

(51) International Patent Classification:
G06F 19/10 (2011.01) *G06F 19/22* (2011.01)(21) International Application Number:
PCT/IB2016/050841(22) International Filing Date:
17 February 2016 (17.02.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1090/CHE/2015 5 March 2015 (05.03.2015) IN(71) Applicant: INDIAN INSTITUTE OF SCIENCE
[IN/IN]; C V Raman Road, Bangalore, Karnataka 560012 (IN).

(72) Inventors: NATARAJAN, Santhi; Centre for Nano Science and Engineering, Indian Institute of Science, C V Raman Road, Bangalore, Karnataka 560012 (IN). PAL, Debnath; Supercomputer Education and Research Centre, Indian Institute of Science, C V Raman Road, Bangalore, Karnataka 560012 (IN). NANDY, S.k.; Supercomputer Education and Research Centre, Indian Institute of Science, C V Raman Road, Bangalore, Karnataka 560012 (IN).

(74) Agent: KHURANA & KHURANA, ADVOCATES & IP ATTORNEYS; E-13, UPSIDC, Site-IV, Behind-Grand Venice, Kasna Road, Greater Noida, National Capital Region, Uttar Pradesh 201310 (IN).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: DATA STREAMING IN HARDWARE ACCELERATOR FOR MAPPING AND ALIGNMENT OF SHORT READS

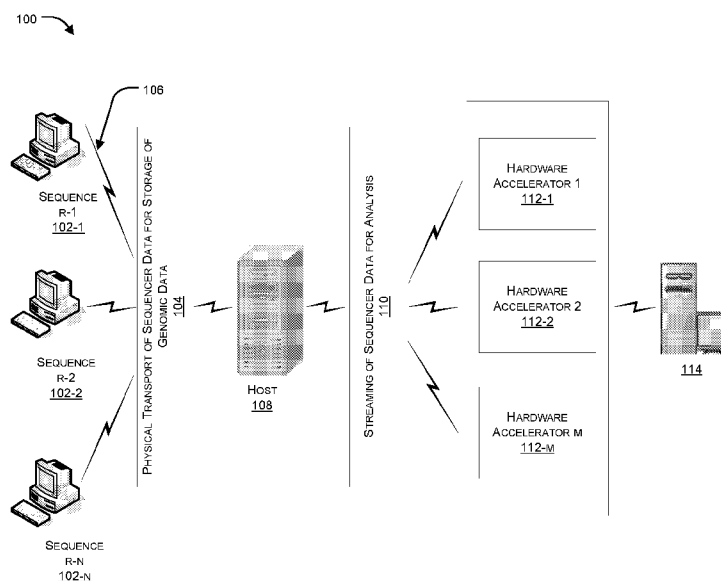


FIG. 1

(57) Abstract: The present disclosure relates to a scalable hardware accelerator for mapping and aligning of short reads with a reference genome that supports streaming of genomic data from host, and drastically reduces the storage requirement within the accelerator platform, thus overcoming the bottleneck created by storage and retrieval of data and making the mapping and aligning process faster and more efficient.

DATA STREAMING IN HARDWARE ACCELERATOR FOR MAPPING AND
ALIGNMENT OF SHORT READS

TECHNICAL FIELD

[1] The present disclosure generally relates to the field of bioinformatics and molecular biology. In particular, it pertains to a scalable hardware accelerator to map and align genomic data.

BACKGROUND

[2] Background description includes information that may be useful in understanding the present invention. It is not an admission that any of the information provided herein is prior art or relevant to the presently claimed invention, or that any publication specifically or implicitly referenced is prior art.

[3] Latest technical advances in genomic sequencing have revolutionized many aspects of biology and medicine. These advances have dramatically lowered the cost and exponentially increased the throughput of DNA sequencing. As a result sequencing technology is now being applied to a rapidly widening array of scientific and medical problems, from basic biology to forensics, ecology, evolutionary studies, agriculture, drug discovery, and the growing field of personalized medicine.

[4] Sequencing machines determine the nucleotide sequence of short DNA fragments, typically a few tens to hundreds of bases, called short reads. With present day sequencing technologies this can be done in a massively parallel manner, yielding much higher throughput than older sequencing technologies – on the order of tens of billions of bases per day from one machine. For comparison, the human genome is approximately 3 billion bases in length.

[5] For most applications, a complete genetic sequence of an organism is not determined de novo. Rather, in most instances, for the organism in question, a “reference” genome sequence has already been determined and is known. Since the short reads are derived from randomly fragmenting genome of one organism for which a reference genome sequence is already known, the first step for data analysis is ordering of all of these fragments to determine the overall gene sequence of the individual sample using the reference genome

sequence effectively as a template, i.e., mapping these short read fragments to the reference genome sequence. In this analysis, a determination is made concerning the best location in the reference genome to which each short read maps, and is referred to as the short read mapping problem.

[6] Short read mapping problem is technically challenging, both due to the volume of data and because sample sequences may not be identical to the reference genome sequence, but as expected, will contain a wide variety of individual genetic variations. Due to the sheer volume of data, e.g., a billion short reads from a single sample, the speed or runtime of the data analysis is significant, with the data analysis now becoming the effective bottleneck in genome sequencing. In addition, successful sequencing should exhibit sensitivity to genetic variations to successfully map sequences that are not completely identical to the reference, both because of technical errors in the sequencing and because of genetic differences between the subject and the reference genome.

[7] Biologists and other researchers use sequence alignment as a fundamental comparison method to find common patterns between sequences, predict protein structure, identify important genetic regions, and facilitate drug design. For example, sequence alignment is used to derive flu vaccines by identifying DNA signatures of pathogens. Since biological sequence alignment is now an essential tool used in molecular biology and biomedical applications it is essential that alignment results are available in a timely manner. The growing volume of genomic data and the complexity of sequence alignment present a challenge in obtaining accurate alignment results in a timely manner.

[8] A number of softwares are available in art that perform short read alignment with the reference genome for example BWA, Novoalign, Bowtie, SOAP2, BFAST, SSAHA2, Mpscan, GASSST, Churchill etc. These software based approaches have number of limitations such as use of heuristic algorithms for mapping that reduces the accuracy as compared to exact algorithms. In addition, they take more time to perform alignment of millions of short reads, making short read mapping the major task affecting the throughput and performance of the sequencing pipeline.

[9] Very few attempts have been made to develop short read mapping accelerators in hardware. One such model for short read mapping has been developed based on research sponsored by Washington Technology Center (WTC) and Pico Computing. This platform performs mapping of 50 million 76 bp short reads from one of paired end Illumina GA IIx

run on human exome data. The hardware is based on a 24- FPGA Pico Computing system. The platform uses BFAST algorithm for indexing, and Smith Waterman and Needleman Wunsch Algorithms for scoring. However this platform is not scalable and time taken for alignment is decided by problem size. Furthermore, the accuracy is compromised due to heuristics involved.

[10] Smith Waterman algorithm is a dynamic programming method for determining similarity between a pair of nucleotide or protein sequences. It ensures the best optimal alignment between sequences. The algorithm is used for performing pairwise local alignment of DNA or protein sequences. A pairwise alignment finds highly related subsequences of two sequences. It identifies subsequences that are preserved during the course of evolution. It is highly useful for dissimilar sequences suspected to contain regions of similarity within their larger sequence context. The alignment need not include entire length of the two sequences. The method is very sensitive in detecting similarity between two sequences sharing evolutionary origin along the entire length, while part of the sequence under strong enough selection pressure to preserve valid similarity.

[11] There have been several attempts to accelerate the Smith Waterman Dynamic programming algorithm in hardware. However, these implementations suffer from various short comings such as sequence length considered for alignment is limited by the hardware size, the architectures are not inherently scalable, they do not perform traceback with forward scan in overlapped mode, their performance is limited by hardware I/O bandwidth, they have severe processing overhead in software when alignment matrix is recalculated. Besides they also have severe memory bottleneck issues.

[12] There is therefore a need for a solution that overcomes the drawbacks of the known methods and provides a hardware accelerator for accurate mapping and alignment of short reads in high throughput sequencing platforms.

[13] All publications herein are incorporated by reference to the same extent as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Where a definition or use of a term in an incorporated reference is inconsistent or contrary to the definition of that term provided herein, the definition of that term provided herein applies and the definition of that term in the reference does not apply.

[14] In some embodiments, the numbers expressing quantities of ingredients, properties such as concentration, reaction conditions, and so forth, used to describe and claim certain embodiments of the invention are to be understood as being modified in some instances by the term “about.” Accordingly, in some embodiments, the numerical parameters set forth in the written description and attached claims are approximations that can vary depending upon the desired properties sought to be obtained by a particular embodiment. In some embodiments, the numerical parameters should be construed in light of the number of reported significant digits and by applying ordinary rounding techniques. Notwithstanding that the numerical ranges and parameters setting forth the broad scope of some embodiments of the invention are approximations, the numerical values set forth in the specific examples are reported as precisely as practicable. The numerical values presented in some embodiments of the invention may contain certain errors necessarily resulting from the standard deviation found in their respective testing measurements.

[15] As used in the description herein and throughout the claims that follow, the meaning of “a,” “an,” and “the” includes plural reference unless the context clearly dictates otherwise. Also, as used in the description herein, the meaning of “in” includes “in” and “on” unless the context clearly dictates otherwise.

[16] The recitation of ranges of values herein is merely intended to serve as a shorthand method of referring individually to each separate value falling within the range. Unless otherwise indicated herein, each individual value is incorporated into the specification as if it were individually recited herein. All methods described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context. The use of any and all examples, or exemplary language (e.g. “such as”) provided with respect to certain embodiments herein is intended merely to better illuminate the invention and does not pose a limitation on the scope of the invention otherwise claimed. No language in the specification should be construed as indicating any non-claimed element essential to the practice of the invention.

Groupings of alternative elements or embodiments of the invention disclosed herein are not to be construed as limitations. Each group member can be referred to and claimed individually or in any combination with other members of the group or other elements found herein. One or more members of a group can be included in, or deleted from, a group for reasons of convenience and/or patentability. When any such inclusion or deletion occurs, the

specification is herein deemed to contain the group as modified thus fulfilling the written description of all Markush groups used in the appended claims.

OBJECTS OF THE INVENTION

[17] An object of the present disclosure is to overcome the drawbacks of existing methods of short read mapping and alignment with a reference genome sequence.

[18] Another object of the present disclosure is to provide a hardware accelerator that overcomes the drawbacks of existing methods of short read mapping and alignment with a reference genome sequence.

[19] Another object of the present disclosure is to provide an accelerator architecture that reduces storage requirement while indexing, mapping and alignment, thus overcoming bottleneck in existing methods.

[20] Another object of the present disclosure is to provide an accelerator architecture that supports streaming of genomic data from host, thereby drastically reducing storage requirements within the accelerator platform and speeds up the mapping and alignment of short reads with a reference genome.

[21] Another object of the present disclosure is to provide a hardware accelerator that can co-exist in a variety of streaming infrastructures.

SUMMARY

[22] Aspects of the present disclosure relate to streaming of genomic data to a hardware accelerator for alignment of short reads with a reference genome in high throughput sequencing platforms. In an aspect, the disclosure provides a hardware accelerator that can map and align short reads as genomic data is streamed, thus speeding up the process.

[23] In an aspect, the present disclosure provides an accelerator architecture that can support streaming of genomic data from a host, thereby drastically reducing storage requirements within the accelerator platform. In an aspect, current software and hardware mapping and alignment techniques require tremendous storage for genomic data, reference genome indices, short reads, hash tables and pointers. In conventional hardware acceleration platforms, these storage requirements continue in the form of memory requirements within

the accelerator (FPGA/ASIC), as rich DDR memories on board and as secondary storage requirements. This results in much of the processing time being spent in retrieving data from storage and handling memory bottleneck issues. The disclosed accelerator architecture overcomes these bottlenecks and speeds up the process of short read mapping and alignment.

[24] In another aspect of the present disclosure, the genomic data can be pre-processed including a onetime indexing of the reference genome to create a Reference Index Table (RIT). Processed data can then be accurately slotted and scheduled within a frame as specified by relevant streaming protocol standards. In another aspect, streaming of the genomic data can be done over any existing protocol, methodology or streaming technologies such as but not limited to built in hardware interconnects like PCIe, physical Ethernet connectivity with configurable speeds over multiple connectivity models, wireless data transmission models or cloud infrastructure.

[25] In another aspect of the present disclosure, the streaming set up can have a cluster of hardware accelerators with a local host storing the genomic data and genomic data server; and the accelerators can co-exist in a variety of streaming infrastructure.

[26] In another aspect, the architecture for hardware accelerator can incorporate a stream receive block (also referred to as Stream RX hereinafter) that receives the pre-processed genomic data, wherein Stream RX can validate the data to authenticate the stream, extract the genomic data including short reads and RITs that are embedded within the stream, and schedule the extracted data to mapping and alignment datapath. In another aspect, the Stream RX block adheres to streaming protocol guidelines and standards.

[27] In another aspect of the disclosure, the hardware accelerator can include multiple datapaths for mapping and alignment, which can run in parallel on the scheduled data. Each datapath can include a Short Read to Reference Mapper (interchangeably referred to simply as mapper) and Short Read Aligner (also referred to simply as aligner). In another aspect, for an exemplary 32bp read-RIT pair, the mapper configured to a preferred embodiment configured to map up to two mismatches, can spend 3 clock cycles to decide the mapping, and aligner configured to a preferred embodiment can spend 16 clock cycles to give score and final alignment. Therefore, in an aspect, the streaming interface can be clocked in tandem when compared with the mapper and aligner so as to maintain the performance and streaming capabilities.

[28] In an aspect, the hardware accelerator can function with independent, decoupled mapper and aligner models as well, where the mapper alone is present in the pipeline initially, which can perform mapping and filtering of read-RIT pairs, which are probable candidates for alignment. Thereafter, the filtered read-RIT pairs can be streamed to the accelerator again, which has only aligner present in the datapath, which can perform alignment on the incoming pairs, and produce the scores and associated results for each pair for streaming as output data.

[29] In another aspect, proposed architecture for hardware accelerator can incorporate a Stream Transmission Block (TX block) that can collect results of alignment from parallel datapaths and prepare them to be streamed to the host. The TX block can first schedule and slot the results to form a payload for the stream as per the standards and guidelines of the streaming protocol chosen, and then embed the payload within a transmit frame, along with the headers. Further, the TX block can validate the transmit-ready frame with a CRC, and stream it out to the host for assessment.

[30] In another aspect, streams from the hardware accelerator containing frames embedding results of alignment can be received by the host, wherein the host can then post-process the received stream to retrieve results and store them for further analysis and interpretation.

[31] Various objects, features, aspects and advantages of the inventive subject matter will become more apparent from the following detailed description of preferred embodiments, along with the accompanying drawing figures in which like numerals represent like components.

BRIEF DESCRIPTION OF THE DRAWINGS

[32] The accompanying drawings are included to provide a further understanding of the present disclosure, and are incorporated in and constitute a part of this specification. The drawings illustrate exemplary embodiments of the present disclosure and, together with the description, serve to explain the principles of the present disclosure.

[33] FIG. 1 illustrates an exemplary block diagram of general computational infrastructure for data transport from sequencing platforms for storage and analysis including the hardware accelerator for genomic data alignment in the data path in accordance with embodiments of the present disclosure.

[34] FIG. 2 illustrates an exemplary flow diagram indicating preprocessing of genomic data for Ethernet streaming adaptation in accordance with embodiments of the present disclosure.

[35] FIG. 3 illustrates an exemplary block diagram indicating hardware accelerator cluster in streaming connectivity in accordance with embodiments of the present disclosure.

[36] FIG. 4 illustrates an exemplary block diagram of hardware accelerator architecture with the streaming interface in accordance with embodiments of the present disclosure.

[37] FIG. 5 illustrates an exemplary block diagram indicating host-hardware accelerator connectivity in accordance with embodiments of the present disclosure.

[38] FIG. 6 illustrates an exemplary block diagram indicating hardware accelerator architecture to support Ethernet streaming in accordance with embodiments of the present disclosure.

[39] FIG. 7 illustrates an exemplary flow diagram indicating post processing of streamed results at host in accordance with embodiments of the present disclosure.

DETAILED DESCRIPTION

[40] The following is a detailed description of embodiments of the disclosure depicted in the accompanying drawings. The embodiments are in such detail as to clearly communicate the disclosure. However, the amount of detail offered is not intended to limit the anticipated variations of embodiments; on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the present disclosure as defined by the appended claims.

[41] Each of the appended claims defines a separate invention, which for infringement purposes is recognized as including equivalents to the various elements or limitations specified in the claims. Depending on the context, all references below to the "invention" may in some cases refer to certain specific embodiments only. In other cases it will be recognized that references to the "invention" will refer to subject matter recited in one or more, but not necessarily all, of the claims.

[42] Various terms as used herein are shown below. To the extent a term used in a claim is not defined below, it should be given the broadest definition persons in the pertinent

art have given that term as reflected in printed publications and issued patents at the time of filing.

[43] Embodiments of the present disclosure relate to streaming of genomic data to a hardware accelerator for mapping and alignment of short reads with a reference genome in high throughput sequencing platforms. In an aspect, the present disclosure provides a hardware accelerator that can map and align short reads as genomic data is streamed, thereby speeding up the process.

[44] Embodiments disclosed herein provide an accelerator architecture supporting streaming of genomic data from host, thereby drastically reducing storage requirements within the accelerator platform. In an aspect, conventional software and hardware alignment techniques require tremendous storage for genomic data, reference genome indices, short reads, hash tables and pointers. In these hardware acceleration platforms, the storage requirements continue in the form of memory requirements within the accelerator (FPGA/ASIC), as rich DDR memories on board and as secondary storage requirements. This results in much of the processing time being spent in retrieving data from storage, and handling memory bottleneck issues. The disclosed accelerator architecture that supports streaming of genomic data from host, drastically reduces the storage requirement within the accelerator platform, thus overcoming the bottleneck created by storage and retrieval of data and making the mapping and aligning process faster and more efficient.

[45] In an embodiment, the genomic data can be pre-processed including one time indexing of the reference genome to create a Reference Index Table (RIT). Processed data can then be accurately slotted and scheduled within a frame as specified by relevant streaming protocol standards. In another aspect, streaming of the genomic data can be done over any existing protocol, methodology or streaming technologies such as but not limited to built in hardware interconnects like PCIe, physical Ethernet connectivity with configurable speeds over multiple connectivity models, wireless data transmission models or cloud infrastructure.

[46] In another aspect of the present disclosure, the streaming set up can have a cluster of hardware accelerators hosting many hardware accelerators that can run in parallel with a local host storing the genomic data and genomic data server wherein the accelerators can co-exist in a variety of streaming infrastructure.

[47] In an embodiment, architecture for the proposed hardware accelerator can include a stream receive block (also referred to as Stream RX hereinafter) that can receive the pre-processed genomic data and validate the data to authenticate the stream. It can further extract the genomic data including short reads and RITs that are embedded within the stream and schedule the extracted data to mapping and alignment datapath. In an aspect, the Stream RX block can adhere to streaming protocol guidelines and standards.

[48] In an embodiment, the hardware accelerator can include multiple datapaths for mapping and alignment that can run in parallel on the scheduled data. Each datapath can further include a Short Read to Reference Mapper and Short Read Aligner.

[49] In an embodiment, the streaming interface can be clocked in tandem to mapper and aligner. For an exemplary 32bp read-RIT pair, the mapper configured to a preferred embodiment can spend 3 clock cycles to decide the mapping configured to map up to two mismatches, and aligner configured to a preferred embodiment can spend 16 clock cycles to give score and final alignment. Therefore, streaming interface that is clocked in tandem to mapper and aligner can maintain the performance and streaming capabilities through multiple datapaths for mapping and alignment running in parallel.

[50] In an alternate embodiment, the hardware accelerator can function with independent, decoupled mapper and aligner models as well, where the mapper alone is present in the pipeline initially, which can perform mapping and filtering of read-RIT pairs, which are probable candidates for alignment. Thereafter, the filtered read-RIT pairs can be streamed to the accelerator again, which has only aligner present in the datapath, which can perform alignment on the incoming pairs, and produce the scores and associated results for each pair for streaming as output data.

[51] In an embodiment, architecture for the hardware accelerator can include a stream transmission block (referred to as TX block hereinafter) that can collect results of alignment from parallel datapaths and prepare them to be streamed to the host. It can first schedule and slot the results to form a payload for the stream, as per the standards and guidelines of the streaming protocol chosen. It can further embed the payload within a transmit frame along with the headers. It can then validate the transmit-ready frame with a CRC and stream it out to the host for assessment.

[52] In an embodiment, the disclosed architecture wherein the hardware accelerator supports streaming of genomic data is inherently scalable because there can be a cluster of

hardware accelerators, and within each hardware accelerator, there can be multiple datapaths for mapping and alignment running in parallel and streaming interface clocked in tandem to mapper and aligner. Furthermore, read length is not limited by the reconfigurable hardware size. Time taken by the accelerator is independent of the problem size, with streaming limitations posed only from interconnections, bus architecture etc.

[53] In an aspect the disclosed hardware architecture is a high availability solution that exploits re-configurability of the target platform on which it is realized. As would be apparent to a person skilled in art, the disclosed coarse grain reconfigurable architecture that hosts the parallel pipeline of hardware kernels can provide the necessary and sufficient features to make the design fault tolerant.

[54] In another embodiment the streams from hardware accelerator containing frames embedding results of alignment can be received by host, wherein the host can then process the received stream to retrieve results and store them for further analysis and interpretation.

[55] System for mapping and aligning a short read with a reference genome of the present disclosure can include a number of functional modules such as a streaming module, a mapping module and an alignment module. The streaming module can be configured to stream the short read and data of the reference genome through system hardware. Prior to streaming, the reference genome can be indexed to generate a Reference Index Table. The short read and reference genome data can be slotted and scheduled within a frame based on a streaming protocol. The streaming can be done over any existing protocol, methodology, or streaming technology.

[56] The streaming module in turn can include a stream receive block to extract the short read and the data of the reference genome that is embedded within the stream and schedule the extracted data onto one or more datapaths. Each of the one or more parallel datapaths can include a mapping module and an alignment module. The streaming module can be clocked in tandem to said mapper module and said alignment module. The streaming module can further include a stream transmission block to collect results of alignment from parallel datapaths and prepare them for streaming to a host.

[57] The mapping module of the system can be configured to receive the short reads and the reference genome data from the stream receive block and map the short read with the reference genome. The alignment module can be configured to align a mapped short read with reference genome and transmit results of alignment to the stream transmission block.

[58] Method for mapping and aligning a short read with a reference genome can include the steps of receiving a stream having the short read and data of the reference genome obtained by one time indexing of the reference genome. The short read and the reference genome data can be extracted from the stream for scheduling them through one or more datapaths. Each of these datapath can incorporate a mapping module and an alignment module where the short read can be mapped with the reference genome followed by alignment of mapped short read with the reference genome. The results of alignment from parallel can be collected and streamed to a host.

[59] FIG. 1 illustrates an exemplary block diagram 100 of general computational infrastructure for data transport from sequencing platforms for storage and analysis including the hardware accelerator for genomic data alignment in the data path in accordance with embodiments of the present disclosure. Sequencing research/clinical facilities such as 102-1, 102-2,..... 102-N (collectively referred to as 102 hereinafter) typically produce terabytes of data, creating a huge demand for data storage coupled with server facility for secured access of data for analysis. These storage facilities such as 104 can be centralized or local storages, which demands physical transfer 106 of raw data from sequencing platform to storages 104. Once stored, the data can be again physically transferred to a network, cloud, or custom made facility for genomic data alignment and visualization and preparation of final report for application.

[60] In an embodiment, such a facility for genomic data alignment and visualization and preparation of final report for application can include a host 108 that can pre-process the genomic data including one time indexing of the reference genome to create a Reference Index Table (RIT). Processed data can then be accurately slotted and scheduled within a frame as specified by relevant streaming protocol standards. In another aspect, streaming of the genomic data can be done over any existing protocol, methodology or streaming technologies such as but not limited to built-in hardware interconnects like PCIe, physical Ethernet connectivity with configurable speeds over multiple connectivity models, wireless data transmission models or cloud infrastructure

[61] FIG. 2 illustrates an exemplary flow diagram 200 indicating preprocessing of genomic data for Ethernet streaming adaptation in accordance with embodiments of the present disclosure. The raw reads and reference genome are typically in the form of text strings that can be pre-processed to embed them in frame format of Ethernet. The raw reads and reference genome can be received and stored at host 108 at steps 202 and 204,

respectively. At step 206, a RIT of the reference genome can be created by one time indexing. At step 208, RIT elements and reads can be scheduled for Ethernet framing, and at step 210, the reads and RIT elements can be encoded in binary followed by encoding of reads and RIT elements in hex at step 212. At step 214, the coded read and RIT elements can be transmitted through transmit interface. One should appreciate that these steps are only exemplary in nature and constitute only one of the many embodiments that are possible, and therefore all such embodiments/variations are completely within the scope of the present disclosure.

[62] Referring back to FIG. 1, the genomic data alignment facility can include a number of hardware accelerators such as 112-1, 112-2,..... 112-M (collectively referred to as 112 hereinafter), wherein the hardware accelerators 112 can act as mapping and alignment clusters that can run in parallel with genomic data streamed 110 to them over any streaming protocol/ methodology.

[63] Conventional techniques have tremendous storage requirements of genomic data, reference genomic indices, short reads, hash tables and pointers. In existing hardware acceleration platforms, storage requirements are met in the form of memory within the accelerator (FPGA/ASIC), as rich DDR memories on board and as secondary storage requirements. This results in much of the processing time being spent in retrieving data from storage, and handling memory bottleneck issues. In an embodiment, the accelerator architecture that supports streaming of genomic data from host can drastically reduce the storage requirement within the accelerator platform and thus can overcome the bottleneck created by storage and retrieval of data and make the mapping and aligning process faster and more efficient.

[64] In an embodiment, genomic alignment cluster can provide alignment data at accelerated speeds to a visualization and analysis engine 114 to derive the reports from alignment. In an aspect, the visualization and analysis engine 114 can be part of the host 108.

[65] FIG. 3 illustrates an exemplary block diagram 300 of hardware accelerator cluster in streaming connectivity in accordance with embodiments of the present disclosure. In this embodiment, the hardware accelerator cluster comprising accelerators 112-1, 112-2, 112-3 ...etc. can be spaced apart wherein the genomic data can be streamed to them through streaming infrastructure 302 such as LAN, WAN, cloud etc. to name a few possibilities. Thus, the disclosed accelerator can coexist in a variety of streaming infrastructure.

[66] FIG. 4 illustrates an exemplary block diagram 400 of hardware accelerator architecture with the streaming interface in accordance with embodiments of the present disclosure. Hardware accelerator 112 can incorporate Stream RX 402 that can receive the stream duly embedded with genomic data from the host 108 and can preprocess them to extract short reads and RITs before transferring the extracted data to short read to reference mapper 404 (referred to as mapper 404 hereinafter).

[67] In an embodiment Stream RX 402 can include sub blocks such as stream validation sub-block 410, genomic data extractor sub-block 412, and genomic data scheduler sub-block 414, wherein the stream validation sub-block 410 can receive the pre-processed genomic data through physical interface and validate the frame by checking frame headers to ensure that the frames are indeed targeted to the accelerator. It can discard the frame if the header is invalid and can wait for the next frame. Once a valid frame is detected, the genomic data extractor sub-block 412 can extract the genomic data - short reads and RITs - embedded within the valid frame. The genomic data scheduler sub-block 414 can schedule the extracted data for mapping and alignment across various mapper and aligner kernels within the hardware accelerator 112. In an aspect the Stream RX block 402 can adhere to streaming protocol guidelines and standards.

[68] The hardware accelerator 112 can incorporate short read to reference mapper 404 (referred to as mapper hereinafter) that can receive the genomic data scheduled by Stream RX 402 and can map an incoming short read, against the reference genome, by looking for a hit in the RIT. A single short read can be paired against each of the elements within RIT, to identify a map. The choice of masks, read length and seed length can decide the complexity of the mapper 404. Mapping can be performed for different conditions such as absolute match of the short read and RIT element, single mismatch, two mismatches, three mismatches and so on. There can be a mask set corresponding to each of these conditions.

[69] The hardware accelerator 112 can also incorporate a short read aligner 406 (also referred to as aligner) that can receive the short read and RIT element pair shortlisted by the mapper 404, and work on them to come up with an optimal alignment and an associated alignment score. The aligner 406 can have multiple parallel kernels.

[70] In an alternate embodiment, the hardware accelerator can function with independent, decoupled mapper and aligner models as well, where the mapper alone is present in the pipeline initially, which can perform mapping and filtering of read-RIT pairs,

which are probable candidates for alignment. Thereafter, the filtered read-RIT pairs can be streamed to the accelerator again, which has only aligner present in the datapath, which can perform alignment on the incoming pairs, and produce the scores and associated results for each pair for streaming as output data.

[71] The hardware accelerator 112 can further incorporate a stream TX block 408 that can collect the results of alignment from parallel datapaths of aligner 406 (described in subsequent paragraphs) and prepare them to stream back to the host 108. The Stream TX block 408 can include sub blocks such as an ‘alignment results to stream scheduler sub-block’ 420, a ‘stream TX frame formation sub-block’ 418 and a ‘stream validation sub-block’ 416. The ‘alignment results to stream scheduler sub-block’ 420 can schedule and slot the results to form a payload for the stream as per the standards and guidelines of the streaming protocol chosen. The ‘stream TX frame formation sub-block’ 418 can embed the payload within a transmit frame along with the headers for the frames as per standards. The ‘stream validation sub-block’ 416 can validate the transmit-ready frame with a CRC engine that can calculate the CRC for the entire frame which goes in to the frame as Frame Check Sequence. Validated frames can thereafter be streamed out to the host 108 for assessment.

[72] FIG. 5 illustrates an exemplary block diagram 500 indicating host-hardware accelerator connectivity in accordance with embodiments of the present disclosure. As depicted, connectivity between host 108 and hardware accelerator 112 can be through a duplex wired or wireless Ethernet link 502. The disclosed architecture can support any streaming speed such as 10/100/1000 mbps and the link 502 having adequate speed can be provided to match the hardware capability. It is to be understood that the streaming of the genomic data can be done over any existing protocol, methodology or streaming technologies such as but not limited to built-in hardware interconnects like PCIe, physical Ethernet connectivity with configurable speeds over multiple connectivity models, wireless data transmission models or cloud infrastructure. One should appreciate that these steps are only exemplary in nature and constitute only one of the many embodiments that are possible, and therefore all such embodiments/variations are completely within the scope of the present disclosure.

[73] In an embodiment, there can be multiple data paths provisioned within the hardware accelerator 112 architecture. FIG. 6 illustrates an exemplary block diagram 600 indicating multiple data paths for mapper and aligner within the hardware accelerator architecture in accordance with embodiments of the present disclosure. As illustrated, there

can be multiple mappers such as 404-1,.....404-N in the hardware accelerator 112, each linked to a corresponding aligner 406 such as 406-1,.....206-N. Thus, the disclosed architecture of the proposed hardware accelerator 112 is scalable and can be scaled up to handle any genome mapping problem irrespective of its size.

[74] FIG. 7 illustrates an exemplary flow diagram 700 indicating post processing of streamed results at host 108 in accordance with embodiments of the present disclosure. In an embodiment, the host 108 can receive the streams containing frames embedded with results of mapping and alignment from hardware accelerator 112. The host 108 can process these frames to retrieve these results and store them for analysis and interpretation. Frame capture utilities can give statistics of frame rates and other relevant information regarding speed, bandwidth, throughput etc. and at step 702, the received frames can be validated by examining the header content, whereas at step 704, results of mapping and alignment can be extracted from valid frames, and at step 706, these extracted results can be stored in the host 108 in appropriate format. Subsequently these results can be analyzed and interpreted at step 708.

[75] While the foregoing describes various embodiments of the invention, other and further embodiments of the invention may be devised without departing from the basic scope thereof. The scope of the invention is determined by the claims that follow. The invention is not limited to the described embodiments, versions or examples, which are included to enable a person having ordinary skill in the art to make and use the invention when combined with information and knowledge available to the person having ordinary skill in the art.

ADVANTAGES OF THE INVENTION

[76] The present disclosure overcomes the drawbacks of existing methods of short read mapping and alignment with a reference genome sequence.

[77] The present disclosure provides a hardware accelerator that overcomes the drawbacks of existing methods of short read mapping and alignment with a reference genome sequence.

[78] The present disclosure provides an accelerator architecture that reduces storage requirement while indexing, mapping and alignment, thus overcoming bottleneck in existing methods.

[79] The present disclosure provides an accelerator architecture that supports streaming of genomic data from host, thereby drastically reducing storage requirements within the accelerator platform and speeds up the mapping and alignment of short reads with a reference genome.

[80] The present disclosure provides a hardware accelerator that can co-exist in a variety of streaming infrastructures.

We Claim:

1. A system for mapping and aligning a short read with a reference genome comprising:
 - a streaming module configured to stream said short read and data of said reference genome through a system hardware, wherein said streaming module comprises:
 - a stream receive block to extract said short read and said data of said reference genome that is embedded within said stream, and schedule the extracted data onto one or more datapaths; and
 - a stream transmission block to collect results of alignment from parallel datapaths and prepare them for streaming to a host,
 - a mapping module configured to receive said short read and said data of said reference genome from said stream receive block and map said short read with said data of said reference genome, and
 - an alignment module configured to align a mapped short read with said reference genome and transmit said results of alignment to said stream transmission block.
2. The system of claim 1, wherein said data of said reference genome is obtained by one-time indexing of said reference genome to generate a Reference Index Table before being received at stream receive block.
3. The system of claim 1, wherein said short read and said data of said reference genome is slotted and scheduled within a frame based on a streaming protocol.
4. The system of claim 1, wherein said streaming is done over any existing protocol, methodology, or streaming technology.
5. The system of claim 1, wherein each of said parallel datapaths comprises said mapping module and said alignment module.
6. The system of claim 1, wherein said streaming module is clocked in tandem to said mapper module and said alignment module.
7. A method for mapping and aligning a short read with a reference genome comprising the steps of:

receiving a stream having said short read and data of said reference genome and extracting said short read and said data of said reference genome from said stream for scheduling said extracted data onto one or more datapaths;

mapping said short read with said data of said reference genome;

aligning a mapped short read with said reference genome to transmit results of alignment; and

collecting results of said alignment from parallel datapaths and streaming said results to a host.

8. The method of claim 7, wherein said data of said reference genome is obtained by one-time indexing of said reference genome to generate a Reference Index Table before being received at stream receive block.
9. The method of claim 7, wherein said short read and said data of said reference genome is slotted and scheduled within a frame based on a streaming protocol.
10. The method of claim 7, wherein said streaming is done over any existing protocol, methodology, or streaming technology.

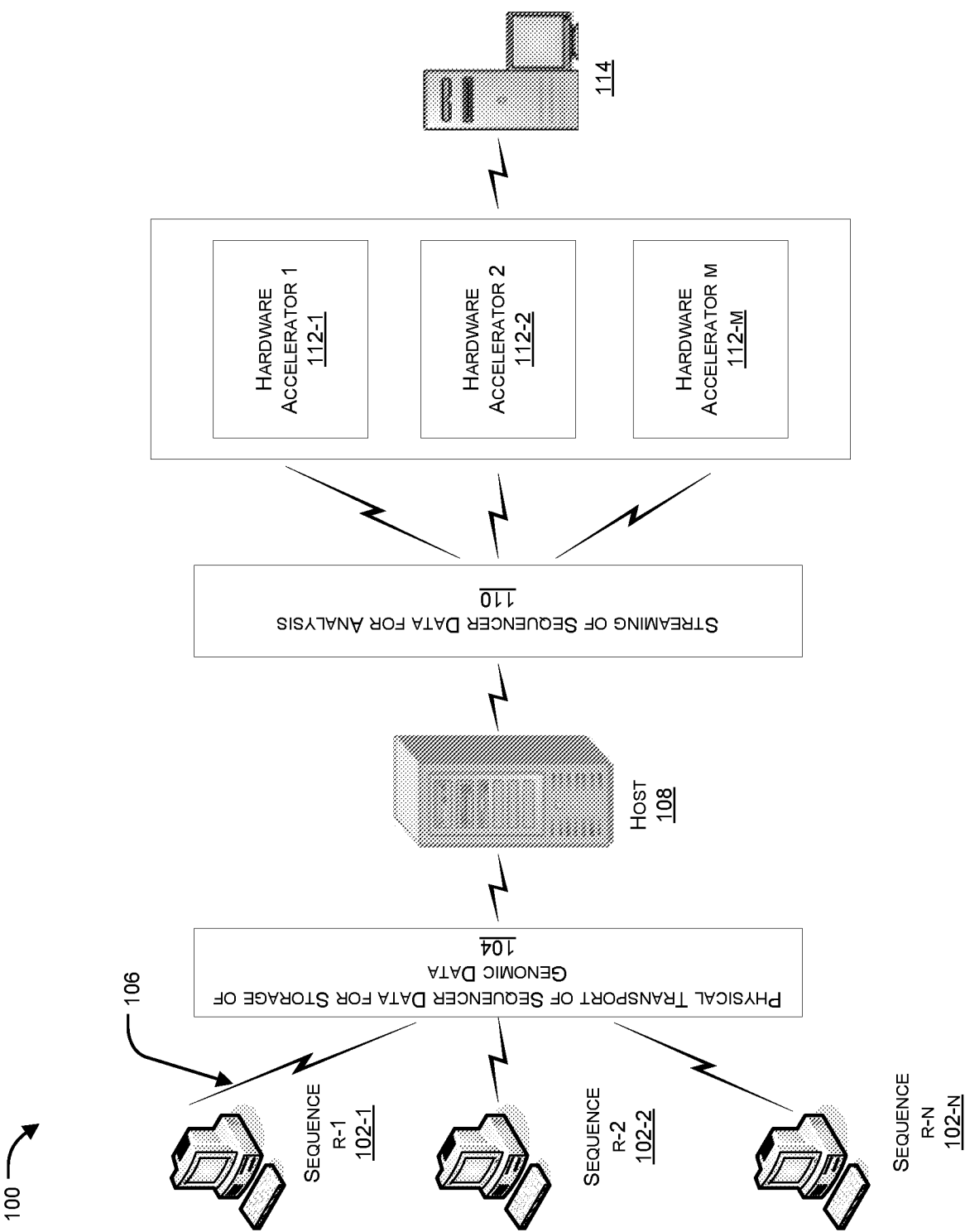


FIG. 1

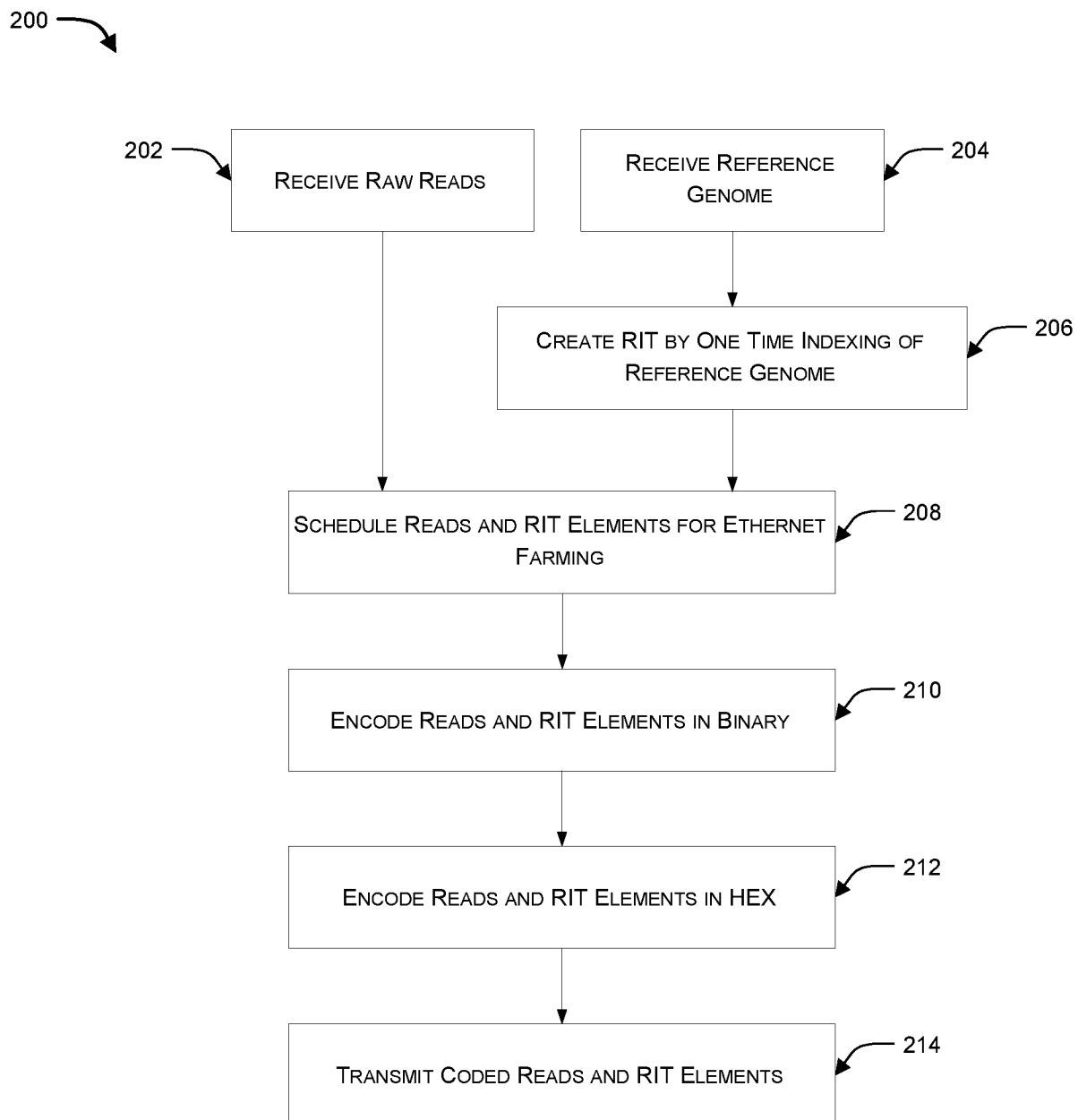


FIG. 2

300 ↗

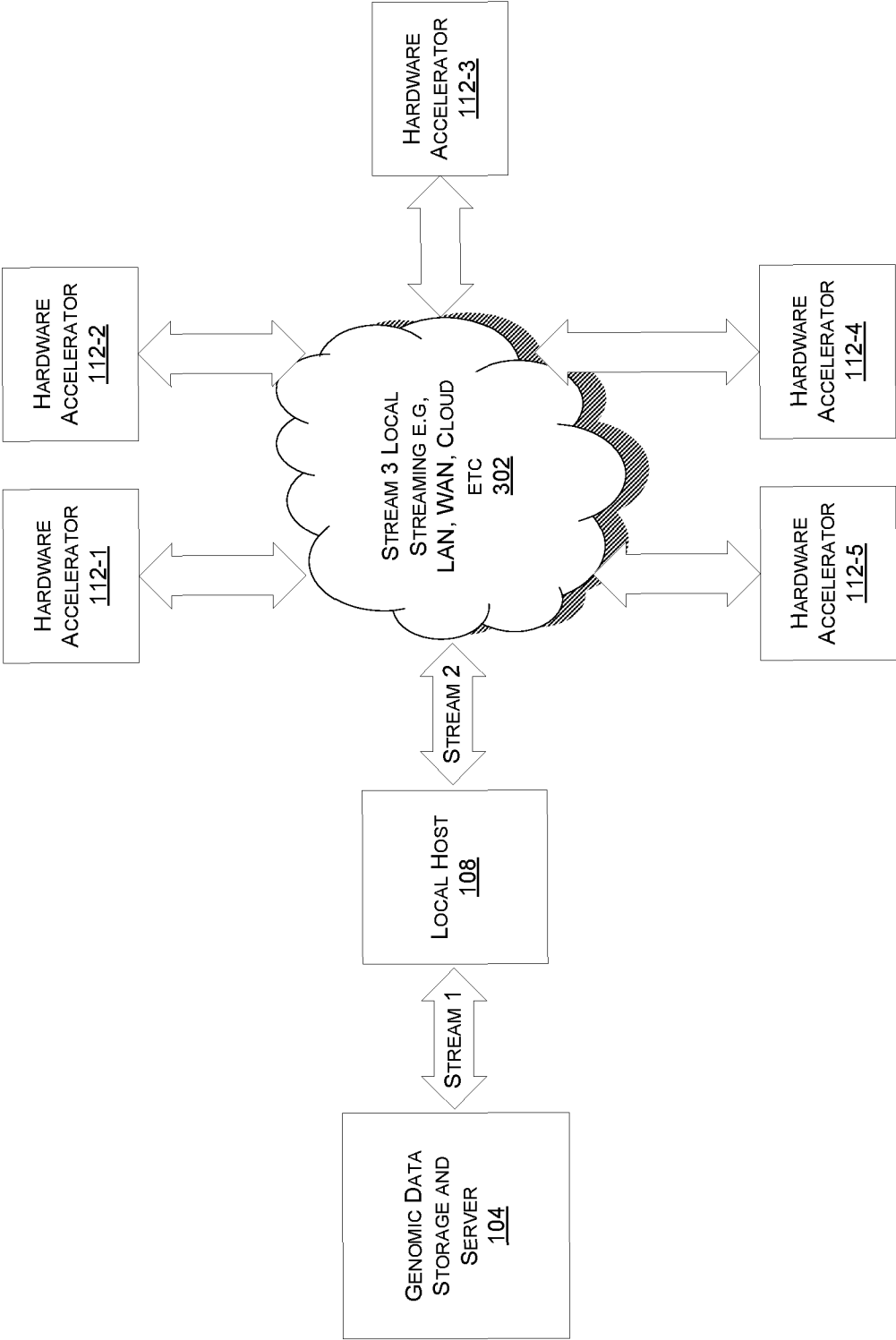


FIG. 3

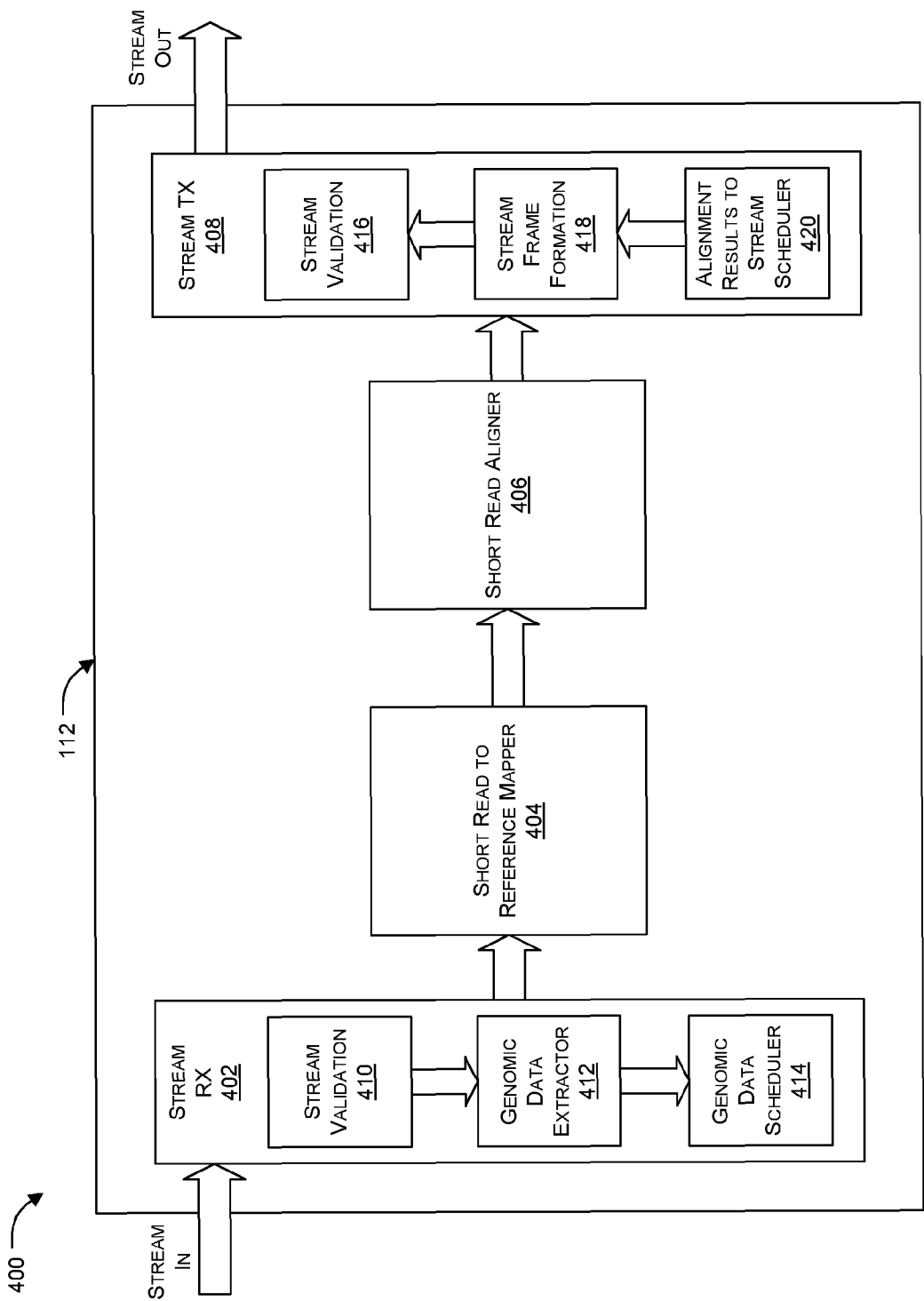


FIG. 4

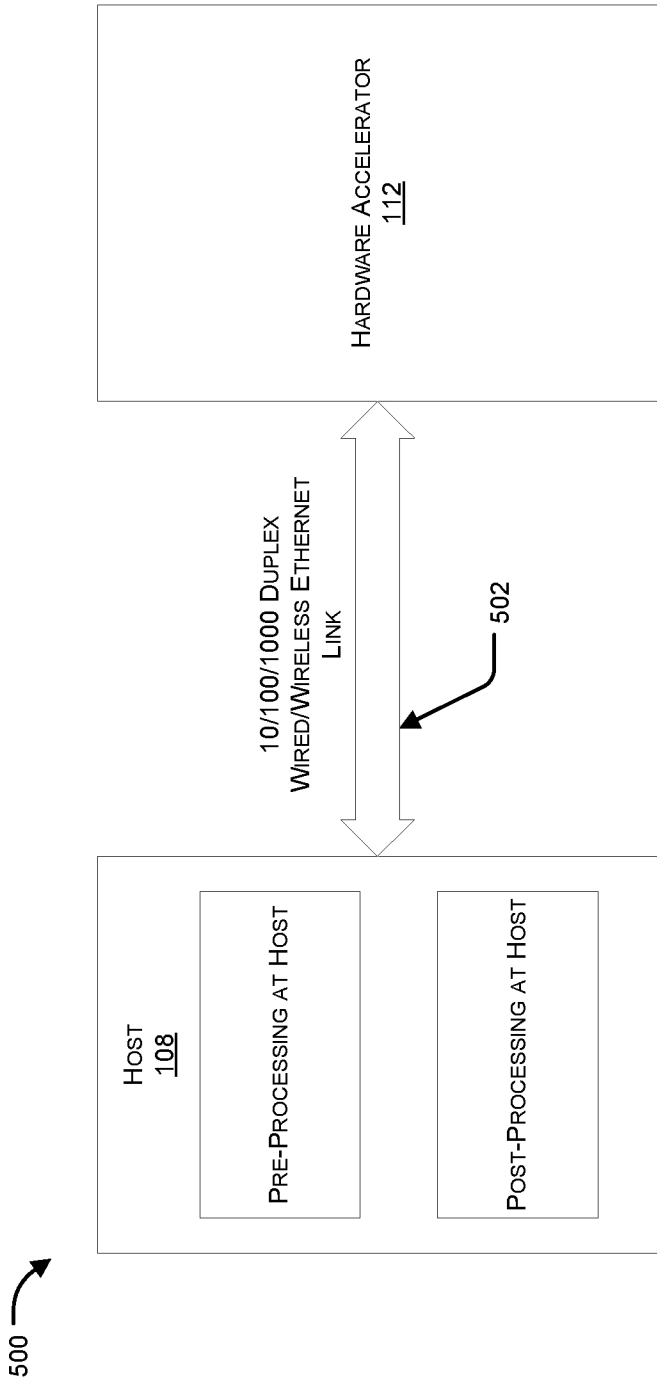


FIG. 5

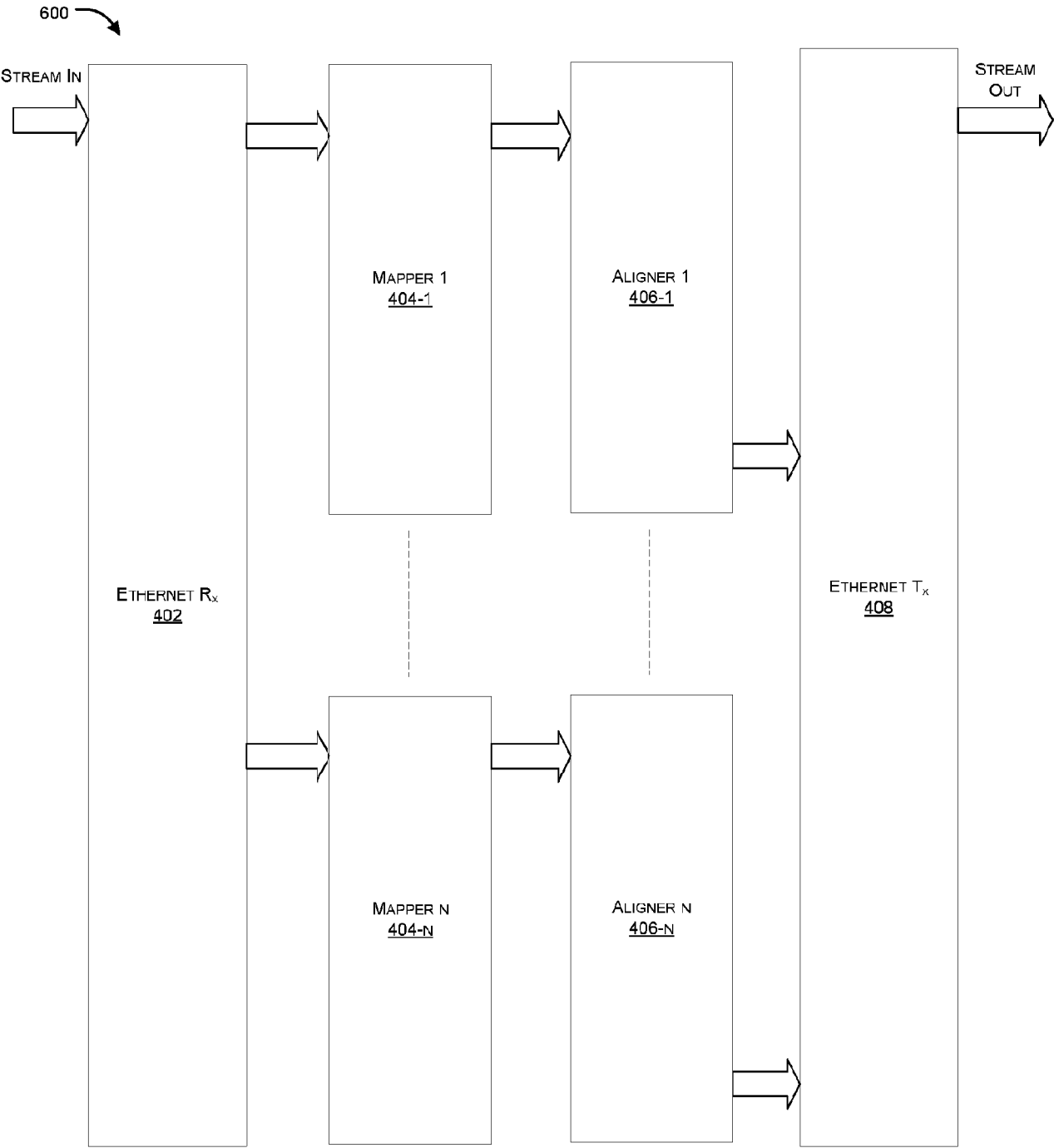


FIG. 6

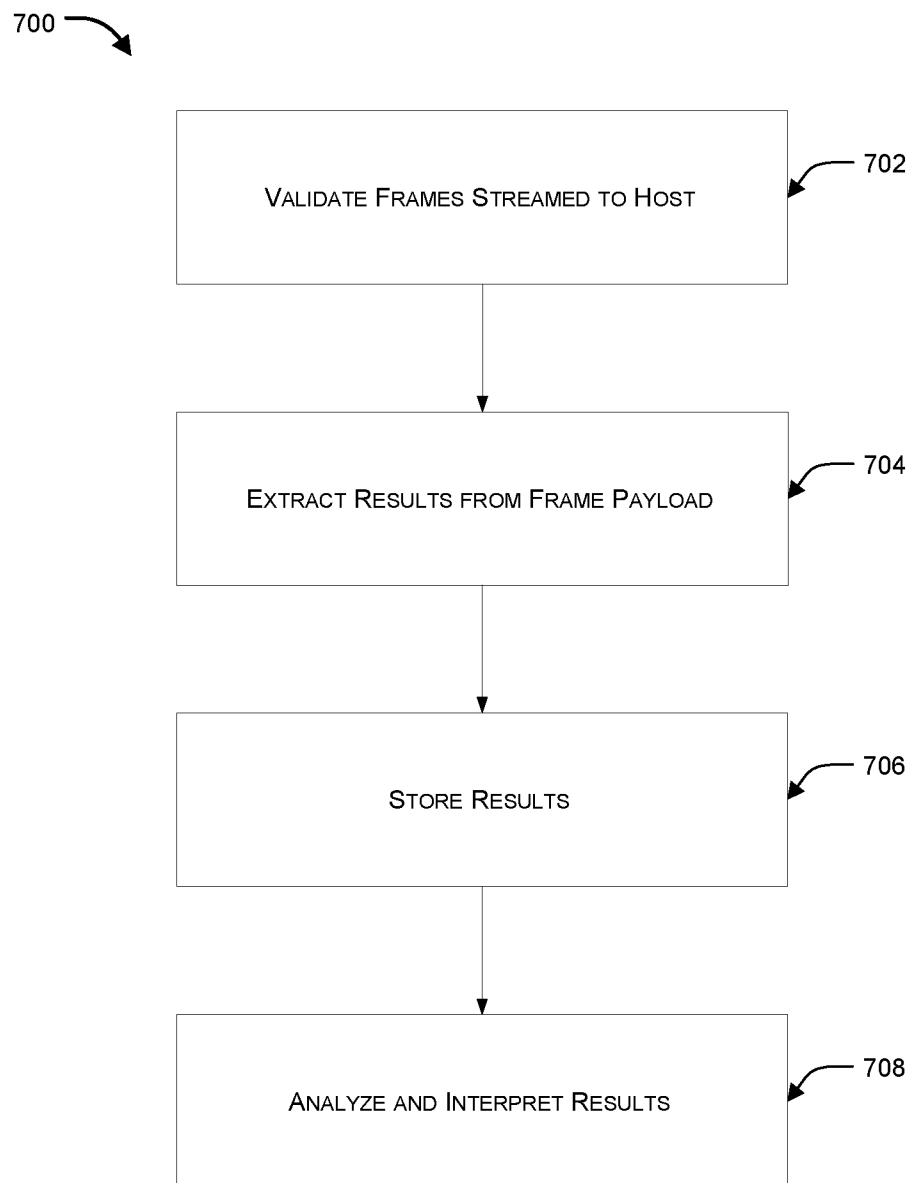


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2016/050841

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/10, G06F19/22 Version=2016.01

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Databases: Patseer, IPO Internal; Search terms: mapping, aligning, short reads, reference genome, hardware accelerator, streaming

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2014/0297196 A1 (PICO COMPUTING, INC.) 2 October 2014 (02-10-2014) Whole document	1-10
X	US 2014/0309944 A1 (EDICO GENOME CORP.) 16 October 2014 (16-10-2014) Paragraphs [0010], [0011], [0020]-[0022], [0026], [0033]-[0035], [0055], [0061], [0064], [0065], [0252], [0253]).	1-10

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 16-05-2016	Date of mailing of the international search report 16-05-2016
Name and mailing address of the ISA/ Indian Patent Office Plot No.32, Sector 14, Dwarka, New Delhi-110075 Facsimile No.	Authorized officer Rakesh Kr Kushwaha Telephone No. +91-1125300200

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2016/050841

Citation	Pub.Date	Family	Pub.Date
US 2014/0309944 A1	16-10-2014	WO 2014113736 A1	24-07-2014