engine[28]

57. In the Cerebras CS-1, each processor core of the set of processor cores corresponds to a different router of the set of routers, and each processor core and corresponding router form a tile.[29]

58. In the Cerebras CS-1, each processor core is communicatively coupled with a corresponding router via the router's set of input ports and set of output ports, each router is communicatively coupled with one or more adjacent routers via the set of input ports and the set of output ports, and each processor core is communicatively coupled with the other processor cores via the set of routers.[30]

---

[28] Kamil Rocki et al., Fast Stencil-Code Computation on a Wafer Scale Processor 3, SC20 (November 9-19, 2020) ("Rocki"), *available at* https://arxiv.org/pdf/2010.03660.pdf (last visited May 4, 2021).

[29] *See, e.g.*, Groeneveld EDPS Pres. at 14:53−15:29 ("The processing element is connected to an NoC router . . . which allows you to send data packages to your left neighbor, your right neighbor, your top neighbor, and your bottom neighbor, or to pull the data into the processing element, or inject it into the network. That's how every processing element looks like. They're all identical."); Rocki at 2 ("The repeated element of the architecture is called a tile. The tile contains one processor core, its memory, and the router that it connects to. The routers link to the routers of the four neighboring tiles.").

[30] *See, e.g.*, Groeneveld EDPS Pres. at 14:53−15:29 ("The processing element is connected to an NoC router . . . which allows you to send data packages to your left neighbor, your right neighbor, your top neighbor, and your bottom neighbor, or to pull the data into the processing element, or

59. In the Cerebras CS-1, each router is operable to receive one or more data packets from the one or more adjacent routers or the processor core corresponding to the router.[31]

60. In the Cerebras CS-1, each router is operable to receive one or more data packets based on a physical destination address of a data packet, each router is operable to send one or more data packets to the one or more adjacent routers or the processor core corresponding to the router, and each router is operable to retain a data packet in the event of a traffic condition.[32]

61. In the Cerebras CS-1, each router implements a static priority routing policy.[33]

---

inject it into the network. That's how every processing element looks like. They're all identical."); Rocki at 3 ("The core connects to a local router that has five bidirectional links, one to each of its four nearest neighbors and one to its own core. The router can move data into and out of these five links, in parallel on every cycle."); Sven Verdoolaege et al., Generating SIMD Instructions for Cerebras CS-1 using Polyhedral Compilation Techniques 2 ("Verdoolaege"), IMPACT 2020 (January 22, 2020), *available at* http://impact.gforge.inria.fr/impact2020/papers/IMPACT_2020_paper_3.pdf (last visited May 4, 2021) ("The target architecture is an MPPA (Massively Parallel Processor Array), consisting of a 2-dimensional grid of PEs [processing elements] that communicate with their nearest neighbors in the four cardinal directions.").

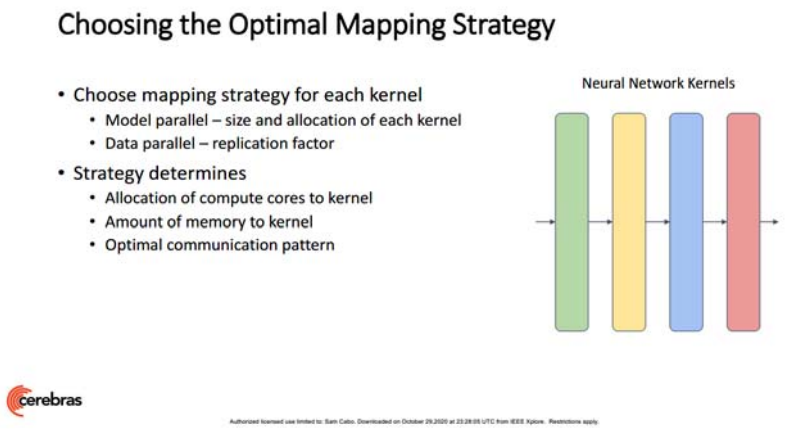
[31] *See, e.g.,* Groeneveld EDPS Pres. at 15:36−16:42 (describing how to send "a data packet or a 'wavelet'"); Rocki at 3 ("The core connects to a local router that has five bidirectional links, one to each of its four nearest neighbors and one to its own core. The router can move data into and out of these five links, in parallel on every cycle.").

[32] *See, e.g.,* Groeneveld EDPS Pres. at 15:36−16:42 (depicting "Wavelet Buffers + forwarding table per buffer"); Rocki at 3 ("The router has hardware queues for its connection to the core and for each of a set of virtual channels, avoiding deadlock."); Manohararajah Pres. at 52:42−53:08 ("[Q:] You sometimes overlay two traffic flows on the same links but you do that by scheduling them clock cycle by clock cycle as part of your compilation? [A:] Correct, uh no that's done but that's done by the hardware. That is not something that the software deals with. [Q:] So it's still buffering and it still figures out when it can send it, the software just figures out it's not going to over allocate it? [A:] Exactly, yeah.").

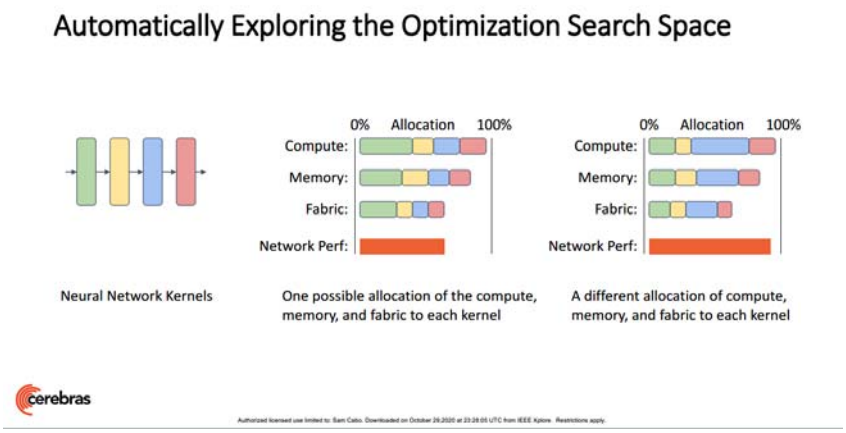
[33] *See, e.g.,* Groeneveld EDPS Pres. at 15:36−16:42 ("Programming these forwarding tables allows us to kind of program any possible routing forwarding in this, on our PEs."); Rocki at 3 ("Communication between potentially distant processors occurs along predetermined routes."); Manohararajah Pres. at 51:52−52:42 ("It's statically configured. It's a virtual channel. So there's some number of virtual channels. Certain virtual channels can talk to certain other virtual channels. So there are some restrictions as to as to how the connectivity is achieved and so that's one of the constraints within the router, so I have to adhere to that constraint").

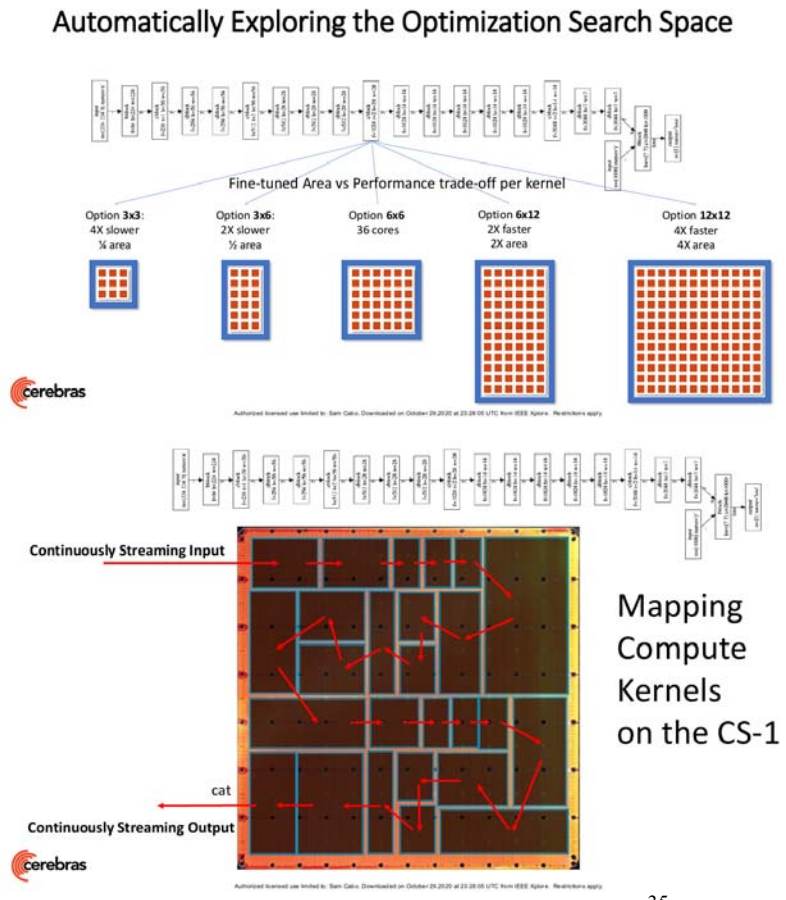
62. The Cerebras CS-1 includes an optimization module configured to determine optimal function (*e.g.*, "kernel") assignment configurations for groups of tiles.[34]

63. The exemplary figures below describe the Cerebras Graph Compiler, which determines optimal function assignment configurations for groups of tiles by "Choosing the Optimal Mapping Strategy" and "Mapping Compute Kernels on the CS-1."



[34] *See, e.g.*, CS-1 Overview at 4 ("Cerebras' software can configure all the cores and routers on the WSE to support a unique, optimized communication pattern for each particular neural network."), 8 ("To translate a deep learning network into an optimized executable, CGC . . . allocates compute and memory to each kernel in the graph and maps every kernel onto a physical region of the computational array of cores. . . . kernel placement is formulated as a multi-constraint problem on 1) memory capacity and bandwidth, 2) computation requirements, and 3) computation costs. The placement engine then takes into account both algorithmic efficiency and compute core utilization to generate a result that maximizes locality, minimizes routing distances, and avoids hotspots.").

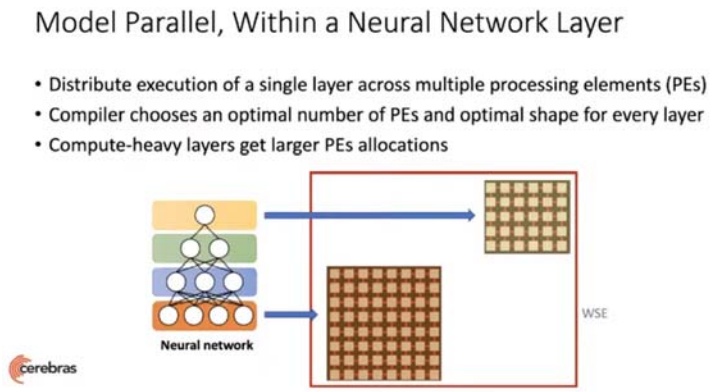


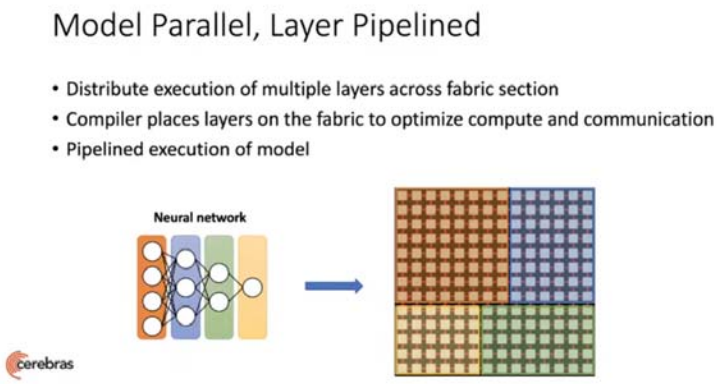


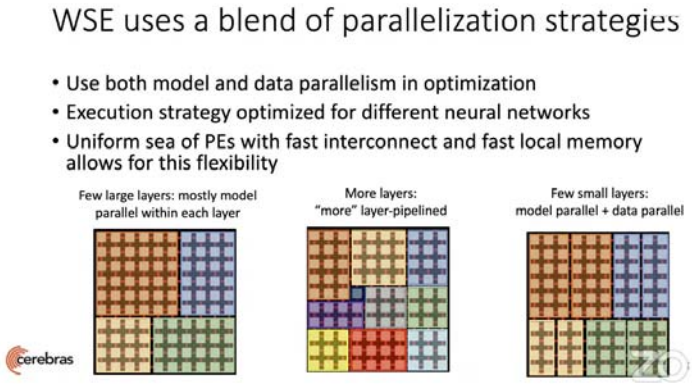
Optimization in the Cerebras CS-1[35]

[35] Cerebras Systems, Software Co-design for the First Wafer-Scale Processor (and Beyond) 8−11 ("Software Co-design"), in 2020 IEEE Hot Chips 32 Symposium (HCS), *available at* https://ieeexplore.ieee.org/document/9220504 (last visited May 4, 2021).

64. The figures below further describe the Cerebras Graph Compiler, which "chooses an optimal number of PEs [processing elements] and optimal shape for every layer" of an application, "places layers on the fabric to optimize compute and communication," and uses an "Execution strategy optimized for different neural networks."







Optimization in the Cerebras CS-1[36]

65. The Cerebras CS-1 optimization module is configured to assign two or more functions, which communicate at least unilaterally more frequently with one another than with other functions, to groups of adjacent tiles based on an optimal function assignment configuration determination.[37]

66. The Cerebras CS-1 optimization module is configured to assign two or more functions . . . wherein the two or more functions are assigned to groups of tiles communicatively coupled in square configurations when the function executes optimally when executed by the groups of tiles communicatively coupled in the square configurations and are assigned to groups of tiles communicatively coupled in linear configurations when the function executes optimally when executed by the groups of tiles communicatively coupled in the linear configurations.[38]

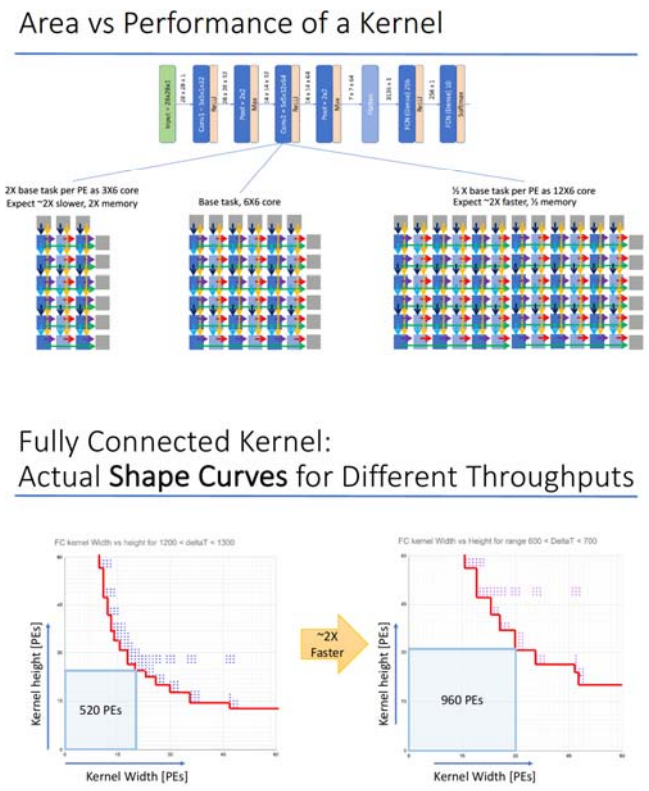
---

[36] Manohararajah Pres. at 17:09–20:12.

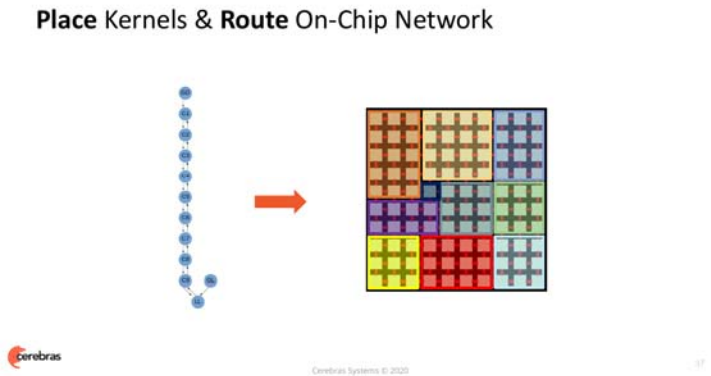
[37] *See, e.g.*, Technical Overview of the Cerebras CS-1 - Neocortex - Pittsburgh Supercomputing Center ("Vassilieva Pres.") at 31:28, https://www.youtube.com/watch?v=4p7Hir6VqZk (last visited May 4, 2021) ("We can maximize their communication between those fractions or those sections of the fabric which should talk to each other for this specific model."), 38:26 ("We configure the network fabric, so we route it to maximize the communication between those sections of the fabric which should talk to each other for this specific model.").

[38] *See, e.g.*, Groeneveld EDPS Pres. at 18:14–18:47 ("Typically a kernel will be implemented as a square or rectangular array of processors on our wafer scale engine."), 19:26–20:16 ("I can either make one that's almost square . . . I can make all kinds of different shapes.").

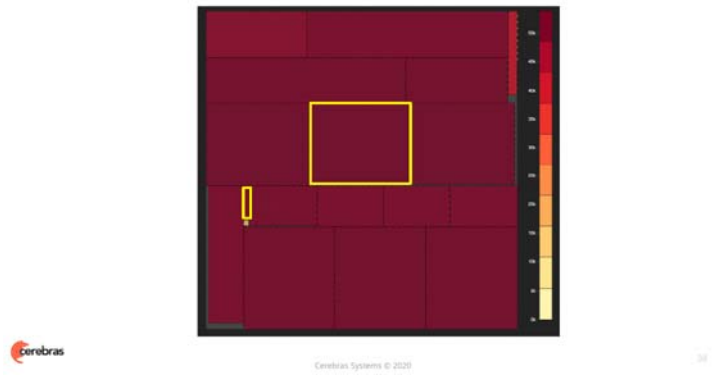
67. The exemplary figures below depict the assignment of functions to groups of tiles having various configurations in the Cerebras CS-1 to optimize execution.



Assignment of Functions in the Cerebras CS-1[39]



[39] Groeneveld at 16, 17.



Assignment of Functions in the Cerebras CS-1[40]

68. On information and belief, with respect to at least claim 1 of the '975 patent, the Cerebras CS-2 includes all relevant features of the Cerebras CS-1. Accordingly, the Cerebras CS-2 satisfies all limitations of at least claim 1 of the '975 patent for the same reasons set forth herein with respect to the Cerebras CS-1. For example, both the Cerebras CS-2 and the Cerebras CS-1 are powered by versions of the Cerebras Wafer Scale Engine and include the Cerebras Swarm communication fabric, which together provide an on-chip network having a set of processor cores and corresponding routers arranged and configured as recited in claim 1 of the '975 patent.[41] In addition, both the Cerebras CS-2 and the Cerebras CS-1 include the Cerebras Graph Compiler, which includes all features of the optimization module recited in claim 1 of the '975 patent.[42]

---

[40] Natalia Vassilieva, Technical Overview of the Cerebras CS-1, the AI Compute Engine for Neocortex 37, 38 (August 19, 2020) ("Vassilieva"), *available at* https://www.cmu.edu/psc/aibd/neocortex/files/neocortex_introcerebras_20200819.pdf (last visited May 4, 2021).

[41] *See* Cerebras CS-2 Overview at 2−4 (describing the Cerebras CS-2 Wafer Scale Engine (WSE-2) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores); CS-1 Overview at 2−4 (describing the Cerebras CS-1 Wafer Scale Engine (WSE) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores).

[42] *See* Cerebras CS-2 Overview at 7−9 (describing the optimization achieved by the Cerebras Graph Compiler, which is present in the Cerebras CS-2); Cerebras CS-1 Overview at 7−9 (describing the optimization achieved by the Cerebras Graph Compiler, which is present in the Cerebras CS-1).

Furthermore, Cerebras' has provided links to several references that describe the similarities between the Cerebras CS-2 and Cerebras CS-1.[43]

69. Cerebras has actively induced others to infringe at least claim 1 of the '975 patent in violation of 35 U.S.C. § 271(b) by instructing others to use certain computer systems, including the Accused Products. Cerebras' active inducement includes, for example and without limitation, marketing, selling, and offering to sell the Accused Products, providing instructions on how to use the Accused Products, and promoting the use of the Accused Products. For example, Cerebras has promoted the use of the Accused Products via its website,[44] and via numerous presentations to customers.[45] Cerebras' customers have directly infringed the '975 patent at least by using the Accused Products.

---

[43] New Archive | Cerebras, https://cerebras.net/news/ (last visited May 4, 2021) (citing Ian Cutress, Cerebras Unveils Wafer Scale Engine Two (WSE2): 2.6 Trillion Transistors, 100% Yield, https://www.anandtech.com/show/16626/cerebras-unveils-wafer-scale-engine-two-wse2-26-trillion-transistors-100-yield (April 20, 2021) ("[T]he second gen WSE will be built into CS-2 systems with a similar design to CS-1"); Nicole Hemsoth, One Giant Leap for Waferscale AI, https://www.nextplatform.com/2021/04/20/one-giant-leap-for-waferscale-ai/ (April 20, 2021) ("[T]he general architecture [of the Cerebras CS-1 and Cerebras CS-2] is the same. . . . there's not a ton of re-engineering that had to go into the CS-2"); Paul Alcorn, Cerebras Second-Gen Wafer Scale Chip: 2.6 Trillion 7nm Transistors, 850,000 Cores, 15kW of Power | Tom's Hardware, https://www.tomshardware.com/news/cerebras-wafer-scale-engine-2-worlds-largest-chip-7nm-850000-cores (April 20, 2021) ("[T]he changes to the first-gen CS-1 system . . . are very minimal in the new CS-2 variant."); Samuel Moore, Cerebras' New Monster AI Chip Adds 1.4 Trillion Transistors, https://spectrum.ieee.org/tech-talk/semiconductors/processors/cerebras-giant-ai-chip-now-has-a-trillions-more-transistors (April 20, 2021) ("The computer system that hosts the WSE 2, called the CS-2, hasn't really changed much either. 'We were able to carry forward significant portions of the physical design [of the Cerebras CS-1],' says Feldman.").

[44] Homepage | Cerebras, https://www.cerebras.net/ (last visited May 4, 2012).

[45] *See, e.g.,* Groeneveld CASS Pres.; Groeneveld EDPS Pres.; Manohararajah Pres.; Vassilieva Pres.; Sean Lie - Wafer-Scale ML, https://www.youtube.com/watch?v=esVqU-hPXxw (last visited May 4, 2021); HC31-S24: ML Training, https://www.youtube.com/watch?v=QF9oObzMBpU&t=3715s (last visited May 4, 2021); Enabling AI's potential through Wafer-scale integration - Andrew Feldman (Cerebras Systems), https://www.youtube.com/watch?v=cfSWv62Hi-o (last visited May 4, 2021); Cerebras' Wafer Scale Engine AI Chip with CEO Andrew Feldman, https://www.youtube.com/watch?v=yso2S2Svdlg; (last visited May 4, 2021).

70. On information and belief, Cerebras has induced such infringement with the specific intent that one or more claims of the '975 patent be infringed, or has been willfully blind to the possibility that its inducing acts would cause the infringing acts.

71. Cerebras has contributed to infringement by others of at least claim 1 of the '975 patent in violation of 35 U.S.C. § 271(c) by selling the Accused Products, each of which is a component of a patented system and which constitutes a material part of the invention in at least claim 1 of the '975 patent. Cerebras has sold the Accused Products knowing the same to be specifically made or especially adapted for use in an infringement of at least claim 1 of the '975 patent, and that the Accused Products are not staple articles or commodities of commerce suitable for substantial noninfringing use.

72. As set forth above, Cerebras has had actual knowledge of the '975 patent prior to the filing of this Complaint. Cerebras has continued to infringe at least claim 1 of the '975 patent. Cerebras' infringement is objectively reckless, knowing, deliberate, and willful.

73. Rex has been damaged as a result of Cerebras' infringing conduct and is entitled to recover damages that adequately compensate it for Cerebras' infringement, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

74. Each of the references cited in Count I of this Complaint in support of the allegations set forth therein was authored, created, prepared, or otherwise provided by Cerebras for the purpose of describing the Accused Products.

COUNT II

(INFRINGEMENT OF U.S. PATENT NO. 10,700,968)

75. Rex realleges paragraphs 1-74 above as if fully set forth herein.

76. The ’968 patent, titled “Optimized Function Assignment in a Multi-Core Processor,” is a valid, enforceable patent that was duly issued by the United States Patent and Trademark Office on June 30, 2020 in full compliance with Title 35 of the United States Code.

77. Rex is the assignee of the ’968 patent with ownership of all substantial rights in the ’968 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

78. The ’968 patent relates to a system and method that includes receiving an application having a set of functions to be executed by a multi-core multiprocessor chip, the chip including a network of tiles, each tile including a core and router pair. An optimal configuration for execution of the set of functions is selected from different configurations, where the different configurations include execution of the functions by different groups of tiles.

79. Cerebras has directly infringed, and continues to directly infringe, at least claim 19 of the ’968 patent in violation of 35 U.S.C. § 271(a) by, for example and without limitation, making, using, offering to sell, selling, and/or importing in and into the United States certain computer systems, including the Accused Products.

80. Claim 19 of the ’968 patent recites a method comprising: receiving a user application, wherein the user application includes a set of functions to be executed by a multi-core microprocessor chip, wherein the multi-core microprocessor chip comprises a set of tiles each including a processor core and a corresponding router, wherein each router: is communicatively coupled with at least one other router to form a network-on-chip grid, and implements the same deterministic static priority routing policy, wherein the deterministic static priority routing policy comprises assigning unchanging priority levels to the input ports of the router and routing outbound data in accordance with the unchanging priority levels; receiving an identification of a

high priority function of the set of functions; identifying one or more tiles with high routing priority according to the static priority routing policy; assigning execution of the high priority function to the one or more tiles with high routing priority; and executing the high priority function in accordance with the assignment.

81. The Cerebras CS-1, for example, is configured to receive a user application, wherein the user application includes a set of functions to be executed by a multi-core microprocessor chip, wherein the multi-core microprocessor chip comprises a set of tiles each including a processor core and a corresponding router.[46]

82. In the Cerebras CS-1, each router is communicatively coupled with at least one other router to form a network-on-chip grid.[47]

---

[46] *See, e.g.,* CS-1 Overview at 1 ("Cerebras is the only company to undertake the ambitious task of designing a system from the ground up to accelerate AI applications."), 8 ("The Cerebras Graph Compiler (CGC) takes as input a user-specified neural network. . . . CGC extracts a static graph representation of the problem from the source language and converts it into the Cerebras Linear Algebra Intermediate Representation (CLAIR). . . . CGC performs a matching and covering operation that matches subgraphs to kernels from the Cerebras kernel library. . . . The result of this matching operation is a kernel graph. CGC then allocates compute and memory to each kernel in the graph and maps every kernel onto a physical region of the computational array of cores. Finally, a communication path, unique to each network, is configured onto the fabric. . . . The final result is a CS-1 executable, customized to the unique needs of each neural network . . . towards accelerating the deep learning application."); Groeneveld EDPS Pres. at 14:53−15:29 ("The processing element is connected to an NoC router . . . which allows you to send data packages to your left neighbor, your right neighbor, your top neighbor, and your bottom neighbor, or to pull the data into the processing element, or inject it into the network. That's how every processing element looks like. They're all identical."); Rocki at 2 ("The repeated element of the architecture is called a tile. The tile contains one processor core, its memory, and the router that it connects to. The routers link to the routers of the four neighboring tiles.")

[47] *See, e.g.,* Groeneveld EDPS Pres. at 14:53−15:29 ("The processing element is connected to an NoC router . . . which allows you to send data packages to your left neighbor, your right neighbor, your top neighbor, and your bottom neighbor, or to pull the data into the processing element, or inject it into the network. That's how every processing element looks like. They're all identical."); Rocki at 3 ("The core connects to a local router that has five bidirectional links, one to each of its four nearest neighbors and one to its own core. The router can move data into and out of these five links, in parallel on every cycle."); Verdoolaege at 2 ("The target architecture is an MPPA

83. In the Cerebras CS-1, each router implements the same deterministic static priority routing policy, wherein the deterministic static priority routing policy comprises assigning unchanging priority levels to the input ports of the router and routing outbound data in accordance with the unchanging priority levels.[48]

84. The Cerebras CS-1 is configured to receive an identification of a high priority function of a set of functions, to identify one or more tiles with high routing priority according to the static priority routing policy, to assign execution of the high priority function to the one or more tiles with high routing priority, and to execute the high priority function in accordance with the assignment.[49]

85. On information and belief, with respect to at least claim 19 of the '968 patent, the Cerebras CS-2 includes all relevant features of the Cerebras CS-1. Accordingly, the Cerebras CS-2 satisfies all limitations of at least claim 19 of the '968 patent for the same reasons set forth herein

---

(Massively Parallel Processor Array), consisting of a 2-dimensional grid of PEs [processing elements] that communicate with their nearest neighbors in the four cardinal directions.").

[48] *See, e.g.,* Groeneveld EDPS Pres. at 15:36−16:42 ("Programming these forwarding tables allows us to kind of program any possible routing forwarding in this, on our PEs."); Rocki at 3 ("Communication between potentially distant processors occurs along predetermined routes."); Manohararajah Pres. at 51:52−52:42 ("It's statically configured. It's a virtual channel. So there's some number of virtual channels. Certain virtual channels can talk to certain other virtual channels. So there are some restrictions as to as to how the connectivity is achieved and so that's one of the constraints within the router, so I have to adhere to that constraint").

[49] *See, e.g.,* Groeneveld EDPS Pres. at 15:36−16:42 ("Programming these forwarding tables allows us to kind of program any possible routing forwarding in this, on our PEs."); Rocki at 3 ("Communication between potentially distant processors occurs along predetermined routes."); Manohararajah Pres. at 51:52−52:42 ("It's statically configured. It's a virtual channel. So there's some number of virtual channels. Certain virtual channels can talk to certain other virtual channels. So there are some restrictions as to as to how the connectivity is achieved and so that's one of the constraints within the router, so I have to adhere to that constraint"), Vassilieva Pres. at 31:28−31:38 ("We can maximize their communication between those fractions or those sections of the fabric which should talk to each other for this specific model."), 38:26−38:35 ("We configure the network fabric, so we route it to maximize the communication between those sections of the fabric which should talk to each other for this specific model.").

with respect to the Cerebras CS-1. For example, both the Cerebras CS-2 and the Cerebras CS-1 are powered by versions of the Cerebras Wafer Scale Engine and include the Cerebras Swarm communication fabric, which together provide an on-chip network having a set of processor cores and corresponding routers arranged and configured as recited in claim 19 of the '968 patent.[50] Furthermore, Cerebras' has provided links to several references that describe the similarities between the Cerebras CS-2 and Cerebras CS-1.[51]

86. Cerebras has actively induced others to infringe at least claim 19 of the '968 patent in violation of 35 U.S.C. § 271(b) by instructing others to use certain computer systems, including the Accused Products. Cerebras' active inducement includes, for example and without limitation, marketing, selling, and offering to sell the Accused Products, providing instructions on how to use the Accused Products, and promoting the use of the Accused Products. For example, Cerebras has

[50] *See* Cerebras CS-2 Overview at 2−4 (describing the Cerebras CS-2 Wafer Scale Engine (WSE-2) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores); CS-1 Overview at 2−4 (describing the Cerebras CS-1 Wafer Scale Engine (WSE) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores).

[51] New Archive | Cerebras, https://cerebras.net/news/ (last visited May 4, 2021) (citing Ian Cutress, Cerebras Unveils Wafer Scale Engine Two (WSE2): 2.6 Trillion Transistors, 100% Yield, https://www.anandtech.com/show/16626/cerebras-unveils-wafer-scale-engine-two-wse2-26-trillion-transistors-100-yield (April 20, 2021) ("[T]he second gen WSE will be built into CS-2 systems with a similar design to CS-1"); Nicole Hemsoth, One Giant Leap for Waferscale AI, https://www.nextplatform.com/2021/04/20/one-giant-leap-for-waferscale-ai/ (April 20, 2021) ("[T]he general architecture [of the Cerebras CS-1 and Cerebras CS-2] is the same. . . . there's not a ton of re-engineering that had to go into the CS-2"); Paul Alcorn, Cerebras Second-Gen Wafer Scale Chip: 2.6 Trillion 7nm Transistors, 850,000 Cores, 15kW of Power | Tom's Hardware, https://www.tomshardware.com/news/cerebras-wafer-scale-engine-2-worlds-largest-chip-7nm-850000-cores (April 20, 2021) ("[T]he changes to the first-gen CS-1 system . . . are very minimal in the new CS-2 variant."); Samuel Moore, Cerebras' New Monster AI Chip Adds 1.4 Trillion Transistors, https://spectrum.ieee.org/tech-talk/semiconductors/processors/cerebras-giant-ai-chip-now-has-a-trillions-more-transistors (April 20, 2021) ("The computer system that hosts the WSE 2, called the CS-2, hasn't really changed much either. 'We were able to carry forward significant portions of the physical design [of the Cerebras CS-1],' says Feldman.").

promoted the use of the Accused Products via its website,[52] and via numerous presentations to customers.[53] Cerebras' customers have directly infringed the '968 patent at least by using the Accused Products.

87. On information and belief, Cerebras has induced such infringement with the specific intent that one or more claims of the '968 patent be infringed, or has been willfully blind to the possibility that its inducing acts would cause the infringing acts.

88. Cerebras has contributed to infringement by others of at least claim 19 of the '968 patent in violation of 35 U.S.C. § 271(c) by selling the Accused Products, each of which is a component of a patented system and which constitutes a material part of the invention in at least claim 19 of the '968 patent. Cerebras has sold the Accused Products knowing the same to be specifically made or especially adapted for use in an infringement of at least claim 19 of the '968 patent, and that the Accused Products are not staple articles or commodities of commerce suitable for substantial noninfringing use.

89. As set forth above, Cerebras has had actual knowledge of the '968 patent prior to the filing of this Complaint. Cerebras has continued to infringe at least claim 19 of the '968 patent. Cerebras' infringement is objectively reckless, knowing, deliberate, and willful.

90. Rex has been damaged as a result of Cerebras' infringing conduct and is entitled to recover damages that adequately compensate it for Cerebras' infringement, which, by law, cannot

---

[52] Homepage | Cerebras, https://www.cerebras.net/ (last visited May 4, 2012).

[53] *See, e.g.,* Groeneveld CASS Pres.; Groeneveld EDPS Pres.; Manohararajah Pres.; Vassilieva Pres.; Sean Lie - Wafer-Scale ML, https://www.youtube.com/watch?v=esVqU-hPXxw (last visited May 4, 2021); HC31-S24: ML Training, https://www.youtube.com/watch?v=QF9oObzMBpU&t=3715s (last visited May 4, 2021); Enabling AI's potential through Wafer-scale integration - Andrew Feldman (Cerebras Systems), https://www.youtube.com/watch?v=cfSWv62Hi-o (last visited May 4, 2021); Cerebras' Wafer Scale Engine AI Chip with CEO Andrew Feldman, https://www.youtube.com/watch?v=yso2S2Svdlg; (last visited May 4, 2021).

be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

91. Each of the references cited in Count II of this Complaint in support of the allegations set forth therein was authored, created, prepared, or otherwise provided by Cerebras for the purpose of describing the Accused Products.

COUNT III

(INFRINGEMENT OF U.S. PATENT NO. 10,127,043)

92. Rex realleges paragraphs 1-91 above as if fully set forth herein.

93. The '043 patent, titled "Implementing Conflict-Free Instructions for Concurrent Operation on a Processor," is a valid, enforceable patent that was duly issued by the United States Patent and Trademark Office on November 13, 2018 in full compliance with Title 35 of the United States Code.

94. Rex is the assignee of the '043 patent with ownership of all substantial rights in the '043 patent, including the right to exclude others and to enforce, sue, and recover damages for past and future infringements.

95. The '043 patent relates to a system and method for implementing very long instruction words (VLIW) having slot instructions that correspond to a set of functional units. Slot instructions include opcodes and value fields having bits that may be allocated to other slot instructions.

96. Cerebras has directly infringed, and continues to directly infringe, at least claim 1 of the '043 patent in violation of 35 U.S.C. § 271(a) by, for example and without limitation, making, using, offering to sell, selling, and/or importing in and into the United States certain computer systems, including the Accused Products.

97. Claim 1 of the ’043 patent recites a system for implementing very long instruction words (VLIW), the system operable to: receive a first very long instruction word (VLIW) comprising a set of slot instructions corresponding to a set of functional units, wherein: each slot instruction includes an opcode identifying an operation to be performed by the set of functional units and value fields related to the operation, wherein a dedicated subset of the value fields include dedicated bits dedicated to the slot instruction and an allocable subset of the value fields include allocable bits allocable to other slot instructions; identify the opcodes of each slot instruction; determine, based on the opcodes, which allocable bits are allocated to which slot instructions; and instruct each functional unit to perform an operation identified by a corresponding slot instruction using the corresponding dedicated bits and any allocable bits determined to be allocated to the slot instruction.

98. On information and belief, the Accused Products are configured to process instructions that correspond to a set of functional units. For example, the Accused Products are configured to process instructions that correspond to arithmetic logic units and/or floating point units included in the processing elements of the Accused Products.

99. On information and belief, those instructions include opcodes that identify an operation to be performed by the set of functional units, and value fields related to the operation.

100. For example, on information and belief, aspects of the Accused Products, including their design, operation, and instruction formats, are described in patent applications filed by Cerebras, including, for example, U.S. Patent Application Publication No. 2020/0380370 to Lie et al. (“Lie”). Lie describes “instructions” having “opcodes” that identify an operation to be

performed, and "operands" that relate to the operation.[54] Functional units included in the Accused Products perform operations based on the opcodes and operands included in those instructions.

101. On information and belief, the opcode and operand fields in the Accused Products' instructions include allocable bits.

102. On information and belief, with respect to at least claim 1 of the '043 patent, the Cerebras CS-2 includes all relevant features of the Cerebras CS-1. Accordingly, the Cerebras CS-2 satisfies all limitations of at least claim 1 of the '043 patent for the same reasons set forth herein with respect to the Cerebras CS-1. For example, both the Cerebras CS-2 and the Cerebras CS-1 are powered by versions of the Cerebras Wafer Scale Engine and include the Cerebras Swarm communication fabric, which together provide an on-chip network having a set of processor cores and corresponding routers arranged and configured as recited in claim 1 of the '043 patent.[55] Furthermore, Cerebras' has provided links to several references that describe the similarities between the Cerebras CS-2 and Cerebras CS-1.[56]

---

[54] *See, e.g.,* U.S. Pat. App. Pub. No. 2020/0380370, ¶¶ [0496]–[0560], [0674]–[0697].

[55] *See* Cerebras CS-2 Overview at 2–4 (describing the Cerebras CS-2 Wafer Scale Engine (WSE-2) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores); CS-1 Overview at 2–4 (describing the Cerebras CS-1 Wafer Scale Engine (WSE) as containing compute cores and describing the Swarm communication fabric as providing a hardware routing engine to each of the cores).

[56] New Archive | Cerebras, https://cerebras.net/news/ (last visited May 4, 2021) (citing Ian Cutress, Cerebras Unveils Wafer Scale Engine Two (WSE2): 2.6 Trillion Transistors, 100% Yield, https://www.anandtech.com/show/16626/cerebras-unveils-wafer-scale-engine-two-wse2-26-trillion-transistors-100-yield (April 20, 2021) ("[T]he second gen WSE will be built into CS-2 systems with a similar design to CS-1"); Nicole Hemsoth, One Giant Leap for Waferscale AI, https://www.nextplatform.com/2021/04/20/one-giant-leap-for-waferscale-ai/ (April 20, 2021) ("[T]he general architecture [of the Cerebras CS-1 and Cerebras CS-2] is the same. . . . there's not a ton of re-engineering that had to go into the CS-2"); Paul Alcorn, Cerebras Second-Gen Wafer Scale Chip: 2.6 Trillion 7nm Transistors, 850,000 Cores, 15kW of Power | Tom's Hardware, https://www.tomshardware.com/news/cerebras-wafer-scale-engine-2-worlds-largest-chip-7nm-850000-cores (April 20, 2021) ("[T]he changes to the first-gen CS-1 system . . . are very minimal in the new CS-2 variant."); Samuel Moore, Cerebras' New Monster AI Chip Adds 1.4 Trillion Transistors, https://spectrum.ieee.org/tech-talk/semiconductors/processors/cerebras-giant-ai-chip-

103. Cerebras has actively induced others to infringe at least claim 1 of the '043 patent in violation of 35 U.S.C. § 271(b) by instructing others to use certain computer systems, including the Accused Products. Cerebras' active inducement includes, for example and without limitation, marketing, selling, and offering to sell the Accused Products, providing instructions on how to use the Accused Products, and promoting the use of the Accused Products. For example, Cerebras has promoted the use of the Accused Products via its website,[57] and via numerous presentations to customers.[58] Cerebras' customers have directly infringed the '043 patent at least by using the Accused Products.

104. On information and belief, Cerebras has induced such infringement with the specific intent that one or more claims of the '043 patent be infringed, or has been willfully blind to the possibility that its inducing acts would cause the infringing acts.

105. Cerebras has contributed to infringement by others of at least claim 1 of the '043 patent in violation of 35 U.S.C. § 271(c) by selling the Accused Products, each of which is a component of a patented system and which constitutes a material part of the invention in at least claim 1 of the '043 patent. Cerebras has sold the Accused Products knowing the same to be specifically made or especially adapted for use in an infringement of at least claim 1 of the '968

---

now-has-a-trillions-more-transistors (April 20, 2021) ("The computer system that hosts the WSE 2, called the CS-2, hasn't really changed much either. 'We were able to carry forward significant portions of the physical design [of the Cerebras CS-1],' says Feldman.").

[57] Homepage | Cerebras, https://www.cerebras.net/ (last visited May 4, 2012).

[58] *See, e.g.,* Groeneveld CASS Pres.; Groeneveld EDPS Pres.; Manohararajah Pres.; Vassilieva Pres.; Sean Lie - Wafer-Scale ML, https://www.youtube.com/watch?v=esVqU-hPXxw (last visited May 4, 2021); HC31-S24: ML Training, https://www.youtube.com/watch?v=QF9oObzMBpU&t=3715s (last visited May 4, 2021); Enabling AI's potential through Wafer-scale integration - Andrew Feldman (Cerebras Systems), https://www.youtube.com/watch?v=cfSWv62Hi-o (last visited May 4, 2021); Cerebras' Wafer Scale Engine AI Chip with CEO Andrew Feldman, https://www.youtube.com/watch?v=yso2S2Svdlg; (last visited May 4, 2021).

patent, and that the Accused Products are not staple articles or commodities of commerce suitable for substantial noninfringing use.

106. As set forth above, Cerebras has had actual knowledge of the '043 patent prior to the filing of this Complaint. Cerebras has continued to infringe at least claim 1 of the '043 patent. Cerebras' infringement is objectively reckless, knowing, deliberate, and willful.

107. Rex has been damaged as a result of Cerebras' infringing conduct and is entitled to recover damages that adequately compensate it for Cerebras' infringement, which, by law, cannot be less than a reasonable royalty, together with interest and costs as fixed by this Court under 35 U.S.C. § 284.

108. Each of the references cited in Count III of this Complaint in support of the allegations set forth therein was authored, created, prepared, or otherwise provided by Cerebras for the purpose of describing the Accused Products.

PRAYER FOR RELIEF

WHEREFORE, Rex respectfully requests the following relief:

A. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has infringed, induced infringement, and contributed to infringement of one or more claims of the '975 patent and declaring that Cerebras' importing, making, using, offering to sell, or selling the Accused Products in the United States are and would be acts of infringement of one or more claims of the '975 patent;

B. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has willfully infringed one or more claims of the '975 patent;

C. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras and its officers, employees, agents, attorneys, affiliates, successors, assigns, and others acting in privity

or concert with it be enjoined from making, using, offering to sell, and selling or inducing or inducing or contributing to others to make, use, offer to sell, or sell any product that infringes the '975 patent, including the Accused Products and from importing the same into the United States;

D. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has infringed, induced infringement, and contributed to infringement of one or more claims of the '968 patent and declaring that Cerebras' importing, making, using, offering to sell, or selling the Accused Products in the United States are and would be acts of infringement of one or more claims of the '968 patent;

E. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has willfully infringed one or more claims of the '968 patent;

F. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras and its officers, employees, agents, attorneys, affiliates, successors, assigns, and others acting in privity or concert with it be enjoined from making, using, offering to sell, and selling or inducing or inducing or contributing to others to make, use, offer to sell, or sell any product that infringes the '968 patent, including the Accused Products and from importing the same into the United States;

G. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has infringed, induced infringement, and contributed to infringement of one or more claims of the '043 patent and declaring that Cerebras' importing, making, using, offering to sell, or selling the Cerebras CS-1 in the United States are and would be acts of infringement of one or more claims of the '043 patent;

H. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras has willfully infringed one or more claims of the '043 patent;

I. The entry of a judgment in favor of Rex, and against Cerebras, that Cerebras and its officers, employees, agents, attorneys, affiliates, successors, assigns, and others acting in privity or concert with it be enjoined from making, using, offering to sell, and selling or inducing or inducing or contributing to others to make, use, offer to sell, or sell any product that infringes the '043 patent, including the Accused Products and from importing the same into the United States;

J. The entry of a judgment awarding Rex damages resulting from Cerebras' infringement in an amount no less than a reasonable royalty, together with pre-judgment and post-judgment interests and costs, as fixed by the Court.

K. The entry of a judgment that the damages be enhanced based on Cerebras' willful infringement;

L. The entry of a judgment declaring that this is an exceptional case and awarding Rex its attorney fees in this matter pursuant to 35 U.S.C. § 285.

M. Such other relief as this Court deems just and proper.

JURY DEMAND

Rex hereby demands trial by jury in this action on all issues so triable.

MORRIS, NICHOLS, ARSHT & TUNNELL LLP

/s/ Brian P. Egan

Brian P. Egan (#6227)
P.O. Box 1347
Wilmington, DE 19899
(302) 658-9200
began@morrisnichols.com

Attorneys for Plaintiff Rex Computing, Inc.

OF COUNSEL:

Michael A. Berta
Nicholas Lee
ARNOLD & PORTER KAYE SCHOLER LLP
777 South Figueroa Street, 44th Floor
Los Angeles, CA 90017-5844
(213) 243-4000

Nicholas M. Nyemah
ARNOLD & PORTER KAYE SCHOLER LLP
601 Massachusetts Ave, NW
Washington, DC 20001-3743
(202) 942-5000

Mark Samartino
ARNOLD & PORTER KAYE SCHOLER LLP
70 West Madison Street Suite 4200
Chicago, IL 60602-4231
(312) 583-2300

May 4, 2021