Hardware Accelerators for Genomic Data Processing

William Wang

1 OVERVIEW

Precision medicine offers the promise of personalized medicine for individual patients [23]. However, computational analysis is holding back the progress towards precision medicine, similar to how insufficient computation power held back deep learning. Hardware accelerators have been proposed to accelerate genomic analysis pipelines. The industry *de facto* standard pipeline for secondary data processing of NGS short reads comprises of BWA-MEM for alignment and GATK for variant calling from Broad Institute. Key algorithms in BWA-MEM and GATK can be accelerated, i.e., SMEM for seeding, Smith-Waterman for extending, and PairHMM for variant calling. In addition, nanopore-based sequencing leverages recurrent neural networks for basecalling, RNN accelerators are now commonplace. Accelerators can be implemented in customized computing technologies, such as FPGA and ASIC, some leverage emerging NVM-based analogue computing as well.

2 FPGA ACCELERATORS

Xilinx [15] gives a good overview on how FPGA can help accelerate precision medicine. Luk et al. [32] give a comprehensive review on recent FPGA accelerators for genomics in the past two decades. Ahmed et al. [1] give a good overview comparing seed-and-extend techniques in DNA read alignment algorithms. In addition, Edico Genome [30, 42] and BlueBee [16, 35] provide FPGA based accelerators as part of their high-performance genomics solutions.

2.1 Super-Maximal Exact Match (SMEM)

Cong et al. present an SMEM implementation in FPGA [8, 10]. Luk et al. present an FM-index implementation in FPGA [4]. Wang et al. present an FM-index accelerator design with ASIC implementation in [44]. Edico DRAGEN takes a different hash-based approach for seeding[30].

2.2 Smith-Waterman

Leong et al. present a Smith-Waterman accelerator based on systolic arrays [48]. Liao et al. present a Banded Smith-Waterman implementation [27]. DRAGEN platform from Edico also features a Banded Smith-Waterman implementation. Bluebee and TUDelft present another implementation on Intel FPGA with OpenCL [16].

2.3 PairHMM

IBM Research presents a PairHMM FPGA implementations with systolic arrays in [22]. Other PairHMM FPGA implementations include those from Politecnico di Milano [36], UIUC [6, 17] and TUDelft [35]. Intel offers an open-source PairHMM implementation for selected Intel FPGA cards [21], such as Arria 10 FPGA based cards, i.e., Inspur F10A. The implementation based on OpenCL is described in [20].

2.4 Recurrent Neural Network (RNN)

Han et al. [14], Chang et al. [7], Gao et al. [13], Wang et al. [43], and Li et al. [26] present FPGA accelerators for RNN that is used for nanopore-based sequencing data alignments [37]. NVIDIA offers an open-source deep learning convolutional neural network accelerator *NVDLA* [34].

2.5 Other Algorithms

In addition to BWA-MEM+GATK, certain cancer genomics pipelines also feature alignment refinement before variant calling. The alignment refinement pipeline typically include sorting, duplicate marking, INDEL realignment, and base quality score recalibration. Wu et al. [45] introduce an FPGA accelerated INDEL realignment accelerator in the AWS Cloud that accelerates INDEL realignment by 81x and reduces cost by 32x.

2.6 Accelerator Cards and Cloud FPGA

Xilinx Alveo [47] and Alpha Data [11] accelerator cards provide ready-to-program FPGA on the accelerator cards that can be directly plugged into servers. Maxeler Data Flow Engine (DFE) [29] also offers ready-to-program FPGA modules coupled with 48GB memory, i.e., MPC-X2000 MAX4. Major data centres also offer FPGA instances [46], such as AWS [3], Nimbix Cloud [33], AliCloud [2], Tencent Cloud [39], Huawei Cloud [19] and Baidu Cloud [5].

3 ASIC ACCELERATORS

Turakhia et al. [40, 41] introduce *Darwin*, a genomic processing accelerator that accelerates GraphMap for third generation genomic data processing by 1000x. Fujiki et al. [12] introduce *GenAx*, an genomic processing accelerator that accelerates BWA-MEM by 31.7x with SMEM plus hash-based seeding and automata-based extending stages.

4 EMERGING NVM-BASED ACCELERATORS

Similar to machine learning accelerators from Syntiant [38] and Mythic [31] with NVM-based analogue computation, Technion [24, 25] presents ReRAM-based analogue computing accelerators for seeding and Smith-Waterman. UCSB [18] introduces a 3D-ReRAM based DNA alignment accelerator architecture. In addition, Zokaee et al. [49] present a short-read alignment in ReRAM, and Lou et al. [28] present a spintronics-based basecalling-in-memory architecture for nanopore sequencing.

5 STARTUPS

Edico Genome offers a DRAGEN Bio-IT accelerator card that accelerates the BWA-MEM plus GATK pipeline for secondary data analysis of Illumina-based short reads. BlueBee provides a high-performance cloud-based genomic data analysis service that features FPGA-based acceleration cards. Falcon Computing also provides a compiler-enabled FPGA accelerator solution that targets genomics [9].

REFERENCES

[1] Nauman Ahmed, Koen Bertels, and Zaid Al-Ars. 2016. A comparison of seed-and-extend techniques in modern DNA read alignment algorithms. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 1421–1428.

[2] Alibaba. 2019. Aliyun FPGA Instances. https://cn.aliyun.com/product/ecs/fpga. (2019).

[3] Amazon. 2019. AWS EC2 F1 Instances. https://aws.amazon.com/ec2/instance-types/f1/. (2019).

[4] James Arram, Thomas Kaplan, Wayne Luk, and Peiyong Jiang. 2017. Leveraging FPGAs for accelerating short read alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 14, 3 (2017), 668–677.

[5] Baidu. 2019. FPGA Cloud Compute. https://cloud.baidu.com/product/fpga.html. (2019).

[6] Subho S Banerjee, Mohamed El-Hadedy, Ching Y Tan, Zbigniew T Kalbarczyk, Steve Lumetta, and Ravishankar K Iyer. 2017. On accelerating pair-HMM computations in programmable hardware. In *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 1–8.

[7] Andre Xian Ming Chang and Eugenio Culurciello. 2017. Hardware accelerators for recurrent neural networks on FPGA. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*. IEEE, 1–4.

[8] Mau-Chung Frank Chang, Yu-Ting Chen, Jason Cong, Po-Tsang Huang, Chun-Liang Kuo, and Cody Hao Yu. 2016. The SMEM Seeding Acceleration for DNA Sequence Alignment. In *Field-Programmable Custom Computing Machines (FCCM), 2016 IEEE 24th Annual International Symposium on*. IEEE, 32–39.

[9] Falcon Computing. 2019. Falcon Accelerated Genomics Pipeline. https://www.falconcomputing.com/falcon-accelerated-genomics-pipeline/. (2019).

[10] Jason Cong, Licheng Guo, Po-Tsang Huang, Peng Wei, and Tianhe Yu. 2018. SMEM++: A Pipelined and Time-Multiplexed SMEM Seeding Accelerator for Genome Sequencing. In *2018 28th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 210–2104.

[11] Alpha Data. 2019. Alpha Data High Performance Reconfigurable Computing. https://www.alpha-data.com/dcp/. (2019).

[12] Daichi Fujiki, Aran Subramaniyan, Tianjun Zhang, Yu Zeng, Reetuparna Das, David Blaauw, and Satish Narayanasamy. 2018. GenAx: A genome sequencing accelerator. In *Proceedings of the 45th Annual International Symposium on Computer Architecture*. IEEE Press, 69–82.

[13] Chang Gao, Daniel Neil, Enea Ceolini, Shih-Chii Liu, and Tobi Delbruck. 2018. DeltaRNN: A Power-efficient Recurrent Neural Network Accelerator. In *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 21–30.

[14] Song Han, Junlong Kang, Huizi Mao, Yiming Hu, Xin Li, Yubin Li, Dongliang Xie, Hong Luo, Song Yao, Yu Wang, et al. 2017. Ese: Efficient speech recognition engine with sparse lstm on fpga. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 75–84.

[15] Brian Hill, Jaclyn Smith, Gans Srinivasa, Kemal Sonmez, Ashish Sirasao, Amit Gupta, and Madhubanti Mukherjee. 2017. Precision medicine and FPGA technology: Challenges and opportunities. In *60th IEEE International Midwest Symposium on Circuits and Systems, MWSCAS 2017*. Institute of Electrical and Electronics Engineers Inc.

[16] Ernst Houtgast, Vlad-Mihai Sima, and Zaid Al-Ars. 2017. High Performance Streaming Smith-Waterman Implementation with Implicit Synchronization on Intel FPGA using OpenCL. In *Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on*. IEEE, 492–496.

[17] Sitao Huang, Gowthami Jayashri Manikandan, Anand Ramachandran, Kyle Rupnow, Wen-mei W Hwu, and Deming Chen. 2017. Hardware acceleration of the pair-HMM algorithm for DNA variant calling. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 275–284.

[18] Wenqin Huangfu, Shuangchen Li, Xing Hu, and Yuan Xie. 2018. RADAR: a 3D-reRAM based DNA alignment accelerator architecture. In *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[19] Huawei. 2019. FPGA Accelerated Cloud Server. https://www.huaweicloud.com/product/fcs.html. (2019).

[20] Intel. 2017. Accelerating Genomics Research with OpenCLâĎć and FPGAs. https://www.intel.co.uk/content/www/uk/en/healthcare-it/solutions/documents/genomics-research-with-opencl-and-fpgas-paper.html. (2017).

[21] Intel. 2019. Accelerated kernel library for genomics. https://github.com/Intel-HLS/GKL. (2019).

[22] Megumi Ito and Moriyoshi Ohara. 2016. A power-efficient FPGA accelerator: Systolic array with cache-coherent interface for pair-HMM algorithm. In *Low-Power and High-Speed Chips (COOL CHIPS XIX), 2016 IEEE Symposium in*. IEEE, 1–3.

[23] J Larry Jameson and Dan L Longo. 2015. Precision medicine-personalized, problematic, and promising. *Obstetrical & Gynecological Survey* 70, 10 (2015), 612–614.

[24] Roman Kaplan, Leonid Yavits, and Ran Ginosar. 2018. RASSA: Resistive Accelerator for Approximate Long Read DNA Mapping. *arXiv:1809.01127* (2018).

[25] Roman Kaplan, Leonid Yavits, Ran Ginosar, and Uri Weiser. 2017. A resistive cam processing-in-storage architecture for dna sequence alignment. *IEEE Micro* 37, 4 (2017), 20–28.

[26] Zhe Li, Caiwen Ding, Siyue Wang, Wujie Wen, Youwei Zhuo, Chang Liu, Qinru Qiu, Wenyao Xu, Xue Lin, Xuehai Qian, et al. 2018. E-RNN: Design Optimization for Efficient Recurrent Neural Networks in FPGAs. *arXiv preprint arXiv:1812.07106* (2018).

[27] Yi-Lun Liao, Yu-Cheng Li, Nae-Chyun Chen, and Yi-Chang Lu. 2018. Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 1–9.

[28] Qian Lou and Lei Jiang. 2018. BRAWL: A Spintronics-Based Portable Basecalling-in-Memory Architecture for Nanopore Genome Sequencing. *IEEE Computer Architecture Letters* 17, 2 (2018), 241–244.

[29] Maxeler. 2019. Maxeler MPC-X2000. https://www.maxeler.com/products/mpc-xseries/. (2019).

[30] Neil A Miller, Emily G Farrow, Margaret Gibson, Laurel K Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, Shane Corder, Lisa Krivohlavek, Adam Walter, et al. 2015. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome medicine* 7, 1 (2015), 100.

[31] Mythic. 2019. Mythic's intelligence processing unit (IPU) adds best-in-class intelligence to any device. https://www.mythic-ai.com/. (2019).

[32] Ho-Cheung Ng, Shuanglong Liu, and Wayne Luk. 2017. Reconfigurable acceleration of genetic sequence alignment: A survey of two decades of efforts. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–8.

[33] Nimbix. 2019. Nimbix Cloud FPGA Instances. https://www.nimbix.net/alveo/,https://www.nimbix.net/intel-fpga/. (2019).

[34] NVIDIA. 2019. The NVIDIA Deep Learning Accelerator NVDLA. https://github.com/nvdla/hw. (2019).

[35] Shanshan Ren, Vlad-Mihai Sima, and Zaid Al-Ars. 2015. FPGA acceleration of the pair-HMMs forward algorithm for DNA sequence analysis. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 1465–1470.

[36] Davide Sampietro, Chiara Crippa, Lorenzo Di Tucci, Emanuele Del Sozzo, and Marco D Santambrogio. 2018. FPGA-based PairHMM Forward Algorithm for DNA Variant Calling. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 1–8.

[37] Damla Senol Cali, Jeremie S Kim, Saugata Ghose, Can Alkan, and Onur Mutlu. 2018. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Briefings in bioinformatics* (2018).

[38] Syntiant. 2019. Always-On Machine Learning Solutions for Battery Powered Devices. https://www.syntiant.com/. (2019).

[39] Tencent. 2019. FPGA Cloud Computing. https://cloud.tencent.com/product/fpga. (2019).

[40] Yatish Turakhia, Gill Bejerano, and William J Dally. 2018. Darwin: A Genomics Co-processor Provides up to 15,000 X Acceleration on Long Read Assembly. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 199–213.

[41] Yatish Turakhia, Kevin Jie Zheng, Gill Bejerano, and William J Dally. 2017. Darwin: A Hardware-acceleration Framework for Genomic Sequence Alignment. *bioRxiv* (2017), 092171.

[42] Pieter Van Rooyen, Michael Ruehle, Robert J McMillen, and Mark Hahm. 2016. Bioinformatics Systems, Apparatuses, And Methods Executed On An Integrated Circuit Processing Platform. (June 16 2016). US Patent App. 15/048,935.

[43] Shuo Wang, Zhe Li, Caiwen Ding, Bo Yuan, Qinru Qiu, Yanzhi Wang, and Yun Liang. 2018. C-LSTM: Enabling Efficient LSTM using Structured Compression Techniques on FPGAs. In *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 11–20.

[44] Yuanrong Wang, Xueqi Li, Dawei Zang, Guangming Tan, and Ninghui Sun. 2018. Accelerating FM-index Search for Genomic Data Processing. In *Proceedings of the 47th International Conference on Parallel Processing*. ACM, 65.

[45] Lisa Wu, David Bruns-Smith, Frank A Nothaft, Qijing Huang, Sagar Karandikar, Johnny Le, Andrew Lin, Howard Mao, Brendan Sweeney, Krste Asanovic, et al. [n. d.]. FPGA Accelerated INDEL Realignment in the Cloud. In *Proceedings of the 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*.

[46] Xilinx. 2019. Accelerated Cloud Services. https://www.xilinx.com/products/design-tools/cloud-based-acceleration.html. (2019).

[47] Xilinx. 2019. Xilinx Alveo Adaptable Accelerator Cards for Data Center Workloads. https://www.xilinx.com/products/boards-and-kits/alveo.html. (2019).

[48] Chi Wai Yu, KH Kwong, Kin-Hong Lee, and Philip Heng Wai Leong. 2003. A Smith-Waterman systolic cell. In *International Conference on Field Programmable Logic and Applications*. Springer, 375–384.

[49] Farzaneh Zokaee, Hamid R Zarandi, and Lei Jiang. 2018. AligneR: A Process-in-Memory Architecture for Short Read Alignment in ReRAMs. *IEEE Computer Architecture Letters* 17, 2 (2018), 237–240.