On the Centrality of HPC for taking NGS to the next frontier: Clinical application at scale

Monica Martinez-Canales, PhD
Principal Engineer
Big Data Solutions Group
monica.martinez-canales@intel.com

Stephen R. Wheat, PhD
Director, HPC Pursuits
HPC and Big Data Business Group
stephen.r.wheat@hpe.com

www.intel.com/healthcare/bigdata
HPC Alliance Accelerates Customer Value
HP and Intel partner to deliver innovation across the entire HPC Solution Framework

HPC Center of Excellence – Customer Experience Innovation
- Code Modernization
- Performance Optimization
- Engineering
- Benchmarking / POC
- Service Engagement
- Resource Utilization
- Business Outcomes
- Rapid Deployment
- Expert Community

HPC Solutions Framework – HP Innovation
- Industry Expertise
  - Geosciences (Oil & Gas)
  - Life Sciences
  - Financial Services
- Ecosystem Optimization
  - ANSYS
  - BIOVIA
  - Gaussian
  - Halliburton
  - Abaqus
  - Paradigm
  - Schlumberger
  - SIMULIA
  - Synopsys
- Solutions Design
  - Purpose built and optimized for HPC workloads
  - End-to-end Integration to simplify deployment and management
  - Uniquely designed for Density, Energy Efficiency, and Performance

HPC System Framework – Intel Innovation
- Intel® Xeon Phi™ coprocessors
- Intel® Omni-Path Architecture
- Intel® Lustre® Software
- Intel® Xeon® processors
- Value
  - Scalability & Resiliency
  - Power Efficiency
  - Price/performance

Customer Value
- Optimal Performance
- Exceptional Efficiency
- Rapid Deployment
## HP Apollo Systems Family

- **Purpose-Build HPC Solutions**

<table>
<thead>
<tr>
<th><strong>Apollo 8000</strong></th>
<th><strong>Apollo 6000</strong></th>
<th><strong>SL4500/Apollo 4000</strong></th>
<th><strong>Apollo 2000</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Supercomputing</td>
<td>Rack Scale HPC</td>
<td>Server Solutions Purpose Built for Big Data</td>
<td>Enterprise Bridge to Scale-Out Compute</td>
</tr>
<tr>
<td><strong>Purpose Built liquid-cooled supercomputing</strong>&lt;br&gt;Delivering the highest level of performance and density for the most demanding workloads</td>
<td><strong>Purpose built rack scale HPC solutions</strong>&lt;br&gt;Delivering shared infrastructure efficiency optimized for specialized workloads</td>
<td><strong>Purpose Built Server Storage Solutions</strong>&lt;br&gt;Delivering Industry leading density, efficiency, and price performance at Hyperscale for Big Data and Object Storage Solutions</td>
<td><strong>Purpose Built Density Optimized solutions</strong>&lt;br&gt;Delivering Hyperscale efficiency and performance for traditional enterprise and SME datacenters</td>
</tr>
<tr>
<td>Liquid Cooled</td>
<td>Rack Scale</td>
<td>Server Based Storage</td>
<td>Traditional Datacenters</td>
</tr>
</tbody>
</table>
What is the optimal platform for code modernization?

The world is going parallel – stick with sequential code and you will fall behind.

### Core Specifications

<table>
<thead>
<tr>
<th></th>
<th>Haswell</th>
<th>Intel® Xeon Phi™ x100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>18</td>
<td>61</td>
</tr>
<tr>
<td><strong>Threads/Core</strong></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td><strong>Vector Width</strong></td>
<td>256-bit</td>
<td>512-bit</td>
</tr>
<tr>
<td><strong>Peak Memory Bandwidth</strong></td>
<td>68 GB/s</td>
<td>352 GB/s</td>
</tr>
</tbody>
</table>

**Future Xeon Phi**
- Knights Landing
  - Cores: >60
  - Threads/Core: 4
  - Vector Width: 512-bit (x2)
  - Peak Memory Bandwidth: >400 GB/s

**Future Xeon**
- Knights Landing
  - Cores: TBA
  - Threads/Core: TBA
  - Vector Width: 512-bit
  - Peak Memory Bandwidth: TBA

Haswell coupled with Knights Corner provides the optimal scale to get performance now and get ready for Knights Landing and future Xeon.
Intel is Making an Enormous Pipeline Investment
Scaling R&D and touching next-gen via Intel® Parallel Computing Centers (IPCC)

*Other names and brands may be claimed as the property of others.
How can I get higher performance for my apps?

Lot of performance is being left on the table

We believe most codes are here

Modernization (i.e. parallelization and vectorization) of your code is the solution

VP = Vectorized & Parallelized
SP = Scalar & Parallelized
VS = Vectorized & Single-Threaded
SS = Scalar & Single-Threaded
Key TCO metrics using HP Apollo 6000 with 16C Xeon† & Xeon Phi†

Code modernized using Xeon Phi delivers compelling TCO metrics for FSI on Apollo 6000 servers.
The Oklahoma Health Prediction Center

Applying NGS-enabled solutions to a community-wide effort to combat cancer.

Tulsa’s Collaborative Health Network Platform has already demonstrated:

- Better Primary Care
- Better Coordination of Care
- Enhanced clinical services to communities with support, trust, and advocacy driven by:
  - Information Security Expertise
  - Bio-Informatics Expertise
  - Health Law Expertise
  - Molecular Biology Capacity
  - And, a community-first focus

Creating a “National Model” for Community-Supported City-wide Health Collaboration and Partnerships.
Intel’s Vision for Precision Medicine

Today: Many disparate data types, streams...

- Meds & labs
- Claims & transactions
- Personal data
- Clinical
- Genomics
- Patient experience

Future: connected clinically-actionable information

- Value-based Care
- Higher Quality at Reduced costs

Leading to better decisions

- Patient-centric approach
- Healthier population outcomes
Computing Innovation for Billions of People Worldwide

Health IT
- Performance‡
- Security
- Manageability

Powering the Health Workforce
- Mobility
- Datacenter
- Cloud
- Health information exchange
- Sensors and Wearables

Healthcare Solutions
- Intel®-powered medical imaging & devices
- Security software (McAfee, Wind River)
- Analytics (Intel® Math Kernel Library, Intel® Analytics Toolkit, Lustre, Cloudera)

Life Sciences
- Extreme computing for big data
- Open, interoperable clouds
- Appliances
- Code optimization

Ecosystem Support
- People-centered R&D
- Policy
- Standards

* Other names and brands may be claimed as the property of others.
© 2015 Intel Corporation
Intel in Health & Life Sciences

Big Data and High Performance Computing

Intel® Xeon® / Intel® Xeon Phi™ • Rack Scale Architecture • Integrated Fabric • Software Defined Network • Solid-State Drives • Workload Optimization

Hadoop® • Lustre® • Cloudbursting • Virtualization • Encryption • Cloud-based Analytics

At the Intersection of Transformative Forces

Enabling extreme-scale computing on massive data sets

Helping enterprises build open, interoperable clouds

Contributing code and fostering the ecosystem

Intel® Architecture is present from sequencers to appliances to high-performance computing cloud.

*Other names and brands may be claimed as the property of others.
Trends & Challenges in Life Sciences

Big Data in Life Sciences

• Sequencer advances – 4x data in 50% less time .5TB/device/day
• 4D molecular imaging produces 2TB/device/day

Burdens of Data Management

• Store, manage, share, ingest and move PBs of research & clinical data
• Need to reliably 'snapshot' pipelines with archive to tiered storage

Innovation Drives Change

• Rapid iteration of algorithms far outpace IT, requiring flexibility, agility
• Most applications do not fully leverage available infrastructure

Converged Infrastructure

• Workloads converging between local and cloud-based HPC/Big Data
• Advanced orchestration required to maximize throughput & efficiency
Optimizing Top Applications and Pipelines
Intel working with industry experts worldwide

- Genomics, Molecular Dynamics and Molecular Imaging applications targeting both Intel® Xeon® processors and Xeon® Phi™ coprocessors
- Fine- and coarse-grained optimization at the node and cluster level
- Work with code authors to release optimizations, disseminate best practices

**Genomics**
- ABYSS*
- BLAST*
- Bowtie*
- TopHat*
- Cufflinks*
- BWA*
- GATK*
- Picard*
- SAMtools*
- MPI-HMMER*
- Velvet*

**Molecular Dynamics**
- AMBER*
- CAS-Soft Sphere*
- CAS-IPE*
- CP2K*
- CPMD*
- DLPOLY*
- GAMESS*
- Gaussian*
- GROMACS*
- LAMMPS*
- NAMD*
- NWChem*
- Quantum Espresso*
- VASP*

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

[www.intel.com/healthcare/optimizecode]
Intel Workflow Profiler
coarse-grained profiling of long-running workflows

- Automates data collection & charts from standard Linux OS tools
- Quickly identify CPU, Memory & I/O Constraints
- Pareto analysis of hotspots in user-defined steps of workflow
- Target areas which will benefit from newer algorithms and technologies

Open Source Distribution:
https://01.org/workflow-profiler
The future of precision medicine

1 patient visit

2 causal genes identified & molecular pathways determined

3 targeted therapeutics & companion diagnostics; treatment begins in earnest

**PRIMARY ANALYSIS**

- Individual
  - 1 to 4 days

- Multiple individuals
  - weeks
  - hours

**SECONDARY ANALYSIS, DNA/RNA PIPELINE + MORE**

- Multiple sample analysis starts here
  - Joint genotyping
  - Variant store
  - Pop/Dis study

**PRECISION MEDICINE**

- Predicted Actionable Variants
- Data Driven Association
- Clinically Actionable Variants
- Clinical trials
- Data curation
- Knowledge Database

**PRECISION MEDICINE**

- Targeted therapeutics & companion diagnostics
- Treatment begins in earnest

**1 to 4 days**

**weeks hours**

**months days**

Precision medicine is the standard of care, integral to wellness by 2020
Applying Precision Treatments for Pediatric Cancer

**Neuroblastoma:**
From 15 days to less than 4 hours for personalized treatment

Patient / Physician diagnosis, treatment, ongoing management

Tumor Sample

Complete molecular characterization of the diseased tumor

Analytical tool for mapping patient data against database for recommended treatment

Integration of scientific & clinical evidence for future research

Treatments with a more reasonable chance of a cure

Minimizing trial and errors

Understanding the individual disease

Accelerating targeted treatment options

Creating platform to scale to 100k+ patients

*Some names and brands may be claimed as the property of others.*
Genomics Data Processing Pipeline

NextGen Sequencing Computing & Storage Environment

Lustre

- 288 TB
- 9.6 GB/sec read
- 7.8 GB/sec write

FDR IB

CIFS to Lustre Gateway

Ethernet Network 10Gb

Clinical Cluster

- 1536 Cores
- FDR Infiniband
- 10 Gb Ethernet
- 1 Gb Ethernet
- 2 x 768 GB RAM SMP Nodes
- 30 TB NFS storage
- 600 TB NAS storage

Images

Intensities

Bases With Quality Scores

Aligned Reads

Mutations

2-7 TB

200 - 300 GB

50 - 100 GB

100 - 200 GB

1 - 2 GB

Register Cluster locations and edges. Select representative pixels

Process intensities to call base. Calculate quality score for base call

Process reference sequence and align reads. Identify no-match and multi-match

Identify SNP, insertions, deletions, rearrangements, translocations, haplotypes etc.

*Some names and brands may be claimed as the property of others.*
An Idea for Clinical Decision Support Applications Combining Clinical, Genetic/Genome, and Family Health History Data

Goal:

Promote widespread use of clinical decision support that will help clinicians/counselors in assessing risk and assist genetic counselors in ordering genetic tests.

Build a scalable CDS that leverages standardized data that includes:

- Family Health History
- Clinical data and Screening
- Genomic Data

Solution Considerations:

- Be agnostic to data collection tools.
- Be scalable to different clinical domains (grow beyond Breast Cancer) and other healthcare institutions.
- Be standards based where they exist
- Work across all EHRs, but start with Cerner
- Leverage Intel technologies (infrastructure, Intel Data Platform, etc.)
- Be flexible to incorporate other data sources (e.g. imaging data, personal device data)
Sample Clinical Workflow with Clinical Decision Support

1. Patient has encounter, fills out initial screening app
2. Patient fills out detailed FHH and medical Hx app
3. If patient is high risk, schedule genetic counselor
4. Screen patient for further testing
5. Case review, order genetic tests for patient and optionally family
6. Order placed with relevant clinical info
7. Return narrative, codified genomic result

Risk Screening Applications | Dynamic Family Health History app | Risk assessment app | Diagnosis and Treatment Recommendation | Genomic Predictive Models w/ machine learning | Genomic inference Engine

CDS

Risk Screening Data | Family Health History | Genomic Repository | CDR | EHR

Solution Considerations

Utilize the following where appropriate:

1. Health Services Platform (HSPG), HealtheDecision, Open CDS
2. Intel Data Platform for Machine Learning, Graph Analytics, Mining
3. HL7 standards, FHIR + SMART Apps for clinician facing applications
## Sample Analytics

<table>
<thead>
<tr>
<th>Areas</th>
<th>Benefits, solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilization and Treatment Analysis</td>
<td>Combine trends with individual treatment analysis</td>
</tr>
<tr>
<td>Treatment Effectiveness</td>
<td>Build large-scale treatment effectiveness monitoring</td>
</tr>
<tr>
<td>Diagnosis and Treatment correlation</td>
<td>Discover diagnosis/treatment connections</td>
</tr>
<tr>
<td>Managed Care Optimization</td>
<td>Optimize resources for managed care</td>
</tr>
<tr>
<td>Diagnosis Treatment and Trends and Predictions</td>
<td>Determine overall trends, but put them at the disposal of individual diagnosticians</td>
</tr>
<tr>
<td>Drug Utilization and Expense Prediction</td>
<td>Build dynamic precise drug utilization prediction models</td>
</tr>
<tr>
<td>Treatment and Outcomes Analysis and Optimization</td>
<td>Predict treatment prognosis, optimize based on individual’s complete picture</td>
</tr>
<tr>
<td>Demand Forecasting</td>
<td>Better demand preparedness</td>
</tr>
<tr>
<td>Price Analysis and Determination</td>
<td>Optimize quality and revenue through price monitoring</td>
</tr>
<tr>
<td>Epidemiology Research</td>
<td>Discover trends by analyzing data from disperse sources</td>
</tr>
<tr>
<td>Provider Ratings and Benchmarking</td>
<td>Use all source data to benchmark and monitor providers</td>
</tr>
<tr>
<td>Patient History and Digital Records Archiving and Analysis</td>
<td>Combine patient history records from disparate sources, greatly improve the quality of patient care</td>
</tr>
<tr>
<td>Contract Optimization</td>
<td>Optimize contract resource utilization</td>
</tr>
</tbody>
</table>

Imagine what is possible

We are working with industry experts like you –

- to overcome systemic challenges,
- maximize the use of available infrastructure, and
- drive innovation through open standards and platforms.

These technologies facilitate more accurate science, enabling scientists & clinicians to imagine what is possible, instead of framing their research in terms of constraints.


www.intel.com/healthcare/bigdata
www.intel.com/healthcare/optimizecode
Intel Across Healthcare: Learn More

Mobile Healthcare
Care Coordination
Secure Cloud
Devices and Imaging

Performance
Big Data
Policy and Standards
Aging

www.intel.com/healthcare
www.intel.com/healthcare/optimizecode
Backup
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 quadratically 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, etc.*