## Architects of the Genome Era Bridging Biology and Compute at Scale

Sneha Goenka Princeton University

## Past 2 decades witnessed significantly reducing cost of sequencing





#### **Genomics growth rate >>> CPU performance**





ATCTGAAATTGTGGCAATAATCAATAGCTTACCAACCAAAAAGAGTCCAGGACCAGATGGATTCACAGCCGAATTCTACCAGAGGTGCAAGGAGGAACTGGTACCATTCCTTCTGAAA CTATTCCAATCAATAGAAAAAGAGGGAATCCTCCCTAACTCATTTTATGAGGCCAGCATGATTCTGATACCAAAGCCTGGCAGAGACACCAAAAAAGAGAAATTTTAGACCAATAT CCTTGATGAACATTGGTGCAAAAATCCTCAATAAAATACTGGCAAAAAAATCCAGCAGCACATCAAAAAGCTTATCCACCATGATCAAGTGGGCTTCAT TCCTGGGATGCAAGGCTG GTTCA ATA TATACA AATCAA TAA ATGCAA TCCAGCATA TAA ACA GAGACA AAGACA AAA ACCACA TGA TTA TCTCAA TAGATGCAGAAA AGGCCTTTGACA AAA TTCAACAACCCTTC ATGCTAAAAACTCTCAATAAATTAGGTATTGGTGGGACGTATCTCAAAATAATAAGAGCTATTTATGACAAACCCACAGCCAATATCATACTGAATGGGCAAAAACTGGAAGTGTTCC CTTTGAAGACTGGCACAAGACAGGGATGCCCTCTCTCACCACTCCTATTCAACATAGTGTTGGAAGTTCTGGCCAGGGCAATTAGGCAGGAGAAGGAAATAAAGGGTATTCAATTAGG AAAGAGGAAGTCAAATT 'TGCAGATGACACGATTGTATATCTAGCAAACCCCCATTGTCTCAGCCCAAAATCTCCTTAAGCTGATAAGCACAACTTCAGCAAAAGTCTCAGGA TA CAA AAT CAA TGTACAAAA ATCACAAGCATT CCTATA CAC CAA CAA CAGACAAACAGAGAGCCAAAT CAT GAGTGAACT CCCATT CACAAT TTCAAAGAGAATAAAATACCTAG GAAT AGGAAG AAGGAGAAC TACAAACCACTGCTCAATGAAA 'AAAAGAGGAT AATCAA TGGAACCAAAAAAGAGCCCGCATCACCAAGTCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCATCACACTACCTGACTTCAAACTATACTACAAGGCTACAGTAACCAAAACAAGCAT GGTACTGGTACCAAAACAGAGATATAGATCAATGGAACAGAACAGAGCCCTCAGAAATAATGCCGCCATATCTACAACTATCTGATCTTTGACAAACCTGGGAAAAACAAGCAATGGGG <u>ETTATA CAAAAA TCAATTCAA GATGGA TTAAAGAC</u> AAAGGATTCCCTATTTAATAAATGATGC TTAAACGTTAGACCTAAAACCATAAAAA GI TTGCAACCTACTCATC CCAAAATTGACAAATGGGATCTAATTAAACTAAA ACAGGCAACCTACAAAATGGGAGAAAAT CCATCAAAAAGTGGGCAAAGGACACAAACAGACACTTCTCAAAAGAAGAC TGACAAAGGGCTAATATCCAGAATCTACAATGAACT AGAAAAAAACAACAACC ATTTATGCAGCCAAAAAACACATGAAAAAATGCTCA CAGAGAAATGCAAA TCAAAACCACAATGAGATACCATCTCACACCAGTTAGAATGGCAATCATTAAAA ••• **...** (m) A AGTCAGGAAACAACAGGTGCTGGAGAGGATGTGGAG G 'GTGGAAGTCAGTGTGGCGATTCCTCAGGGA TCTAGAACTAGAAATACC ACAATAGCACAGACTTGGAACCAA CGCAA ATGTCCAACAATGATAAA CTGGATTAA GAA AATGTGGCA CATATA CACCATGGA ATA CTATGCA GCCATAAAAAATGATGAGTTCATGT A GTAAACTATCCCAACAACAAAAAACCAAACAAATGAGATCACA TGGACACAGGAAGGGGAACA TAGTGGGTGCAGTGCAC GCAACAACAACAAAAAAAAATTTAAAACATGAGCAAAGGTTTTAAAATACACATTTCTCTAAAGAAGATATGCCAAGTAAGCACAGGATAAGGTGCTCAGCATCACTAAT TGCTGGATGTGGAGAAGTCAGGACTCCTGCACACTGCTGGTGGGAAAGTAAGATGGCACAGCCACTGTGGAAAACAGTAATCACCATATGATCCAG' LTGGGTATA TA CCCCAAAAAACTGAAAGTGGGAATTTGTACACCCATGTTTATAGCAGCATTCACAAGAGCCAAAGGGTAGAAAAAGCCCAAATCTCCATCTACAGATGAATGGATAAGCAAATATGAT GCCATATGTAGAAAGCTGAAACTGGATCCCTTCCTTACACCTTATACAAAAATCAAGTCAAGATGGATTAAAGACTTAGACCTTAGACCTTAAAACCCTTAGAACCCTAGAAGAAAACCC TAGGCAATACCATTCAGGACATAGGCATGGGCAAGGACTTCATGTCTAAAACACCAAAAGCAATGGCAACAAAAGCCAAAATTGACAAATGGGAT BRANGEFON CACAGCAAAAGAAACTACCATCAGAGTGAACAGGCAACCTACAAAATGGGAGAAAATTTTTTGCAACCTACTCATCTGACAAAGGGCCTAATATCCAGAA NIVERSITY

### **Comparative genomics enables genome interpretation**

Article Published: 25 February 2021

A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity

Chimp Genome Helps Scientists Learn More About Human DNA

News Published: September 8, 2005



**Prediction of functional elements** 



#### **Thousand-genome era is already here**



4 million babies born in the US annually

12% babies admitted to the NICU

1/3 NICU hospitalizations have a genetic cause

#### 40%

longer hospitalizations when disorders are genetic

ICU hospitalizations cost \$15-20k/day



#### Faster genetic diagnosis has significant impact



**Genetic Diagnosis Turn-around Time** 



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



#### **Clinical Genomics**

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



#### **Fragmented Acceleration in Genomics**

#### SquiggleFilter: An Accelerator for Portable Virus Detection

Tim Dunn\* timdunn@umich.edu University of Michigan Ann Arbor, MI, USA Harisankar Sadasivan\* hariss@umich.edu University of Michigan Ann Arbor, MI, USA

Jack Wadden jackwadden@gmail.com University of Michigan Ann Arbor, MI, USA

Kush Goliya kgoliya@umich.edu University of Michigan Ann Arbor, MI, USA

#### CiMBA: Accelerating Genome Sequencing Through On-Device Basecalling via Compute-in-Memory

William Andrew Simon<sup>(0)</sup>, Member, IEEE, Irem Boybat<sup>(0)</sup>, Member, IEEE, Riselda Kodra<sup>(0)</sup>, Student Member, IEEE, Elena Ferro<sup>(0)</sup>, Student Member, IEEE, Gagandeep Singh<sup>(0)</sup>, Mohammed Alser<sup>(0)</sup>, Member, IEEE, Shubham Jain<sup>(0)</sup>, Member, IEEE, Hsinyu Tsai<sup>(0)</sup>, Senior Member, IEEE, Geoffrey W. Burr<sup>(0)</sup>, Fellow, IEEE, Onur Mutlu<sup>(0)</sup>, Fellow, IEEE, and Abu Sebastian<sup>(0)</sup>, Fellow, IEEE

Home / Magazines / IEEE Micro / 2021.04

#### IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

Jul.-Aug. 2021, pp. 39-48, vol. 41 DOI Bookmark: 10.1109/MM.2021.3088396

#### SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

Damla Senol Cali<sup>1</sup> Konstantinos Kanellopoulos<sup>2</sup> Joël Lindegger<sup>2</sup> Zülal Bingöl<sup>3</sup> Gurpreet S. Kalsi<sup>4</sup> Ziyi Zuo<sup>5</sup> Can Firtina<sup>2</sup> Meryem Banu Cavlak<sup>2</sup> Jeremie Kim<sup>2</sup> Nika Mansouri Ghiasi<sup>2</sup> Gagandeep Singh<sup>2</sup> Juan Gómez-Luna<sup>2</sup> Nour Almadhoun Alserr<sup>2</sup> Mohammed Alser<sup>2</sup> Sreenivas Subramoney<sup>4</sup> Can Alkan<sup>3</sup> Saugata Ghose<sup>6</sup> Onur Mutlu<sup>2</sup>



#### **Fragmented Acceleration in Genomics**

SquiggleFilter: An Accelerator for Portable Virus Detection

CiMBA: Accelerating Genome Sequencing Through

#### Most accelerators lack software frameworks — and when they exist, they're built in isolation with no shared ecosystem

#### Intensive Applications

Jul.-Aug. 2021, pp. 39-48, vol. 41 DOI Bookmark: 10.1109/MM.2021.3088396 Damla Senol Cali<sup>1</sup> Konstantinos Kanellopoulos<sup>2</sup> Joël Lindegger<sup>2</sup> Zülal Bingöl<sup>3</sup> Gurpreet S. Kalsi<sup>4</sup> Ziyi Zuo<sup>5</sup> Can Firtina<sup>2</sup> Meryem Banu Cavlak<sup>2</sup> Jeremie Kim<sup>2</sup> Nika Mansouri Ghiasi<sup>2</sup> Gagandeep Singh<sup>2</sup> Juan Gómez-Luna<sup>2</sup> Nour Almadhoun Alserr<sup>2</sup> Mohammed Alser<sup>2</sup> Sreenivas Subramoney<sup>4</sup> Can Alkan<sup>3</sup> Saugata Ghose<sup>6</sup> Onur Mutlu<sup>2</sup>



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



#### **Clinical Genomics**

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



### **Understanding whole genome alignment**



13 **PRINCETON** UNIVERSITY

## Seeding finds small, local matching base-pairs



## Seeding finds small, local matching base-pairs



#### Filtering aligns ~100bp around seed hits





### **High-scoring Segment Pair reduced to Anchor**





### **Extension results in the final alignments**

#### **Dynamic Programming Equations**

$$I(i,j) = \max \{H(i,j-1) - o, I(i,j-1) - e\}$$
  

$$D(i,j) = \max \{H(i-1,j) - o, D(i-1,j) - e\}$$
  

$$H(i,j) = \max \begin{cases} 0\\I(i,j)\\D(i,j)\\H(i-1,j-1) + W(r_i,q_j) \end{cases}$$

#### Alignment

human mouse	1 <mark>AGG</mark> T <mark>AG</mark> CAA <mark>GGGGGACAGGA</mark> G
human	26 AGGAGGGGACAGGAG - TG <mark>GCC</mark> AGGAGTGGCCAGGA
mouse	36 AGGAGGGGGCAGGAAACAGCCTGCAGGGGT - AGGA
human	60 GGGGGCAGG
mouse	70 GGGGGCAGG





### Filtering stage dominates the runtime



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



**Clinical Genomics** 

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



## SegAlign system for single chromosome pair





## Naïve approach allocates 1 seed hit/thread

 Considerably varying seed hit positions -> inefficient uncoalesced memory accesses within a warp

[Warp - basic unit for scheduling execution and memory accesses]



2. Divergent branches within a warp due to the dynamic X-drop condition for each thread



#### SegAlign allocates 1 seed hit/warp



1. Efficient bandwidth gains with coalesced memory accesses

2. Exploiting data locality within each partition using parallel prefix scan



#### 13.5x-14x speedup at ~2x cost improvement





# SegAlign's Ungapped extension kernel now in NVIDIA GenomeWorks library

https://github.com/clara-parabricks/GenomeWorks

#### GenomeWorks



#### **Overview**

GenomeWorks is a GPU-accelerated library for biological sequence analysis. This section provides a brief overview of the different components of GenomeWorks. For more detailed API documentation please refer to the documentation.



## Seed-Ungapped Filter-Extend isn't sensitive enough



Improved search heuristics find 20 000 new alignments between human and mouse genomes a

Martin C. Frith 🖾, Laurent Noé

Nucleic Acids Research, Volume 42, Issue 7, 1 April 2014, Page e59, https://doi.org/10.1093/nar/gku104 Published: 31 January 2014 Article history ▼ Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation 3

Virag Sharma, Michael Hiller 💌

Nucleic Acids Research, Volume 45, Issue 14, 21 August 2017, Pages 8369–8377, https://doi.org/10.1093/nar/gkx554 Published: 21 June 2017 Article history ▼



# Increasing indel frequency => increasing need for gapped filtering





## Seed-Gapped Filter-Extend



Replacing ungapped filtering by gapped filtering slows down the software by 200x!



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



<u>Clinical Genomics</u>

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



#### **Specialized Operations**



**On 14nm CPU** 35 ALU ops, 15 load/store 37 cycles 81nJ On 40nm Special Unit 1 cycle (*37x speedup*) 3.1pJ (*26,000x efficiency*) 300fJ for logic (*remainder is memory*)



30

## **Exploiting inner-loop parallelism with systolic arrays**







### **Can we adapt this architecture for extension?**





# Utilizing local memory – size is prohibitive for larger compute





### **Overlapped extension uses constant on-chip memory**

- Tiled (tile size T, overlap O) implementation inspired by GACT in Darwin
- Origin of the next tile lies at the intersection of the current traceback path with the overlap



Outside tile overlap
 Inside tile overlap
 On tile overlap border



## **Overlapped extension uses constant on-chip memory**

• Extension along a direction continues until a tile is encountered with a nonpositive maximum score





#### **Banded + extension arrays = Outer-loop parallelism**



#### **Extension** arrays





## Darwin-WGA finds genes that LASTZ does not



dp4 1 GTGAAGCCTGGTGCTGCATC	20
	12
	12
dm6 57 cccgttcccgttcccgttccctttcccgttTCCATTTGCATTTCCATTATCCCCGA dp4 21 TCCAccatttccattgccatctccatttccatttccatgtccgtttccGTTCA	73
Seed hit 1	
dm6 113 CCCTCAGCGATATAGATTTGAACAACTTGTGCATCGATTTGGGTCG	58
dp4 74 CAATCAAAGATATGGACTTGAATAACTTTGGCATCGACGTGGAGCGCCTGTGGCTG	29
Seed hit 2	
dm6 159 GGAAIGIGCGGGAGCCGAGCIGCGIIICAAIIICAGCGAGIIAIAGIIIGGCIC	12
dp4 130 CGCATGTGGG1GGGCGCCGAGCTGCGTTTCACTGAATCGAAGGGCAATCGGaact	85
dm6 213 TGGATGAGGATTCGAA	41
dp4 186 tgaactcaaattcaaattcaaattcaaattcaaattagcgtccgtc	41
dm6 242 TATCGCA	63
dp4 242 TATCGCATTCGTCCTCCACGGCGTCGACGGCAGCAGCGGCGGCAGGGGGCGGCGGT2	.94
Seed hit 3	
Seed Int S	
dm6 264 TGGCACCGCGCTAGCACTTTTGTAGTGCAAACCGTTTTCGGCCATCTTATCTAGGC	19
dp4 295 - GGTATAGCG GCATTTTTAAAATGAAAACGTTTTCGGCTGGCT ATC GGT	44
dm6 320 GGCTCCTATGGCCACAGTCACtgttattgttgttgttgttgttgttgttgCACATGGCCAGA	75
dp4 345 GGTGCCGTTGCTATTGTTGTTGTTGCACATT-CCAGA	80

Indels (shown by arrows) around each seed hit – dropped by ungapped filtering (*LASTZ*) but retained by gapped filtering (*Darwin-WGA*)



## **Darwin-WGA is more sensitive than LASTZ**

Species pair	Top-10 Alignment Chain Scores	Matching Base-pairs within Alignments	Number of Aligning Exons (protein-coding genes)	
dm6-droSim1	16-droSim1 +0.03% 1.		+0.20%	
dm6-droYak2	dm6-droYak2 +0.05% 1.41x		+0.09%	
dm6-dp4	+1.86%	1.42x	+0.41%	
ce11-cb4	+5.73%	3.12x	+2.70%	
	Represent <u>orthologous</u> <u>sequences (</u> derived from "speciation")	Represent <u>paralagous</u> <u>sequences</u> (derived from "duplication")	Represent <u>functionally</u> <u>relevant orthologous</u> <u>sequences</u> , under some selective pressure (at	

False positive rate (2-mer shuffled genome): 0.0007%



least in the target species)

## Darwin-WGA (FPGA) is 20x faster than iso-sensitive software

	LASTZ (CPU)	Iso-sensitive software		HW config	Cost per hour
Darwin–WGA (FPGA)	0.1x (slowdown)	20x (speed and cost)	Darwin- WGA (FPGA)	1 Xilinx Virtex Ultrascale+ FPGA + 8vCPUs	\$1.65



## While Darwin-WGA uses the same software framework, it features a different interface





#### How about a new language?

Bioinformaticians focus on filters and recurrences.

We provide a domain-specific language to manage load balancing and scheduling.

Architects can then optimize these stages for performance.



## **Introducing FILTR**



Pipeline: sequence of producer-consumer relationships as self-contained, reusable stages

HSP: encapsulates intermediate hits/points along with metadata such as scores, positions, and alignment status

Dataflow control: built-in balancing policies or define custom strategies



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



#### **Clinical Genomics**

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



- 13 years
- Two-week history of dry cough, decreased appetite, chest pain and severe fatigue
- On admission to hospital in Oregon: weak heart
- Rapid deterioration
- Airlifted to Stanford hospitals
- Placed on ECMO
- Possible causes included potentially reversible (myocarditis) and irreversible (genetic heart disease)
- How to make transplant decision?
  - Biopsy?
  - Genetic testing?





#### **State-of-the-art turn-around times**





#### **DNA Sequencing**



**Base calling** 





Alignment

Human Reference Genome



## **Existing fastest turn-around in 14.5 hours**





Illumina HiSeqX



47 PRINCETON UNIVERSITY

#### **Compute dominates the new pipeline with 30hour turnaround time**





**Nanopore PromethION** 

- Longer reads => more complicated computation
- Higher error rate => double the amount of sequencing
- No custom ASIC/FPGA



## **Enabling biologists to leverage sequencing**

#### **Comparative Genomics**



SegAlign: A Scalable GPU-based whole genome aligner [Supercomputing Conference]



**Clinical Genomics** 

An ultra-rapid workflow for clinical whole genome sequencing [New England Journal of Medicine Nature Biotechnology]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments [HPCA]



## **Traditional computational pipeline**





#### Nanopore's "real-time" advantage

- Signal files are generated as soon as the strand passes through the nanopore
- Ideally, we can start base calling right away





## Modified pipeline – overlap base calling and sequencing





#### Challenge1: transfer TB data to cloud

- 2.3 TB of data in 1.5 hour = 3.4 Gbps
- Utilizing available bandwidth



#### Challenge1: transfer TB data to cloud

- 1. VBZ compression for raw signal file 30% less file size
- 2. Optimize file size for
  - (a) number of parallel uploads
  - (b) latency overhead for each new file





**Cloud Storage** 

**Signal Files** 



#### Near real time I/O

### Challenge 2: Optimized distributed system

- Support streaming dataflow
- Minimize orchestration/inter-node communication
- Make sure all resources are fully utilized based on rate of data generation



#### Near real time I/O

#### Challenge 2: Optimized distributed system

- 1. Analysis for specific set of flow cells assigned to each instance
- 2. Stateless pull architecture
- 3. Pipelining different compute stages



## Near real time pipeline





### **Variant Calling**







#### **10.5-hour compute-optimized workflow**





#### Diagnosis made in

## 11.3 hrs

Genetic Dilated Cardiomyopathy

The patient was urgently listed for transplant and received a new heart 21 days later.







### **Co-design across stages**



Negligible impact of sample identification step





#### **Co-design across stages**

Close collaboration with genetic counselors reduces the time for curation with more accurate variant calling

Negligible impact of sample identification step



Non Barcoded



## Ultra-rapid pipeline with 8 hour turnaround time



#### **Diagnosis in 8 hours.. Or less**

![](_page_63_Figure_1.jpeg)

![](_page_63_Picture_2.jpeg)

#### Conclusions

#### Performance in domain-specific design significantly depends on

- Co-Design through integrated systems architecture and algorithmic re-design
- Deep understanding of the domain

#### Landscape for genomics as a domain

- Technological Agility Must keep pace with rapid tech and algorithmic evolution
- Hardware Ramifications GPUs/FPGAs are better platforms; Software framework are critical
- Community Dynamics Community engagement drives adoption of innovations

![](_page_64_Picture_8.jpeg)

## Thank you!

## Workloads in LASTZ v/s Darwin-WGA

![](_page_66_Figure_1.jpeg)

![](_page_66_Picture_2.jpeg)

TSMC 40nm DC synthesis (not a chip prototype)

		Configuration	Area (mm <sup>2</sup> )	Power (W)	
Filtering	Logic	64 x (64PE array)	16.6	25.6	~60% chip power
Extension	Logic	12 x (64PE array)	4.2	6.72	
	Traceback SRAM	12 x (64PE x 16KB/PE)	15.1	7.92	~40% chip area
DRAM	DDR4-2400R	4 x 32GB	-	3.10	
TOTAL			35.9	43.34	

![](_page_67_Picture_3.jpeg)

## Darwin-WGA is 2 orders of magnitude faster than iso-sensitive software

	LASTZ (CPU)	Iso-sensitive software		HW config	Cost per hour
			LASTZ	36 vCPUs	\$1.59
Darwin–WGA (FPGA)	0.1 x (slowdown)	20x (speed and cost)	Darwin- WGA (FPGA)	1 Xilinx Virtex Ultrascale+ FPGA + 8vCPUs	\$1.65
Darwin–WGA (ASIC)	1.5x	300x (1500x perf/Watt)	Darwin- WGA (ASIC)	36 mm <sup>2</sup> , 43 Watt, 40nm TSMC	

![](_page_68_Picture_2.jpeg)