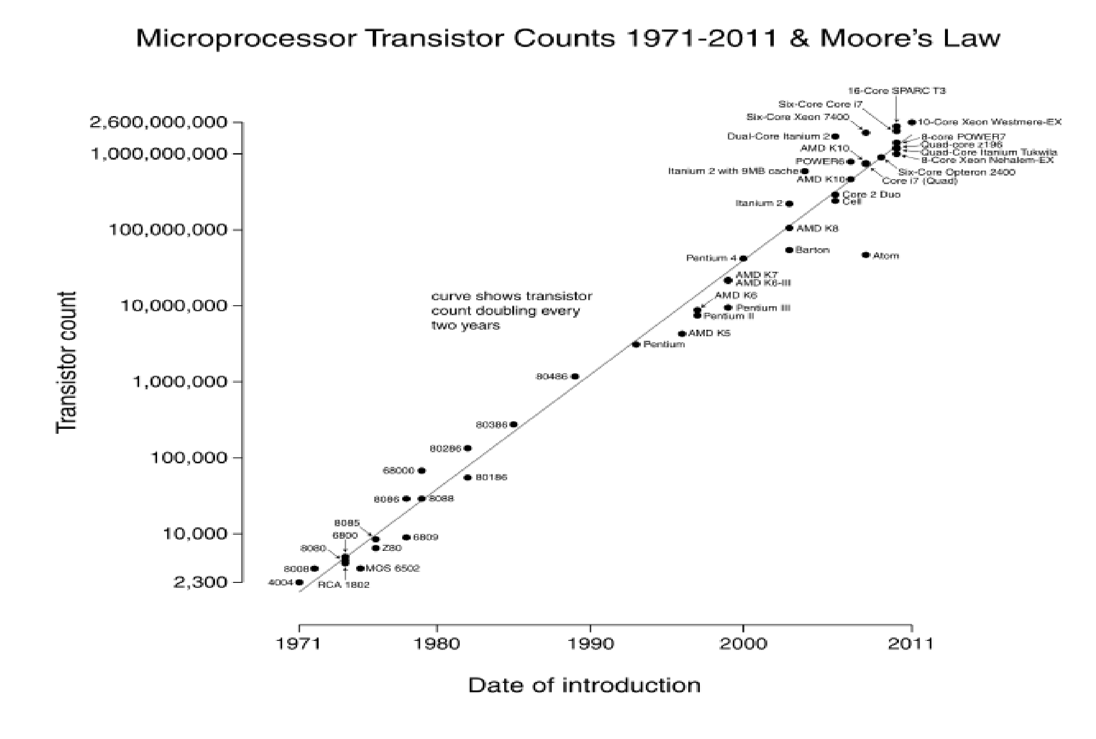
# Personalized Medicine

Big Data "IT" in Health and Life Sciences

Paolo Narvaez
Principal Engineer
Health and Life Sciences

# How we started

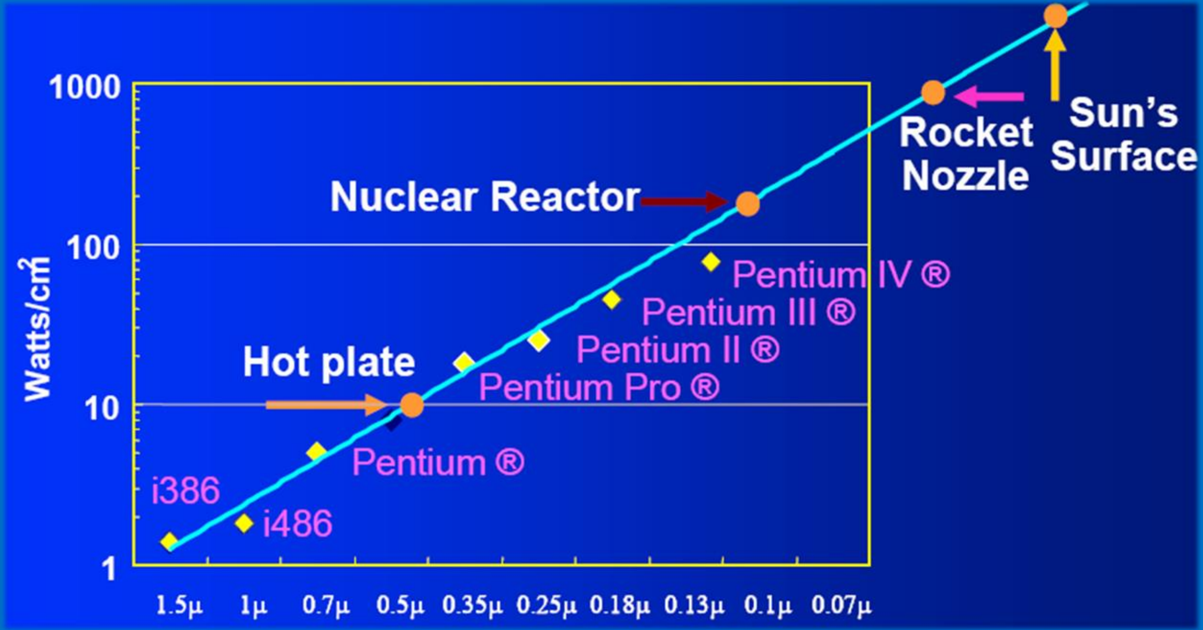


Microprocessor Transistor Counts 1971-2011 & Moore's Law

Source: Wikipedia

Moore's Law is awesome, but…

# Tectonic Shift

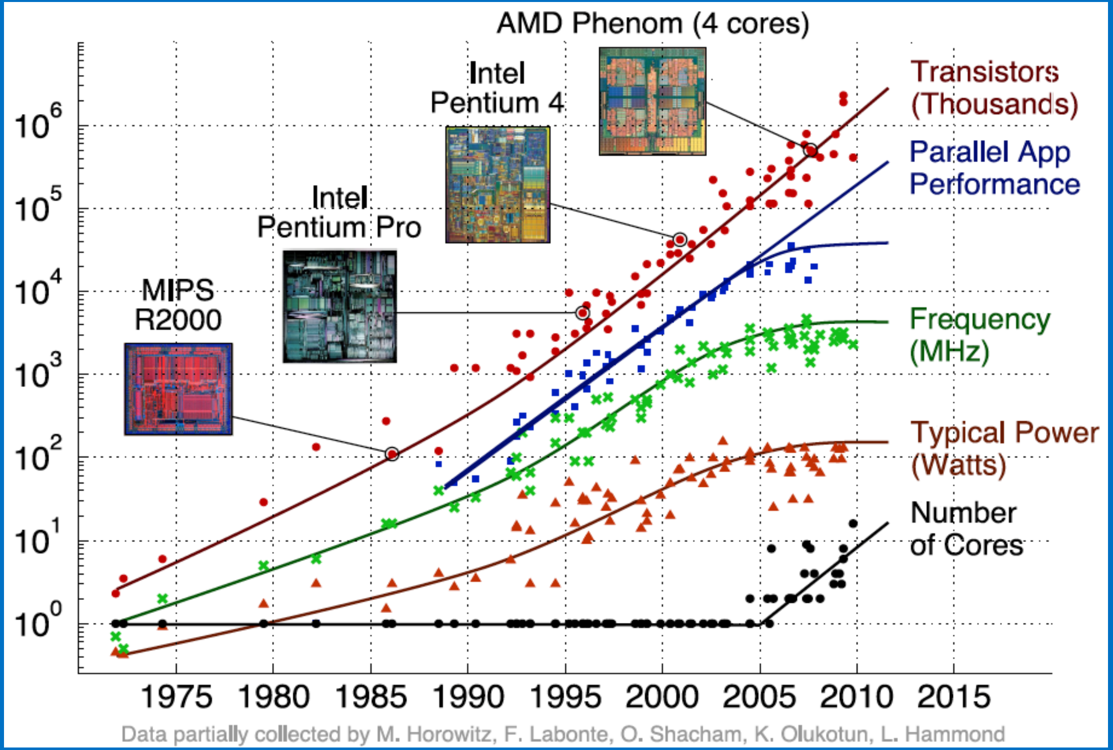


Source: Fred Pollack, Keynote – MICRO'32

Pollack's rule: performance increase due to microarchitecture advances is roughly proportional to square root of increase in complexity [area]

Power Consumption limits single-thread performance

# New Computing Paradigm



Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond

UC PAR Lab Presentation – Krste Asanovic – May24,2010

Future improvements in performance will require taking advantage of parallelization and specialization techniques.

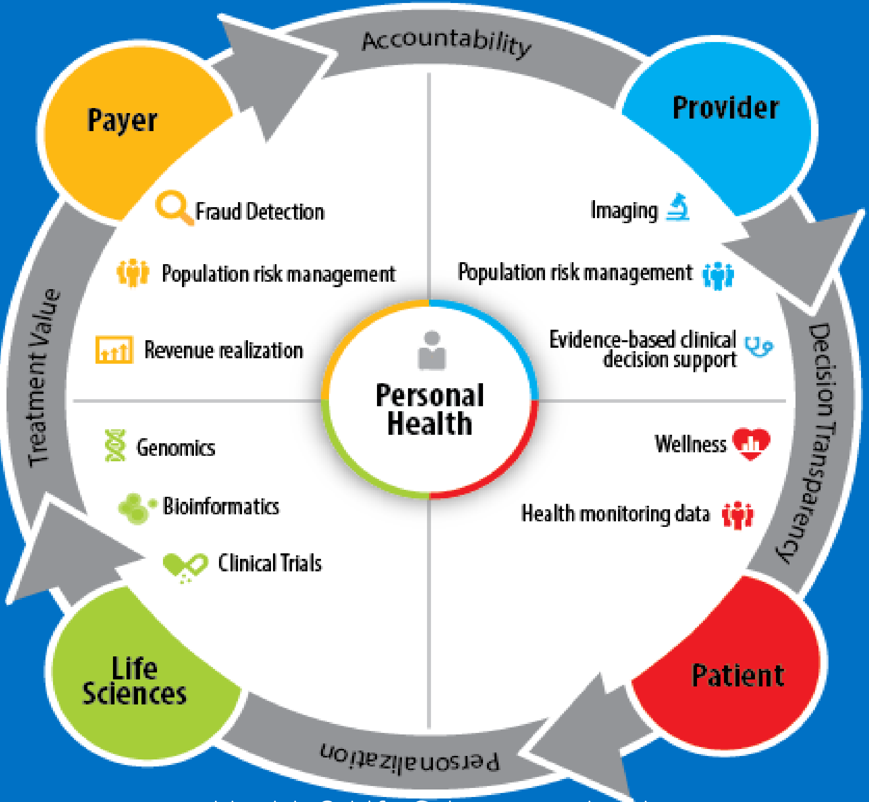
# Parallelization and Specialization

- Parallelization – Run computation on many low-power cores
- Specialization – Run computation on most energy-efficient hardware

- Hardware Repertoire
  - Symmetric Multiprocessing
  - Vector Units – Single Instruction, Multiple Data (SIMD)
  - New specialized instructions (e.g., AES-NI)
  - Integrated graphics processor
  - Heterogeneous Computing - Co-processor
    - GPUs, Xeon Phi
  - Fixed logic accelerators - Offload
  - Programmable logic - FPGA

Designing for this complex ecosystem requires deep understanding of workloads and tighter collaboration with domain experts and software developers.

# Personalized Medicine =
# Complex Big Data and Compute Ecosystem



Health & Life Sciences at Intel
Where information and care meet

# Life Sciences :: Key Industry **Challenges and Solutions**

- Many (most) applications are single-threaded, single address space

  ***Intel is delivering optimizations working with open source community, developing NGS+HPC curriculum***

- Some algorithms scale poorly with the size of the problem. Large data sets exceed available memory and storage

  ***Innovations in acceleration, compute, storage, networking, security, and *-as-a-service.***

- International collaboration is an imperative, bioinformatics expertise is scarce

- ***Intel is working closely with the ecosystem to address enterprise to cloud transmission of terabyte payloads***

- Databases are distributed, data is siloed and will likely stay that way

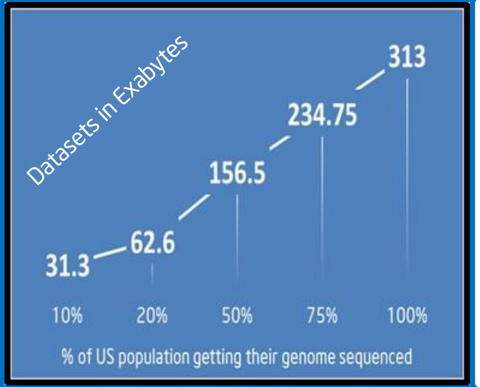
  ***Tools like Hadoop, Lustre, Graphlab, In-Memory Analytics, Security etc.***

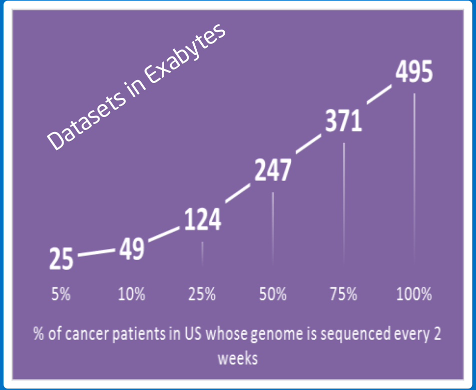
## Need for Efficient Compute Ecosystem

*Other names and brands may be claimed as the property of others.

# Recent Collaborations

# Genomics - Big Data Problem





313 *Exabytes*
if everyone in the US has their genes sequenced

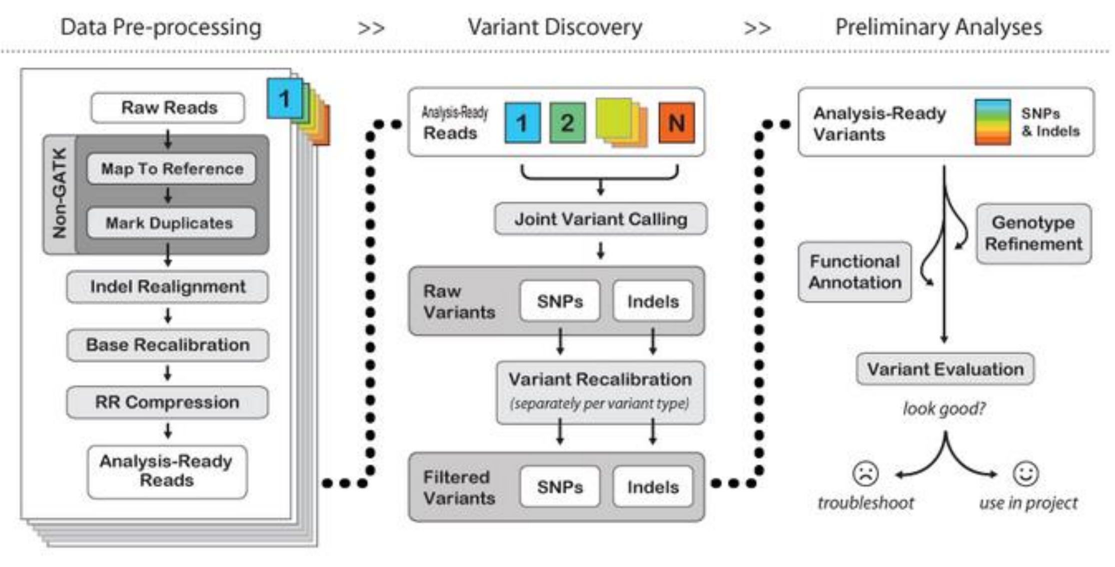
495 *Exabytes*
if every cancer patient in the US has their genes sequenced every 2 weeks.

Source: Knights Cancer Institute, Oregon Health Sciences University & Intel

This is a key area with a large growth potential.
Goal is to anticipate demand for compute, provide efficient solutions, and help grow the market.

**Energy and Total Cost of Operation are key**

# DNA Pipeline - GATK Best Practices



*Other names and brands may be claimed as the property of others.

Look Inside

# DNA Pipeline: BWA+GATK

## Whole Genome Sample: ~65x Coverage

| Step Tool | # of Threads | Runtime (hours) |
|---|---|---|
| Read Alignment (bwa) | 16 | 8 |
| Sampe (bwa) | 1 | 24 |
| Import  (samtools) | 1 | 11 |
| Sort + Index (samtools) | 1 | 14.5 |
| MarkDuplicates (picardtools) + Index | 1 | 11.5 |
| UnifiedGenotyper* (GATK) | 16 | 7.5 |
| SomaticIndelDetector (GATK) | 1 | 3 |
| RealignerTargetCreator (GATK) | 16 | 0.8 |
| IndelRealigner* (GATK) + Index | 1 | 17.5 |
| BaseRecalibrator*(GATK) | 1 | 62 |
| PrintReads* (GATK) + Index + Flagstat | 1 | 25 |
| TOTAL (hours) |  | 177 |

| Step | # of Threads | Runtime (hours) |
|---|---|---|
| Read Alignment (bwa mem) | 24 | 7 |
| View (samtools) | 24 | 2 |
| Sort + Index (samtools) | 24 | 3 |
| MarkDuplicates (picardtools) + Index | 1 | 11 |
| RealignerTargetCreator (GATK) | 24 | 1 |
| IndelRealigner* (GATK) + Index | 24 | 6.5 |
| BaseRecalibrator*(GATK) | 24 | 1.3 |
| PrintReads* (GATK) + Index + Flagstat | 24 | 12.3 |
| TOTAL (hours) |  | 44 |

🟩 Algorithmic Improvement

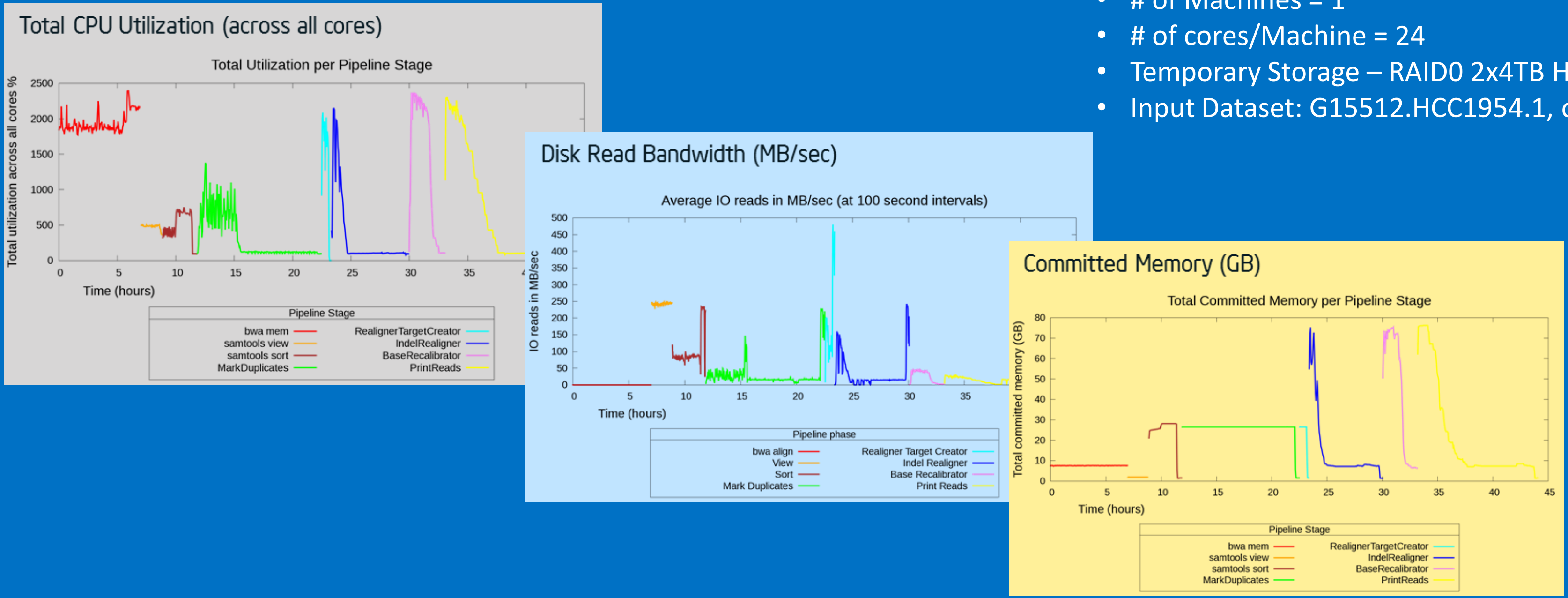
🟧 Thread-level Parallelism

🟦 Cluster-level Parallelism

# Profiling: Single Instance Run of GATK

*GATK: Genome Analysis Toolkit*

- # of Machines = 1
- # of cores/Machine = 24
- Temporary Storage – RAID0 2x4TB HDD
- Input Dataset: G15512.HCC1954.1, coverage: 65x



Total CPU Utilization (across all cores)

Disk Read Bandwidth (MB/sec)

Committed Memory (GB)

Average CPU utilization is very low. Most cores not being used

Average I/O bandwidth is very low. Application not I/O bound

Average memory footprint is small. Application not using memory available in newer systems

**There is a lot of room to improve**

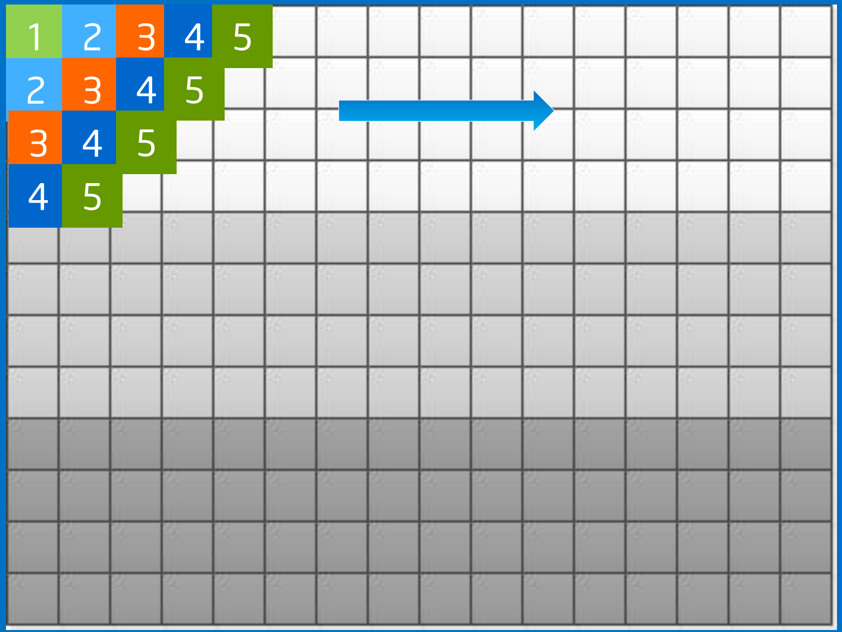
*Other names and brands may be claimed as the property of others.

intel Look inside

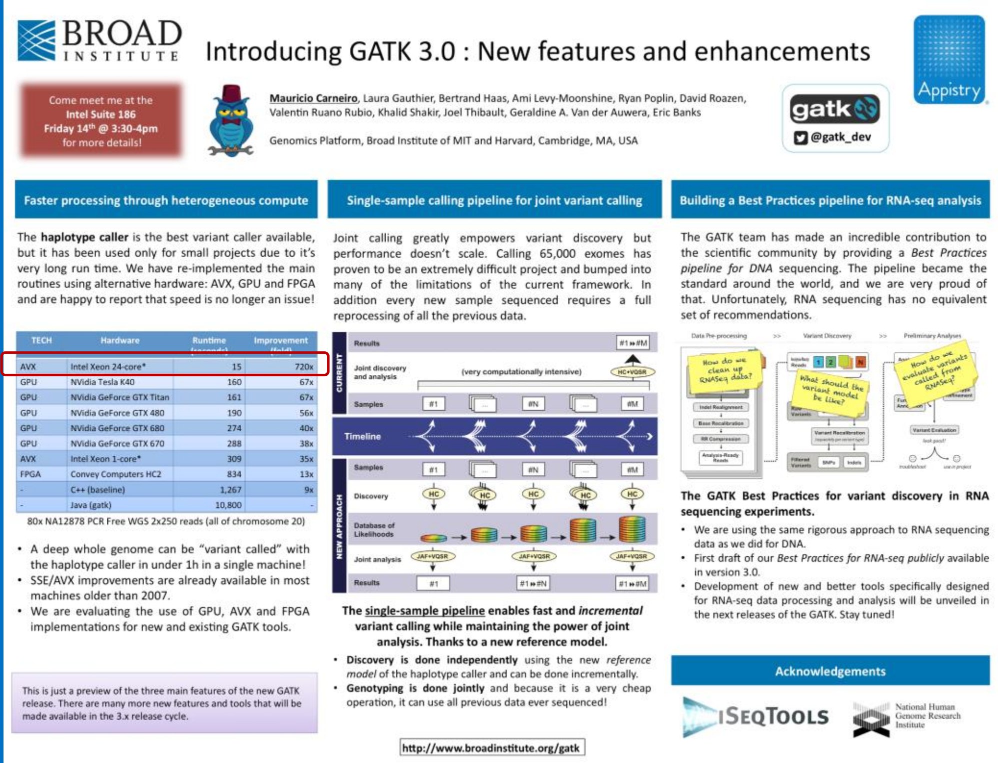
# PairHMM Computation Kernel in Java

```
 /**
     * Updates a cell in the HMM matrix
     *
     * The read and haplotype indices are offset by one because the state arrays have an extra column to
hold the
     * initial conditions

     * @param indI             row index in the matrices to update
     * @param indJ             column index in the matrices to update
     * @param prior            the likelihood editing distance matrix for the read x haplotype
     * @param transition        an array with the six transition relevant to this location
     */
    protected void updateCell( final int indI, final int indJ, final double prior, final double[]
transition) {

        matchMatrix[indI][indJ] = prior * ( matchMatrix[indI - 1][indJ - 1] * transition[matchToMatch] +
                                                 insertionMatrix[indI - 1][indJ - 1] *
transition[indelToMatch] +
                                                 deletionMatrix[indI - 1][indJ - 1] *
transition[indelToMatch] );

        insertionMatrix[indI][indJ] = matchMatrix[indI - 1][indJ] * transition[matchToInsertion] +
                                                 insertionMatrix[indI - 1][indJ] *
transition[insertionToInsertion];

        deletionMatrix[indI][indJ] = matchMatrix[indI][indJ - 1] * transition[matchToDeletion] +
                                                 deletionMatrix[indI][indJ - 1] *
transition[deletionToDeletion];
    }
```

# PairHMM Wave-Front Computation in AVX



# Improvements in GATK 3



BROAD INSTITUTE

## Introducing GATK 3.0 : New features and enhancements

Come meet me at the
**Intel Suite 186**
**Friday 14th @ 3:30-4pm**
for more details!

**Mauricio Carneiro**, Laura Gauthier, Bertrand Haas, Ami Levy-Moonshine, Ryan Poplin, David Roazen, Valentin Ruano Rubio, Khalid Shakir, Joel Thibault, Geraldine A. Van der Auwera, Eric Banks

Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA

gatk
@gatk_dev

Appistry

### Faster processing through heterogeneous compute

The **haplotype caller** is the best variant caller available, but it has been used only for small projects due to it's very long run time. We have re-implemented the main routines using alternative hardware: AVX, GPU and FPGA and are happy to report that speed is no longer an issue!

| TECH | Hardware | Runtime (seconds) | Improvement (fold) |
|---|---|---|---|
| AVX | Intel Xeon 24-core* | 15 | 720x |
| GPU | NVidia Tesla K40 | 160 | 67x |
| GPU | NVidia GeForce GTX Titan | 161 | 67x |
| GPU | NVidia GeForce GTX 480 | 190 | 56x |
| GPU | NVidia GeForce GTX 680 | 274 | 40x |
| GPU | NVidia GeForce GTX 670 | 288 | 38x |
| AVX | Intel Xeon 1-core* | 309 | 35x |
| FPGA | Convey Computers HC2 | 834 | 13x |
| - | C++ (baseline) | 1,267 | 9x |
| - | Java (gatk) | 10,800 | - |

80x NA12878 PCR Free WGS 2x250 reads (all of chromosome 20)

- A deep whole genome can be "variant called" with the haplotype caller in under 1h in a single machine!
- SSE/AVX improvements are already available in most machines older than 2007.
- We are evaluating the use of GPU, AVX and FPGA implementations for new and existing GATK tools.

This is just a preview of the three main features of the new GATK release. There are many more new features and tools that will be made available in the 3.x release cycle.

### Single-sample calling pipeline for joint variant calling

Joint calling greatly empowers variant discovery but performance doesn't scale. Calling 65,000 exomes has proven to be an extremely difficult project and bumped into many of the limitations of the current framework. In addition every new sample sequenced requires a full reprocessing of all the previous data.

**The single-sample pipeline enables fast and *incremental* variant calling while maintaining the power of joint analysis. Thanks to a new reference model.**

- **Discovery is done independently** using the new *reference model* of the haplotype caller and can be done incrementally.
- **Genotyping is done jointly** and because it is a very cheap operation, it can use all previous data ever sequenced!

http://www.broadinstitute.org/gatk

### Building a Best Practices pipeline for RNA-seq analysis

The GATK team has made an incredible contribution to the scientific community by providing a *Best Practices pipeline for DNA* sequencing. The pipeline became the standard around the world, and we are very proud of that. Unfortunately, RNA sequencing has no equivalent set of recommendations.

**The GATK Best Practices for variant discovery in RNA sequencing experiments.**

- We are using the same rigorous approach to RNA sequencing data as we did for DNA.
- First draft of our *Best Practices for RNA-seq publicly* available in version 3.0.
- Development of new and better tools specifically designed for RNA-seq data processing and analysis will be unveiled in the next releases of the GATK. Stay tuned!

Acknowledgements

ISEQTOOLS

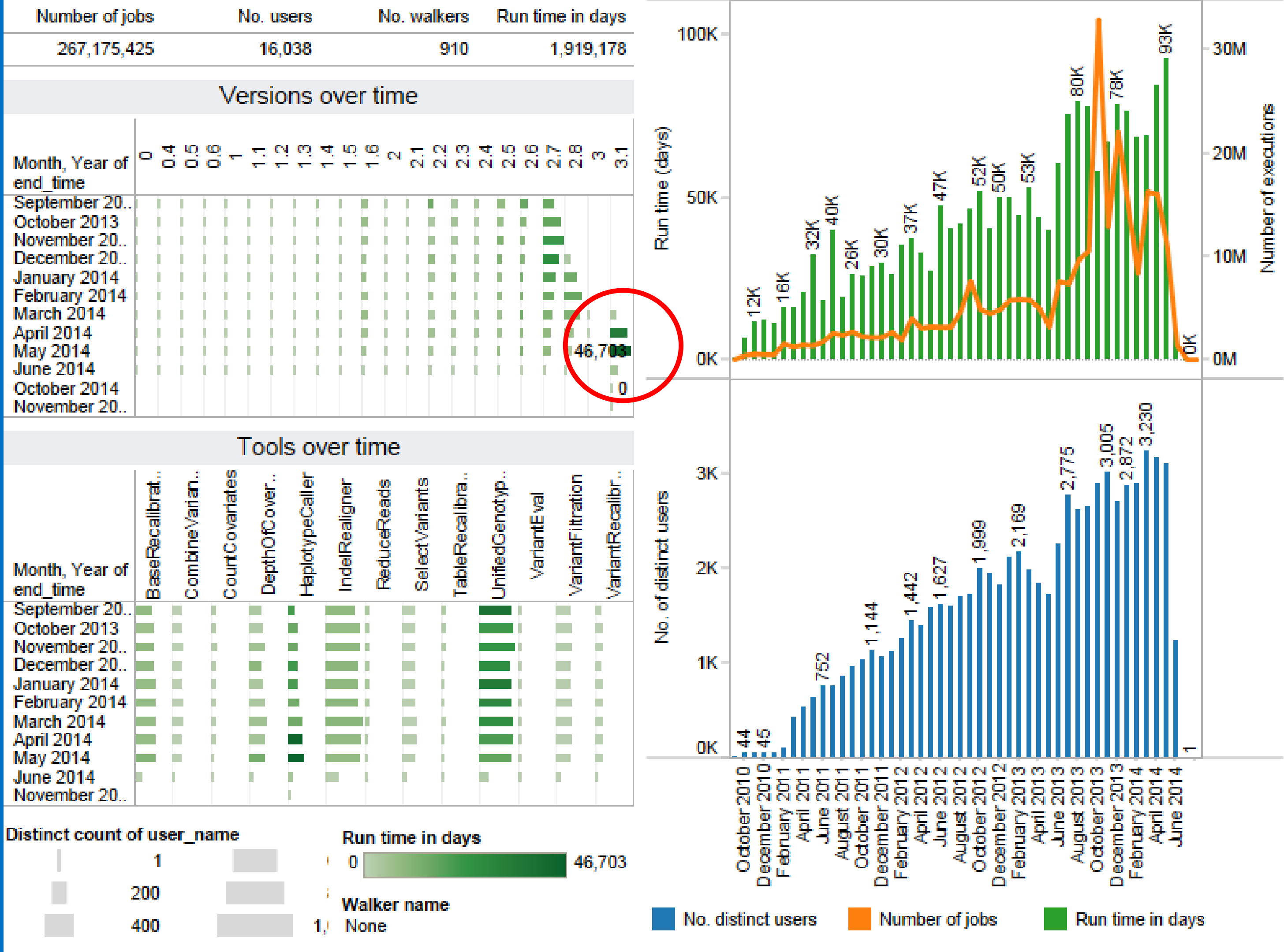
National Human Genome Research Institute

- Pair HMM Acceleration using Intel® AVX resulted in **970x speedup**
  - Computation kernel and bottleneck in GATK Haplotype Caller
  - AVX enables 8 floating point SIMD operations in parallel

| | Time (seconds) | Speedup C++/Java |
|---|---|---|
| Serial C++ | 1540 | 1x / 9x |
| 1 core with AVX (Intra) | 340 | 4.5x / 40.7x |
| 1 core with AVX (Inter) | 285 | 5.4x / 48.6x |
| 24 cores with AVX (Inter) | 14.3 | 108x / 970x |
| 24 cores hybrid (Inter) | 15.7 | 98x / 882x |

*Other names and brands may be claimed as the property of others.

# GATK downloads over time.



| Number of jobs | No. users | No. walkers | Run time in days |
|---|---|---|---|
| 267,175,425 | 16,038 | 910 | 1,919,178 |

# Applications and Workloads **Optimized** on Intel Architecture

- Focus on improving genomics, molecular dynamics pipelines
- Optimize individual applications (node and cluster); Work with code authors to release optimizations

| DOMAIN | Applications | Intel® Architecture Target |
|---|---|---|
| Genomics | **Bowtie 1*, Bowtie 2*** | Xeon® processor |
| | **BWA*** | Xeon® processor |
| | **BLAST*** | Xeon® processor |
| | **GATK*** | Xeon® processor |
| | **HMMER*** | Xeon® processor<br>Xeon® Phi™ coprocessor |
| | **Abyss*** | Xeon® processor |
| | **Velvet*** | Xeon® processor |

| DOMAIN | Applications | Intel® Architecture Targets |
|---|---|---|
| Molecular Dynamics/ Chemistry | **AMBER*** | Xeon® processor<br>Xeon® Phi™ coprocessor |
| | **NAMD*** | |
| | **GROMACS*** | |
| | **GAMESS*** | |
| | **Quantum Espresso*** | |
| | **Gaussian*** | |
| | **VASP*** | |
| | **CP2K*** | |
| | **QBOX*** | |
| | **CPMD*** | |
| | **LAMMPS*** | |

*Other names and brands may be claimed as the property of others.

AYASDI Cure

*Turning Data into Therapies*

Biomarker Discovery · Drug Target Discovery · Precision Medicine

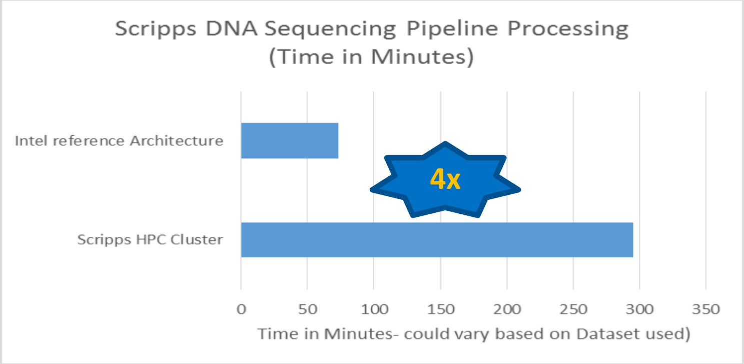
- **Challenge**: Ayasdi Cure™ analyzes highly complex, large data sets and relies on fast computation times to provide real-time output.
- **Solution**:
  - Intel® AVX instructions - **four double-precision floating-point operations in parallel** vs. one.
  - Intel® MKL library - accelerate filter computations
- **Benefits: 400% performance increase in distance computation.**

Scripps Translational Science Institute

Scripps ADVISER Annotation and Distributed Variant Interpretation Server

## Scripps DNA Sequencing Pipeline

- **Challenge**: Processing times, Logistical Delays, Cluster complexity
- **Solution**: Intel® Xeon® E7-4800 series using SSDs
- **Benefits**: **~4x Improvement on processing times**



*Other names and brands may be claimed as the property of others.

## Ultra High-Speed Networking Optimizations

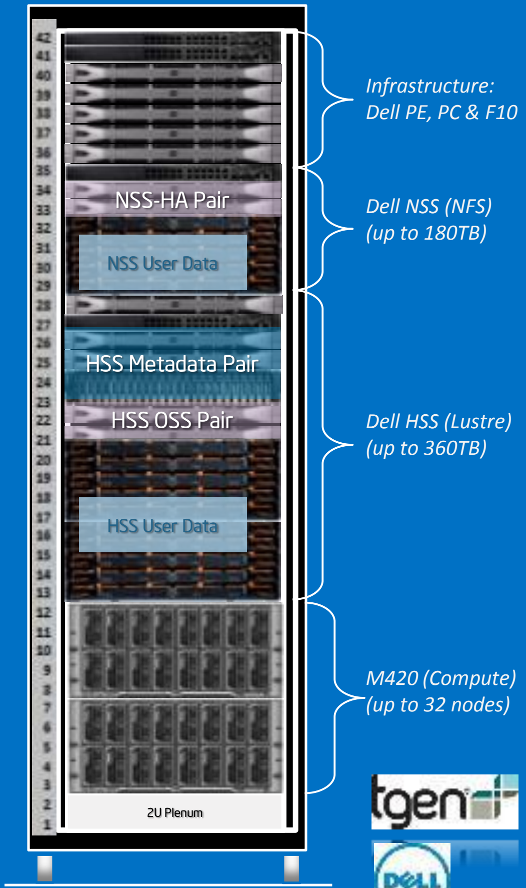
- **Challenge:** Improving big data transfer to and from the backend data center
- **Solution:**
  - Optimize ultra high-speed (10+ Gbps) data transfer solutions built on Aspera's FASP ™ technology
  - Intel® Xeon® E5-2600 (DDIO, SR-IOV)
- **Benefits:**
  - **300% improvement in transfer throughput**
  - **Physical or virtual, LAN or WAN – same transfer speeds**

## High Performance Scale-out Storage Challenge:

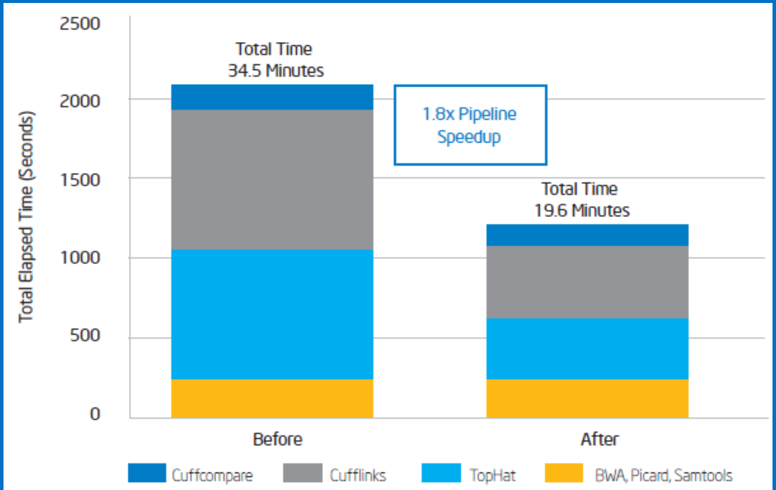
- **Challenge**: 10-15TB data added weekly, small fraction of overall storage capacity and need a system to scale, be flexible and efficient
- **Solution**: HPC-class storage, powered by Intel® Enterprise Edition for Lustre* software
- **Benefits**:
  - Openess, global namespace
  - Performance of upwards of 1 TB/s
  - Virtually unlimited file system and per file sizes, and management simplicity

*Other names and brands may be claimed as the property of others.

# HPC **Appliances** for Life Sciences



- **Challenge:** Experiment processing takes 7 days with current infrastructure. Delays treatment for sick patients
- **Solution:** Dell Next Generation Sequencing Appliance
  - Single Rack Solution; 9 Teraflops, Lustre File Storage; Intel SW tools
- **Benefits:** RNA-Seq processing reduced to **4 hour**
- Includes everything you need for NGS - compute, storage, software, networking, infrastructure, installation, deployment, training, service & support



*Actual placement in racks may vary.*

** 2-socket Intel(R) Xeon(R) CPU E5-2687W / 3.1 GHz

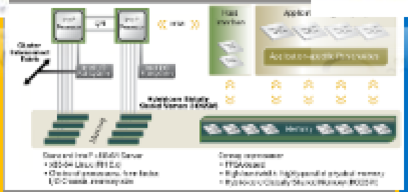
*Other names and brands may be claimed as the property of others.

*Other names and brands may be claimed as the property of others.

# Genomics & Clinical **Analytics Appliances**

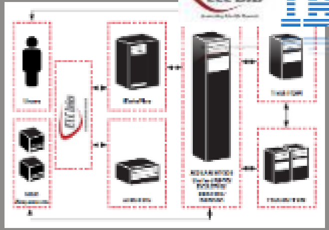
BIOTEAM
Enabling Science

CONVEY
COMPUTER



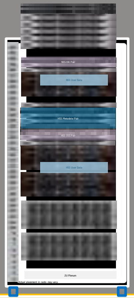
CLC bio

IBM



GENALICE
TECHNOLOGY FOR PEOPLE & SCIENCE

DELL



sgi

Knome

YOUR
LOGO
HERE

*Other names and brands may be claimed as the property of others.

# Let us all make Personalized Medicine mainstream by 2020 ..

- www.intel.com/healthcare/bigdata
- Paolo.Narvaez@intel.com