CLOUD-BASED HPC PLATFORM TO ACCELERATE GENOME ANALYSIS

KURT FLORUS, CTO BLUEBEE

6/4/2016

BLUEBEE
High Performance Genomics

INTRODUCING BLUEBEE

BLUEBEE
High Performance Genomics

About Bluebee

WHAT ?

- Spinoff of the Delft University of Technology and Imperial College London
- Active at the crossroads of High Performance Computing and Genomics
- HPC based cloud platform to enable fast, secure and affordable genome analytics

WHY ?

- Reduce cost and complexity of genetic analysis
- Provide the industry with a global, scalable and sustainable solution

HOW ?

- HPC + Cloud + Genomics = High Performance Genomics
- Combine ease of use with clinical grade data integrity, security and privacy

The people behind Bluebee

Prof. Dr. Koen Bertels
Founder

Chair and Professor at the Computer Engineering Lab at Delft University of Technology

Prof. Dr. Edwin Cuppen
Scientific Advisor

Professor of Human Genetics and Head of Research at the University of Utrecht

Universitair Medisch Centrum Utrecht

Hans Cobben
Co-founder and CEO

Seasoned entrepreneur, senior executive, and co-founder of a number of globally successful companies. Strong background in IT and deploying large scale mission-critical platforms.

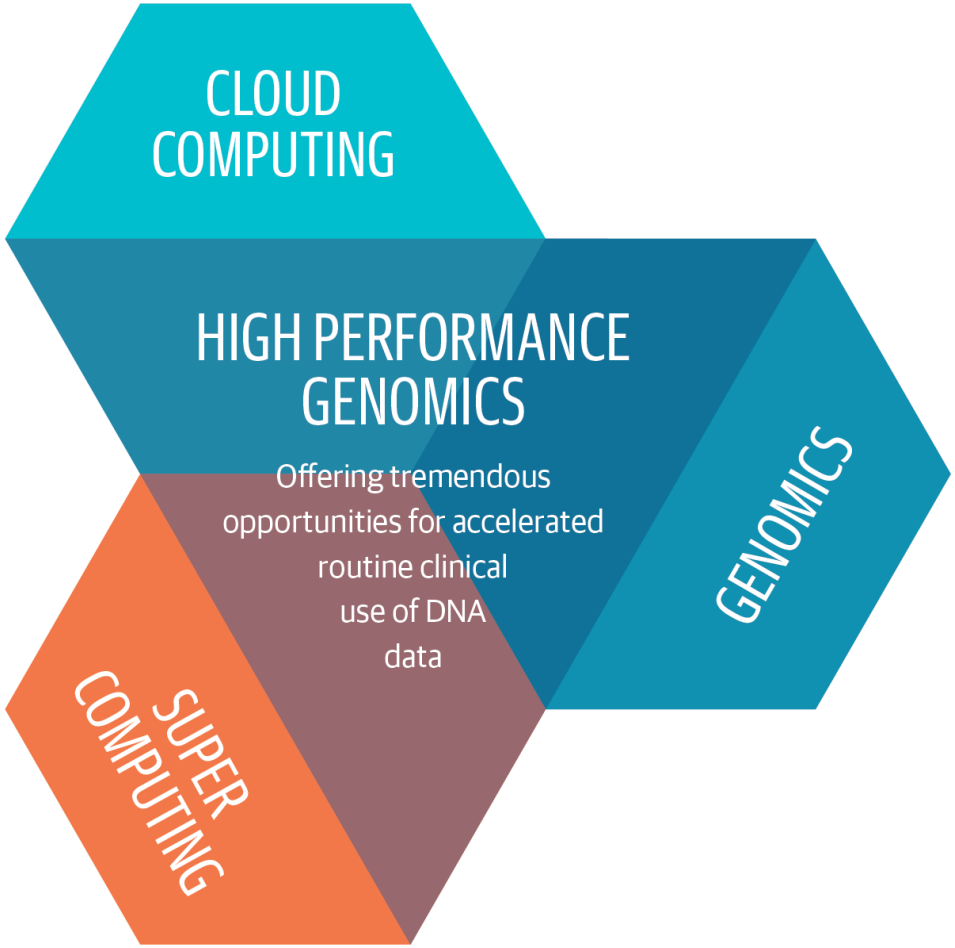
Dr. Peter Hofstee
Scientific Advisor

Senior researcher at IBM Austin Research Laboratory on workload optimized and hybrid systems

Based in Delft, the Netherlands

BLUEBEE
High Performance Genomics

Bluebee Genomics Platform



What ?

- Accelerated private cloud platform for genome data analysis
- A highly secured platform that
  - offers fast & affordable processing
  - provides full control for configuration
  - allows for managed sharing and collaboration
  - supports global roll-out and guarantees local data residency
- Scientific validation and transparency is the foundation of our offering
  - Gold standard algorithms
- A network of global partners that share expertise through the platform

How ?

Bluebee addresses the genome analysis challenge by providing an HPC based private cloud solution for accelerated processing of mass volumes of NGS data

High Performance

- ✓ Leverage our high performance computer infrastructure to drastically reduce processing time
- ✓ Use gold standard algorithms without compromise

Available & Secure

- ✓ Unlimited up scaling and on-the-fly provisioning of computational and data storage capacity
- ✓ Top notch security and data integrity

Convenient

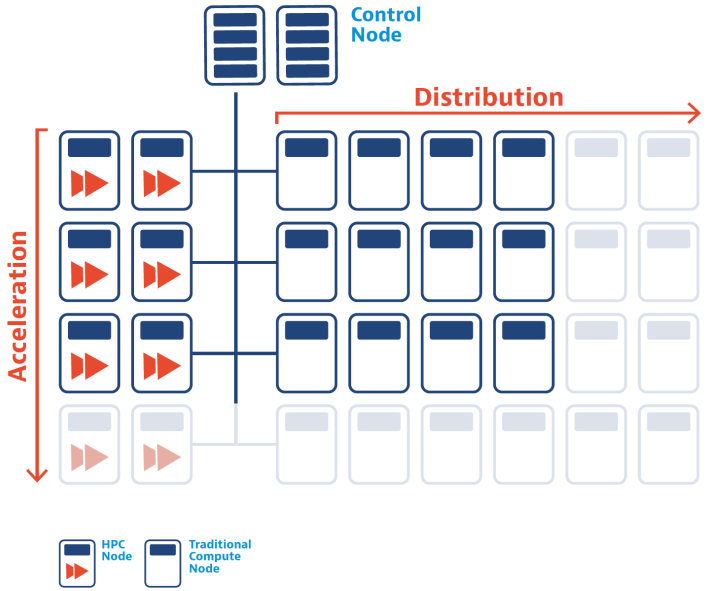
- ✓ Convenience for use, sharing and collaboration
- ✓ Flexible configuration and permission based access

Cost effective

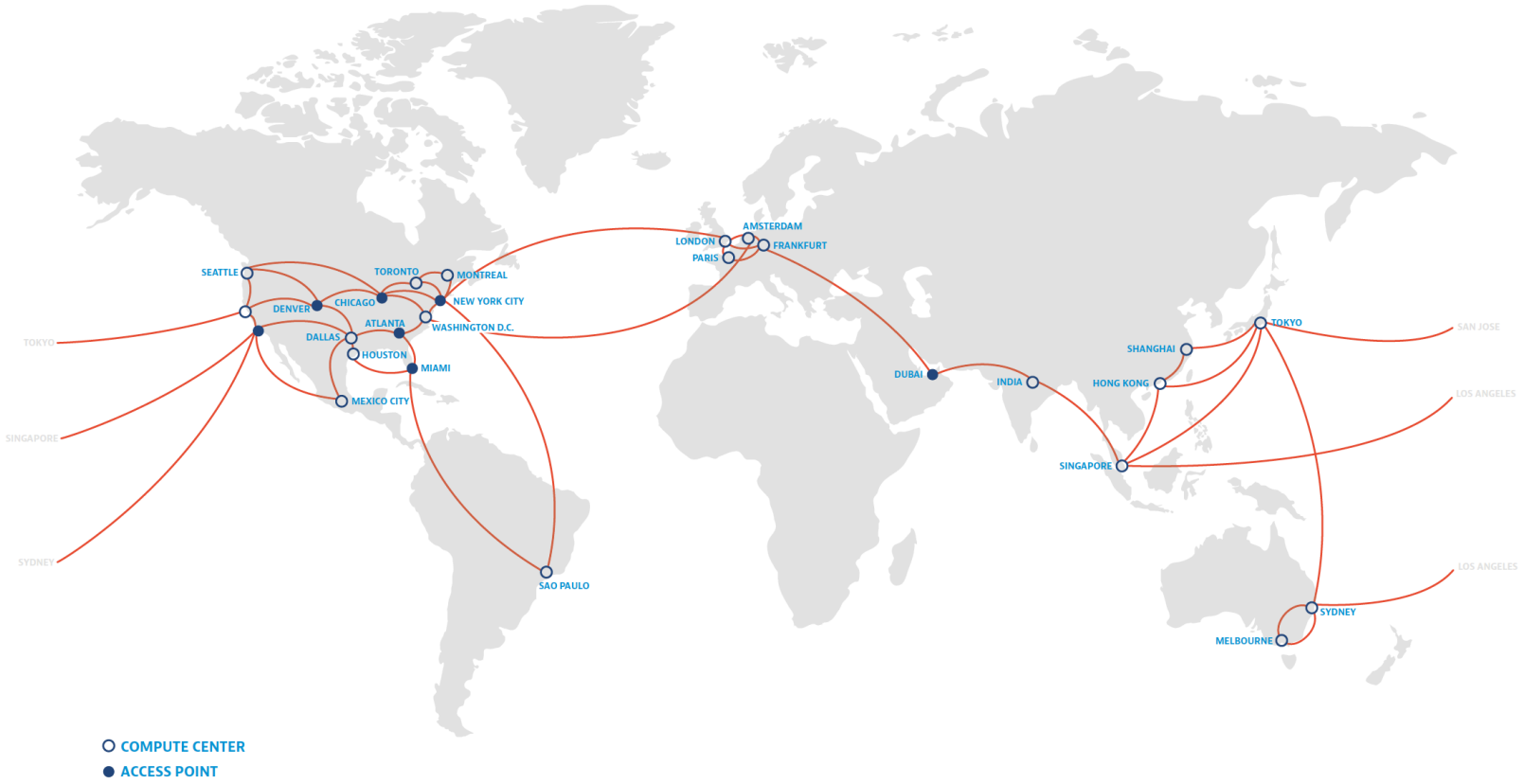
- ✓ Save time and improve efficiency through automation, substantially decrease TCO

High performance

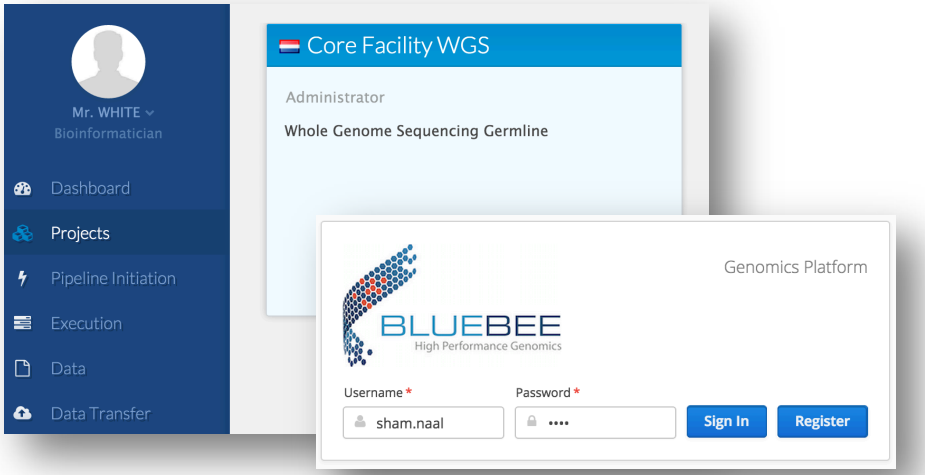
- Faster processing and increased throughput
  - By combining accelerated hardware with distributed computing, mass NGS data volumes are processed in the shortest possible time frames
- Performance is tuned to the need
  - Processing time vs. throughput
- On-the-fly provisioning of computational and data storage capacity

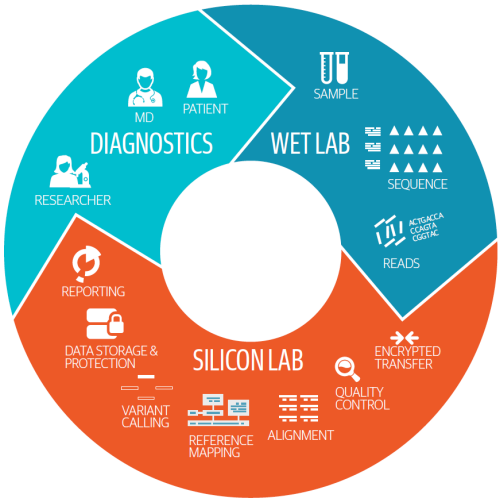


Data locality



Bluebee Genomics Platform

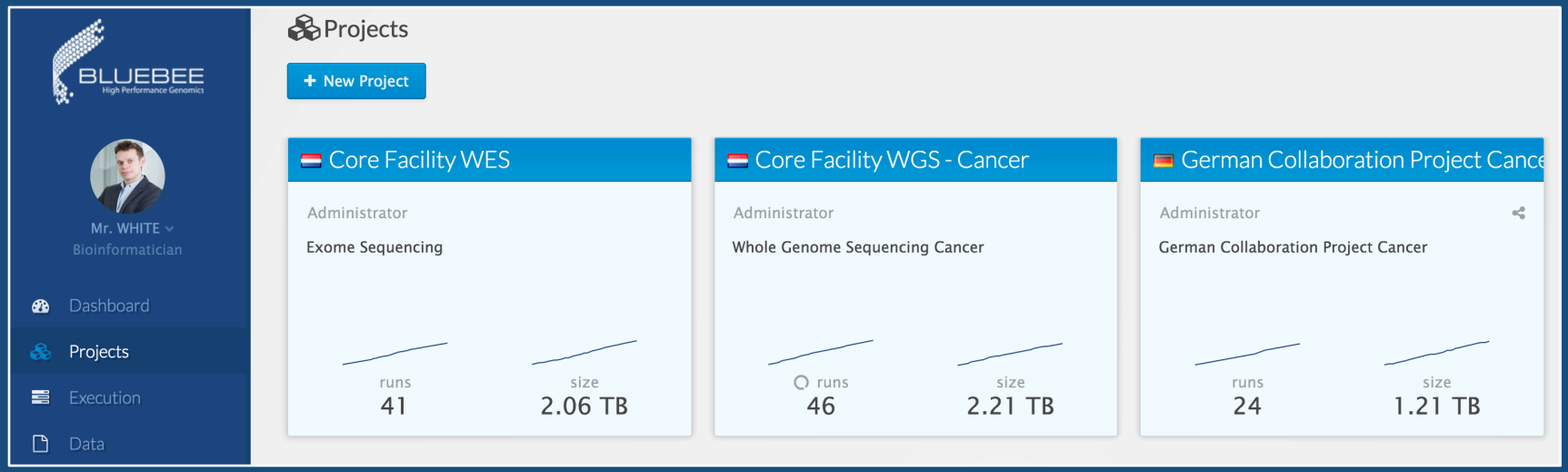


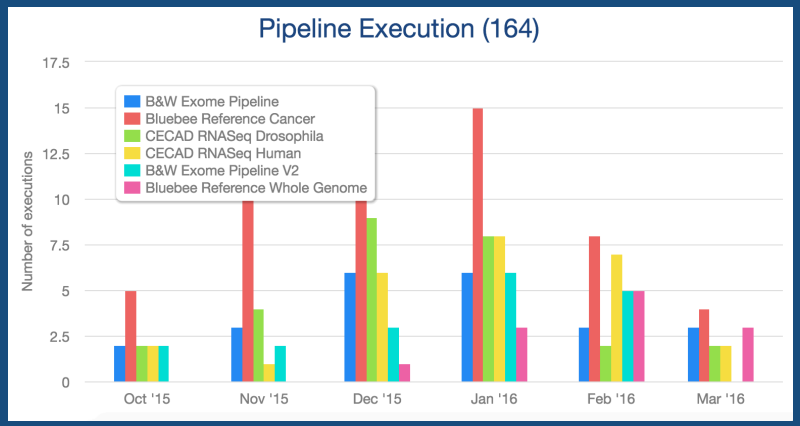


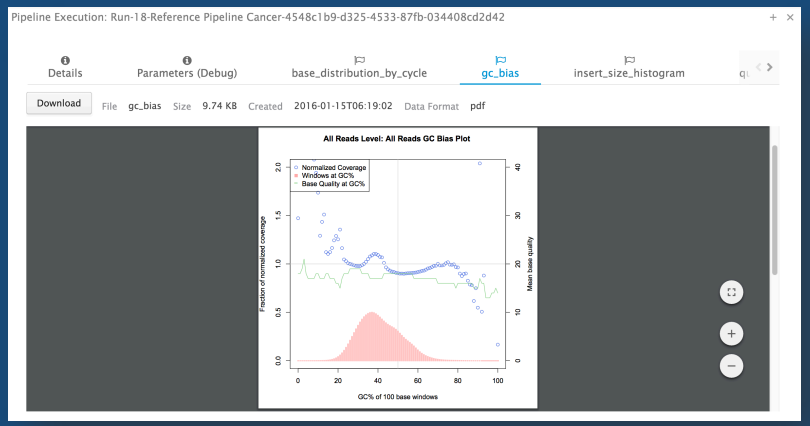
Private Cloud

- state of the art security
- long term data storage
- anonymized data
- data compression
- 4 to 6 eyes control & approval workflow

Request a demo or a trial, talk to our experts at booth 426-428







dkfz.

GERMAN CANCER RESEARCH CENTER
HEIDELBERG, GERMANY

BLUEBEE
High Performance Genomics

dkfz. German Cancer Research Center

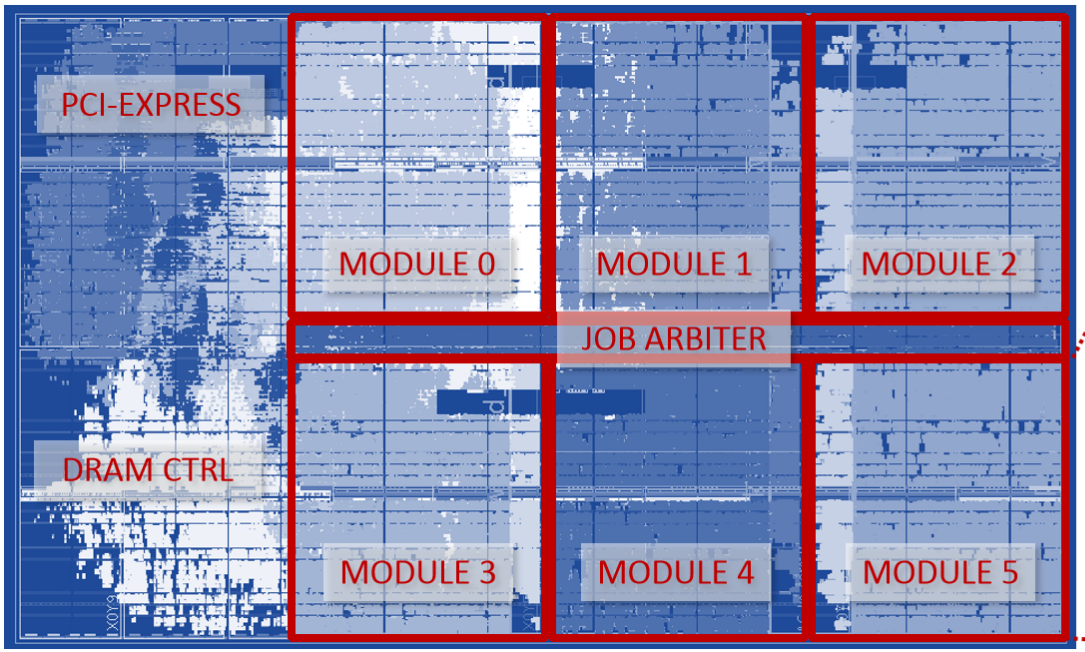
- Location: Heidelberg, Germany
- Research Staff: 2800
- Projects and Topics:
  - International Cancer Genome Consortium, ICGC
    - Cancer genome bioinformatics for 78 different tumor types and subtypes, 5 of which are studied at Heidelberg.
  - ICGC PanCancer
    - Reanalysis of whole genome sequencing data of ~3000 tumor-normal pairs from 25+ cancer types.
  - Personalized Oncology in collaboration with Heidelberg University Medical Center
    - Last year, over 1800 patients were part of the program, a number that will increase to 4000 by 2017 when the full capacity of the XTen sequencers is reached.
- Sequencers
  - Exome: Illumina HiSeq 2000 and HiSeq 2500
  - Whole genome: Illumina HiSeq X Ten
- Data Generation
  - Growing to 3TB per day, adding-up to ~1PB per year.

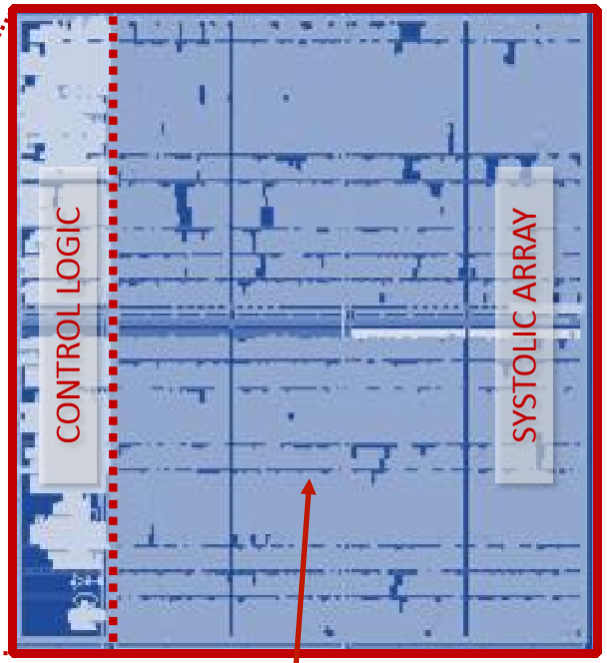
Project focus

- Background
  - Introduction of Illumina X10 Sequencers
    - Read lengths increased to 150bp
    - Required migration from BWA-ALN to BWA-MEM
  - BWA-MEM has been introduced in the PanCancer project and will be the standard for all further WGS sequencing projects at DKFZ
- The Challenge
  - BWA was taking up to 80% of CPU time in the data centre's pipeline
  - Because BWA was already optimised for standard computer infrastructure, further speed up on the existing cluster was not possible
- Clear requirement
  - The BWA was chosen for the alignment step after an evaluation process
    - Although a number of commercial aligners offered by different parties held the promise of higher speed, the gain in speed did not offset against the loss of (backward) compatibility of the results.

BWA-MEM physical layout

AlphaData (Xilinx Virtex-7) FPGA implementation contains six modules:





Local alignment
(Smith-Waterman)

From Houtgast, Sima, Marchiori, Bertels, Al-Ars
(FCCM 2016)

BWA-MEM Acceleration

SYSTEM	FPGA ENABELED	THROUGHPUT (MBASES/S)
DKFZ baseline (traditional implementation of BWA-MEM)	No	1
Convey xxxxx (describe config)	Yes	3,5
Intel Core i7-4770 CPU @ 3.40GHz + AlphaData 7V3	Yes	3.73
IBM POWER8 v2 @ 3.5 GHz, Tyan Habanero TN71-BP012	No	4.31
IBM POWER8 v2 @ 3.5 GHz, Tyan Habanero TN71-BP012 + AlphaData 7V3	Yes	7.18
Dual Socket IBM POWER8 v2.1 @ 3.425 GHz, IBM pSeries S822L (8247-22L1)	No	7.23
Dual Socket IBM POWER8 v2.1 @ 3.425 GHz, IBM pSeries S822L (8247-22L1) + 2x AlphaData 7V3	Yes	12.73

BWA-MEM Acceleration

Host System	MBASES/S	
	Base	FPGA enabled
IBM POWER8 v2 @ 3.5 GHz, Tyan Habanero TN71-BP012	4.31	7.18 (AlphaData 7V3)
Dual Socket IBM POWER8 v2.1 @ 3.425 GHz, IBM pSeries S822L (8247-22L1)	7.23	12.73 (2x AlphaData 7V3)

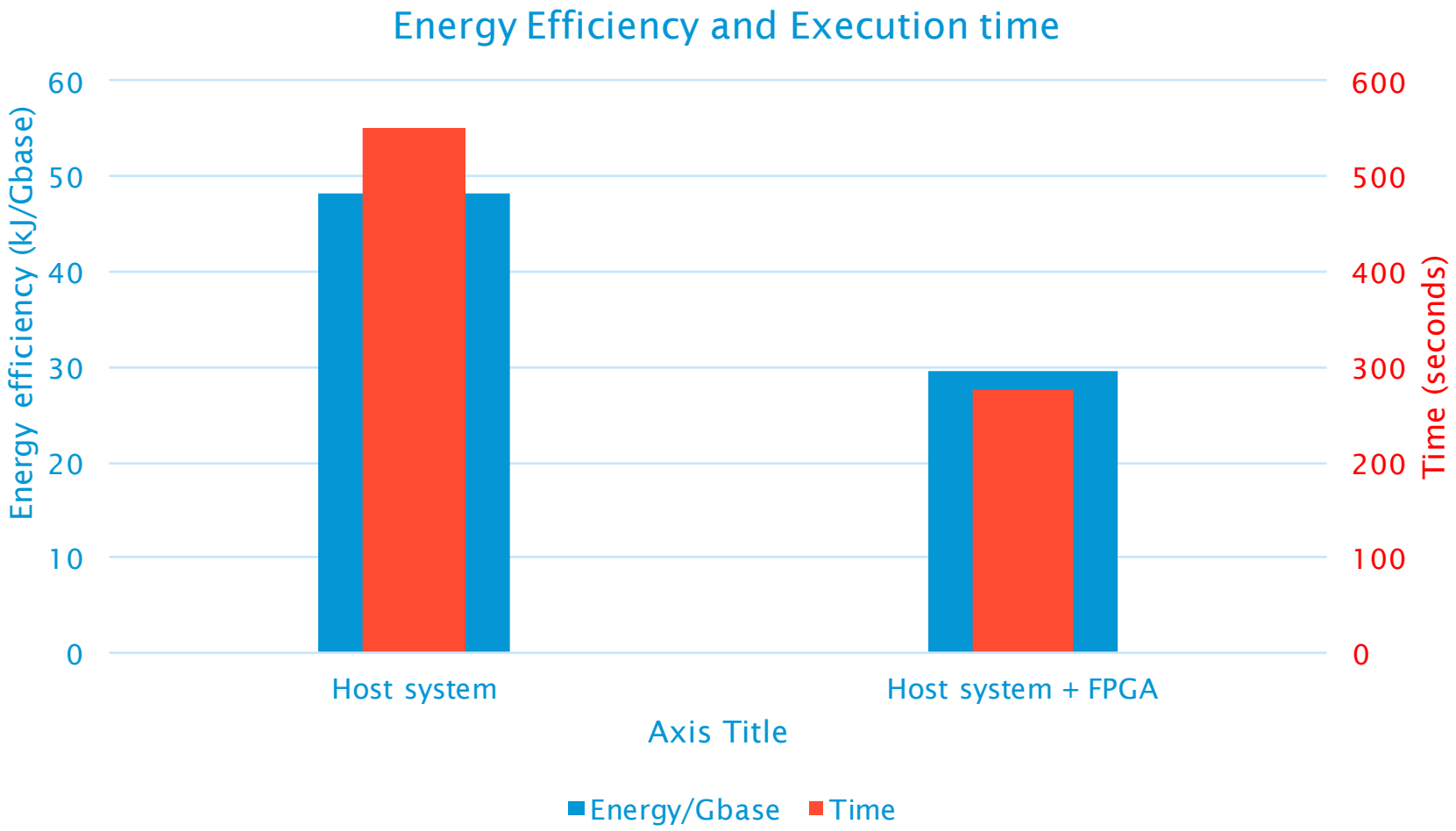
Acceleration benefits

- Current implementation on Convey : 3 – 4 x speed-up depending on data
- New release on P8+2x Alpha Data : additional speed-up of XXX
- Identical result compare to non-accelerated version
- 800 CPU-hours saved per whole genome pair (tumor and matched normal)
- Removed bottleneck for onboarding Illumina X Ten sequencers

> *"Depending on the data, we are typically seeing between 10-20x acceleration for BWA-ALN and 3-4x acceleration for BWA-MEM, which is substantial, and our entire bioinformatics workflow is reduced by 50%, which directly translates in doubling our capacity."*
> Dr. Barbara Hutter, Team Leader Clinical Bioinformatics at DKFZ

Power Efficiency

FPGAs provide twice the performance of the CPU for half the energy requirements

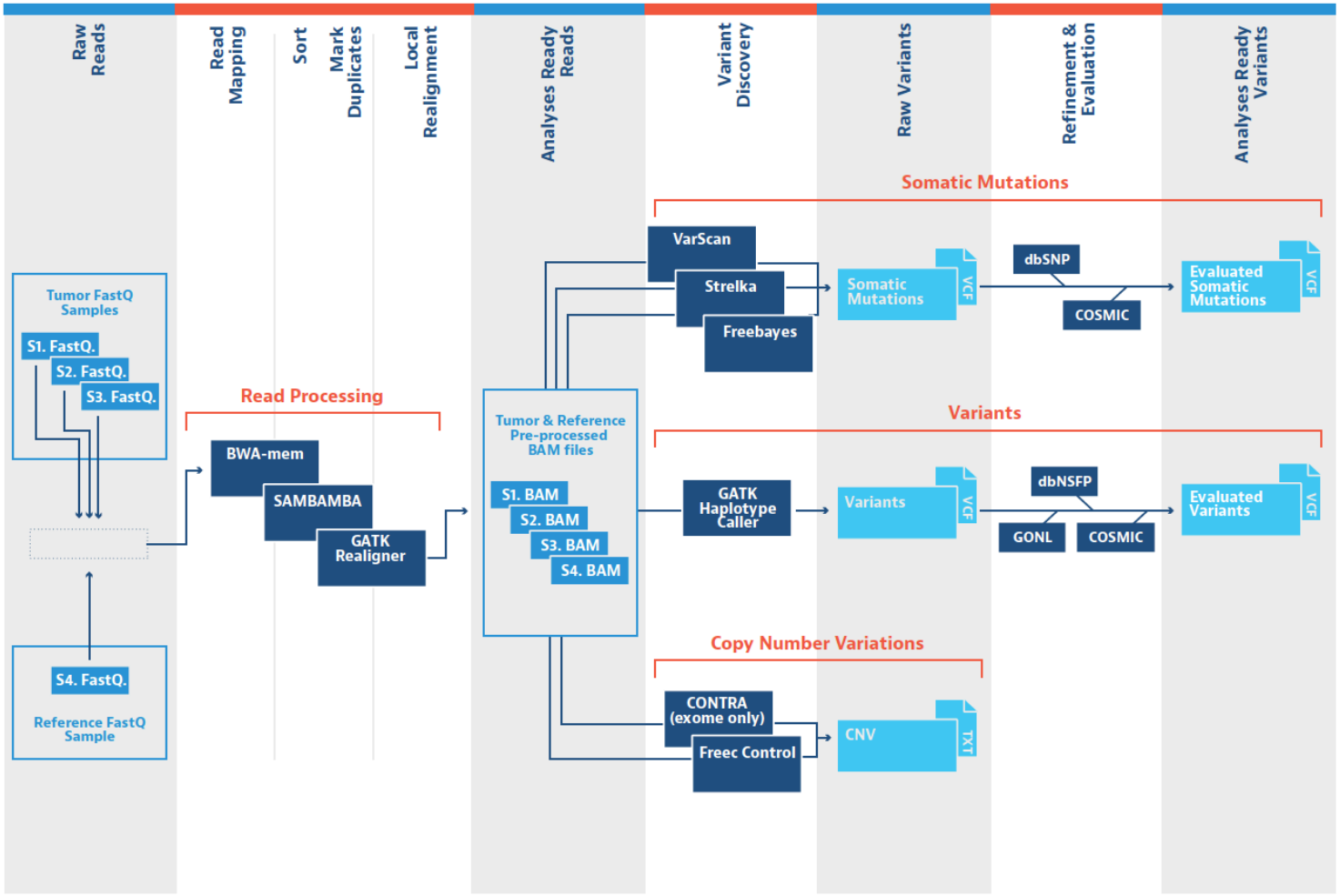


HARTWIG MEDICAL FOUNDATION
AMSTERDAM, THE NETHERLANDS

Hartwig Medical Foundation

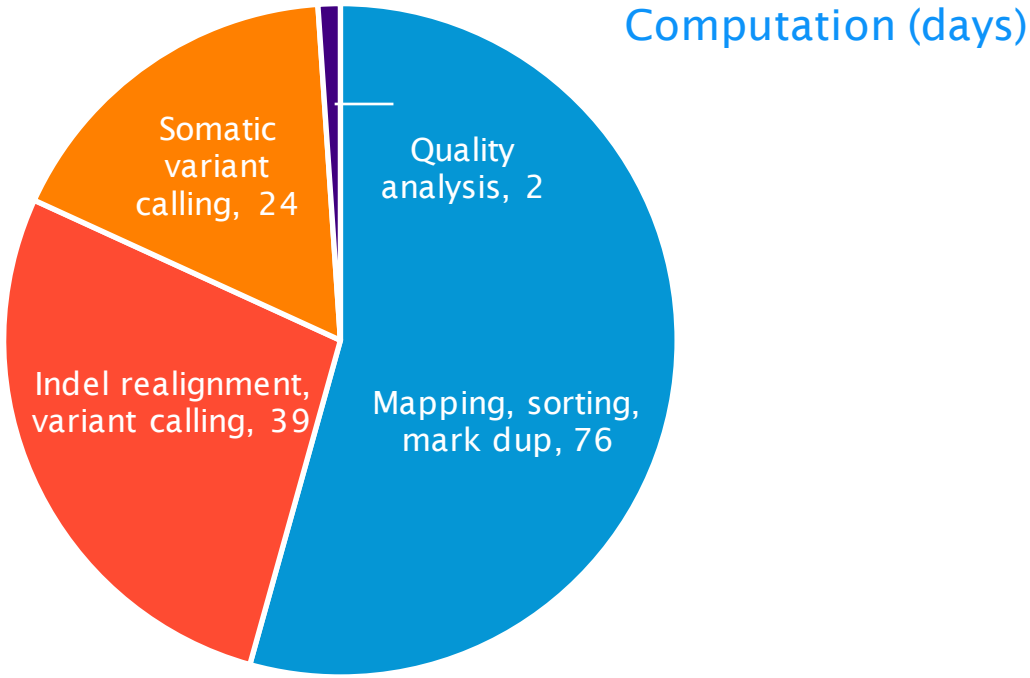
- Location: Amsterdam, The Netherlands
- Projects and Topics:
  - Center for Personalized Cancer Treatment
    - Whole genome sequencing and transcriptome analysis of pre-treatment tumor biopsies. Establishment of a database with clinical, genetic and treatment outcome data.
  - Clinical Genetics Consortium
    - Collaboration between five out of the nine clinical genetics centers in the Netherlands to implement Whole Genome Sequencing for routine germline diagnostics
  - National sequencing infrastructure for research.
    - Facilitating WGS for a broad range of research projects
- Sequencers
  - Exome: Illumina NextSeq
  - RNA-seq: Illumina Next Seq
  - Targeted panels: IonTorrent PGM
  - Whole genome: Illumina Xten
- Data Generation
  - 8-10 Tb/week

Pipeline

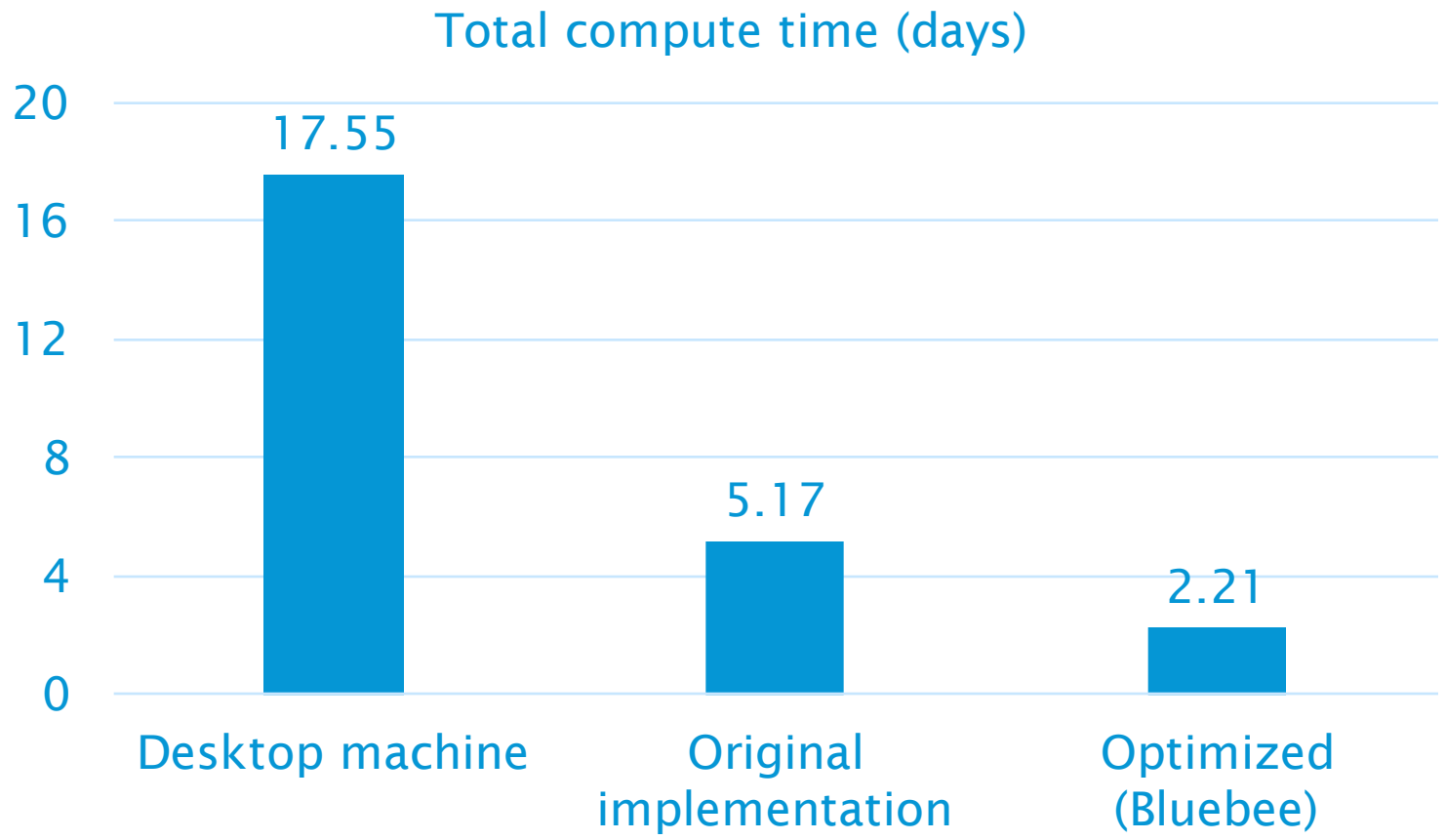


Input data

- Libraries sequenced on an Illumina HiSeq XTEN machine
- Total size of 266 Gbases (or 356 Gbytes of fastq.gz)
- Dataset is tumor/normal set of 90/30x based on NA24385 with a 30% spike in of NA12878



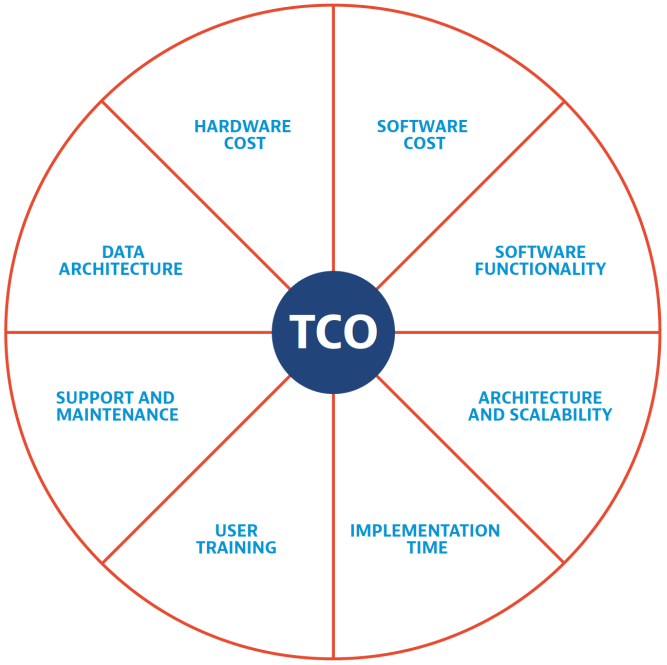
Acceleration achieved



Benefits

- Faster diagnose
- Predictable turn around times
- Scalability, guaranteed turn around times regardless of the number of patient
- Unlimited storage capacity
- Fast and predictable data transfer
- Fixed cost per patient

=> Substantially decreased TCO



Questions?

BLUEBEE
High Performance Genomics

Thank you!

kurt.florus@bluebee.com

BLUEBEE
High Performance Genomics