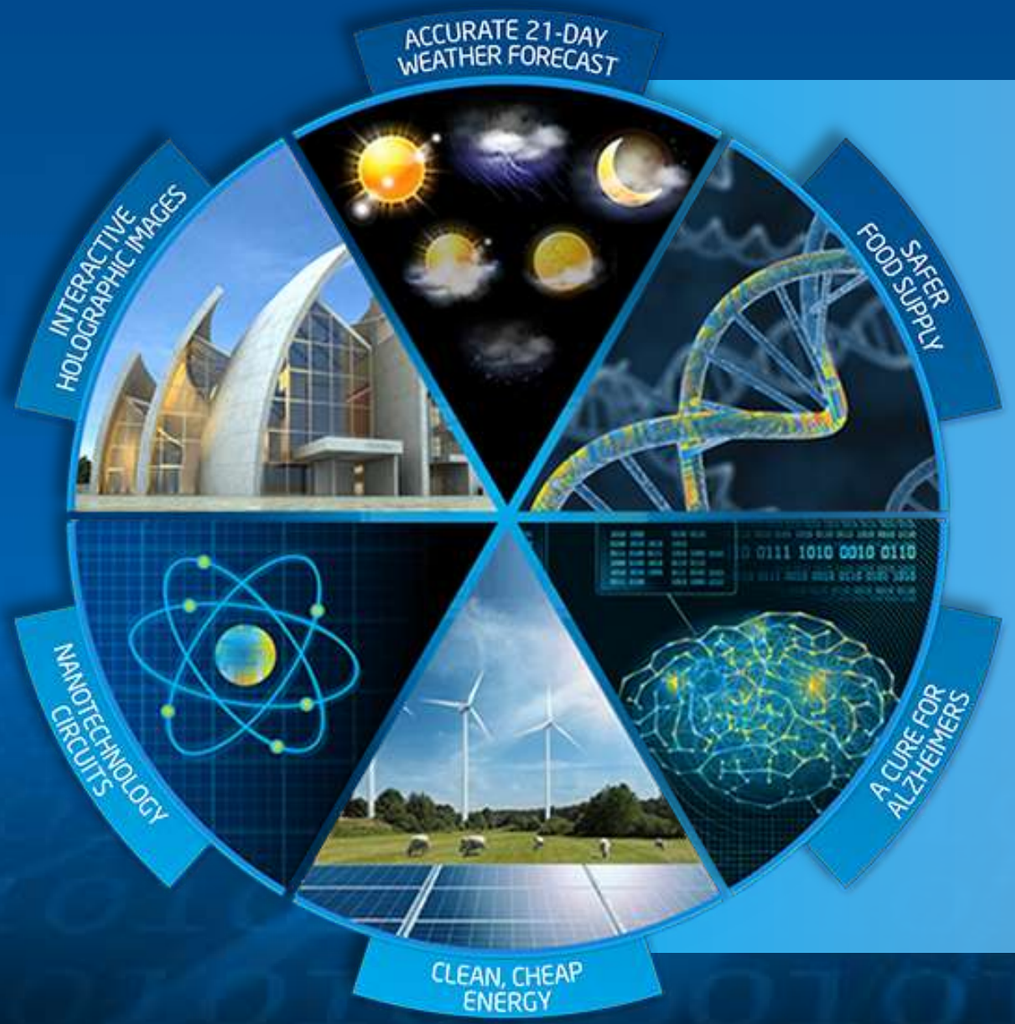
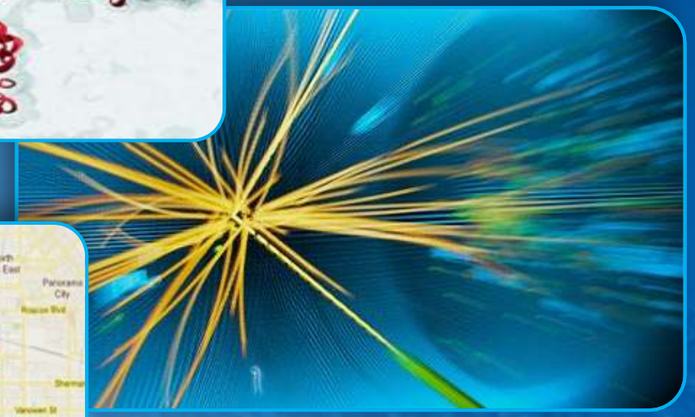
Big Data Use Cases: “The Size of the Data does not define Big Data”

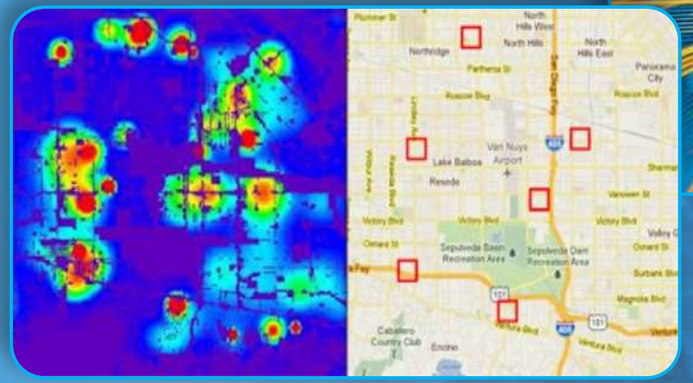
John A. Hengeveld
Director of HPC Marketing,

Technical Computing Group,
Intel

Uncharted Territory on Path to Discovery In Science and Engineering







We Struggle to Define Big Data..

	Traditional View of Data	Big Data
Volume	Gigabytes to Terabytes	Petabytes and beyond But does THIS MATTER?
Velocity	Occasional Batch – Complex Event Processing	Real-Time Data Analytics
Variety	Centralized, Structured i.e. Database	Distributed, Unstructured Multi-format
Value	The Data itself is the end (a lot of HADOOP stuff)	Insight Derived from Data (Find the Cancer Cure)

Not Unlike we Struggle to Define HPC!

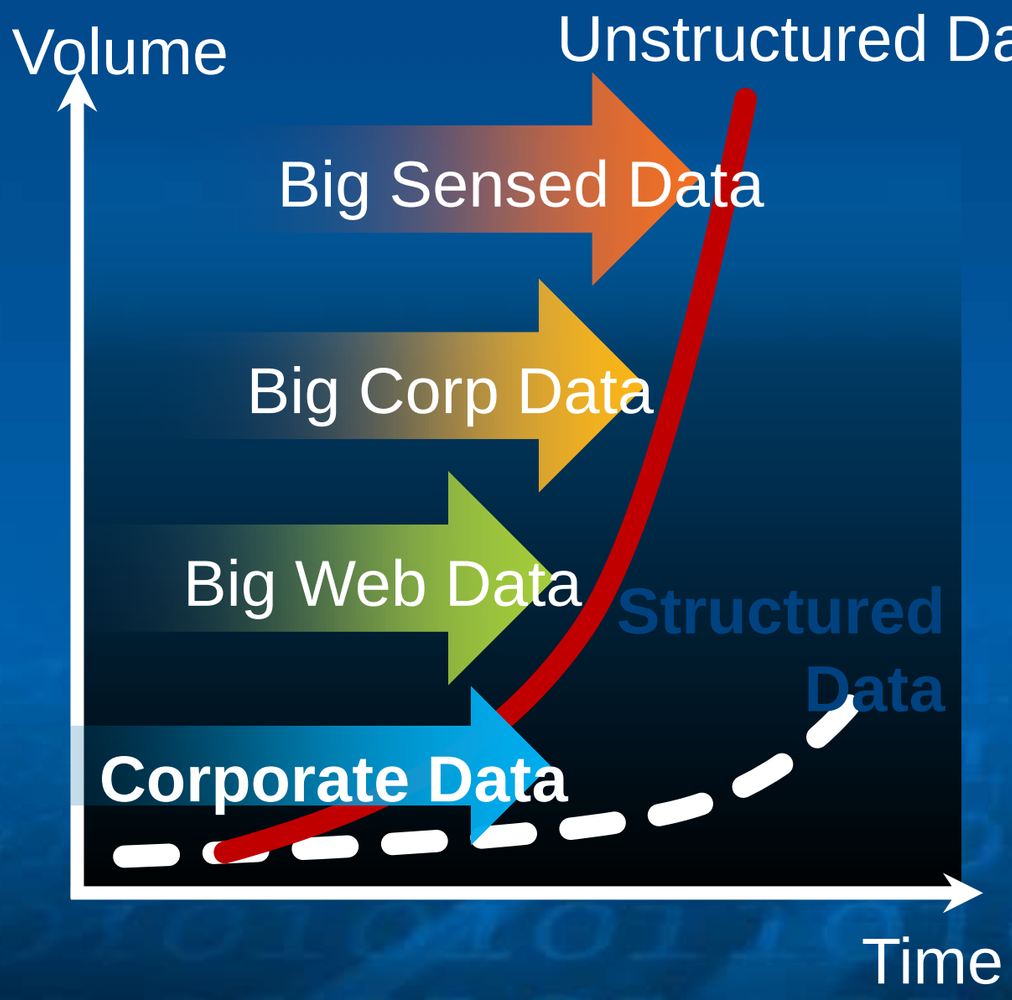
When is a computation requirement “high performance”?

The canonical definitions are problematic…
because they ignore temporal contextuality

HPC is defined by not only the processing…
But the fact that processing is connected together
to work on a single problem seamlessly.

…. Interconnection (fabric) is essential to HPC’s nature.

What will be big tomorrow?



Does Big Data having anything to do with the size or structure of the data?

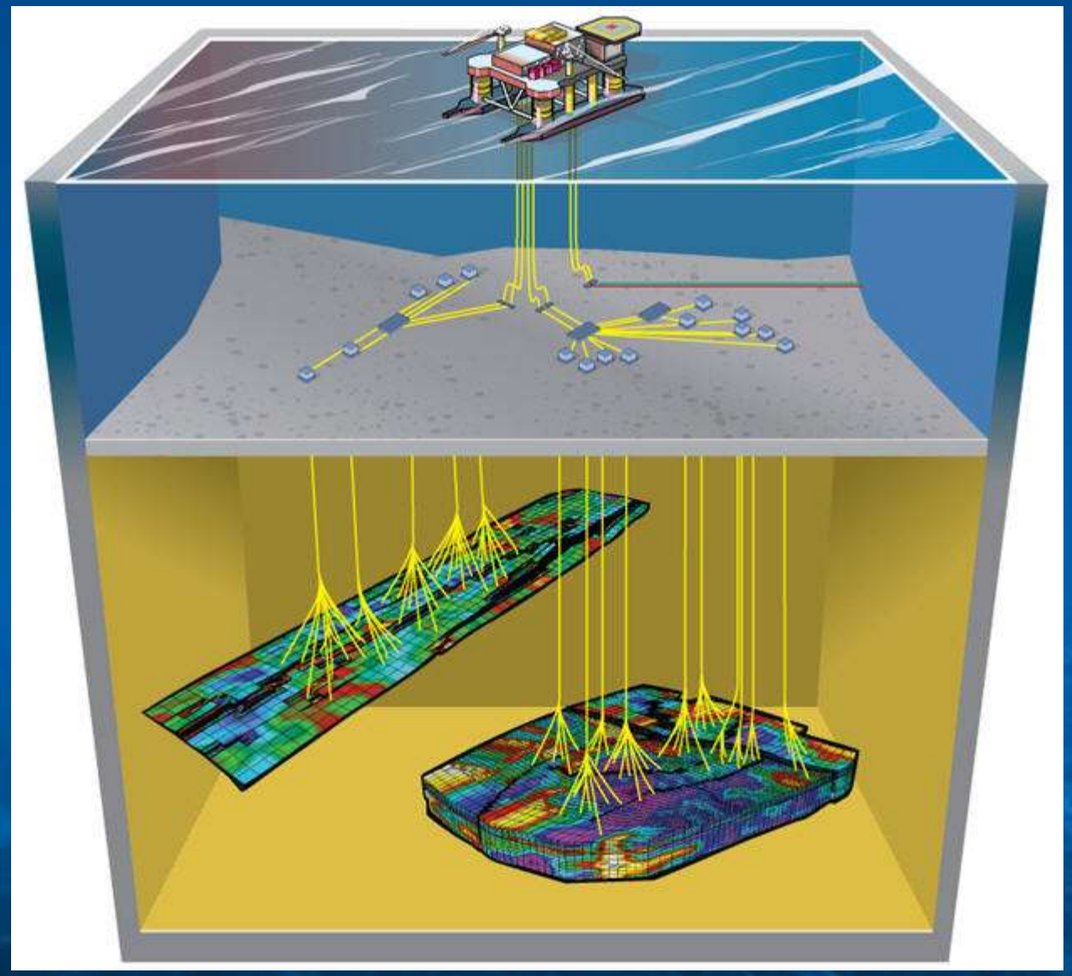
We struggle because we try to define The opportunity from an irrelevant attribute of the storage architecture!

More important than HOW BIG..

Is the impact HOW BIG has to WHERE the data is.. And how we reach it. (interconnect is essential to nature)

John's Definition of Big Data

"Big Data" is a class of insight opportunities where the data brought to the problem is multisourced, distributed or of such a scale that significant processing capability is required to derive insight from it.



(An Oil field is defined by how you reach the oil)

HPC is Essential to Big Data

Four “Use Case Classes”:

Data Aggregation:

Find a specific group of records that comply with a set of requirements.

Filesystem Application

Data Analytics:

Computation of common characteristics or key regression characteristics in data aggregated for the analysis.

HPC + Filesystem

Data Visualization:

Construct a Visual representation of a computationally-derived model of data aggregated for analysis.

HPC + Filesystem + Visual Analytics

Interactive Visualization and Simulation

Command adaptive analysis through manipulation and simulation of aggregated data. Insight from Human Derived Optimization Points.

HPC + Filesystem + Visual Analytics + Interactive Design

Big Data – A Foundation For Delivering Big Value

Intel Leverages the Power of Big Data

Chip Design Validation:

Cut Product Time to Market by 25%

Faster analysis process for validating results

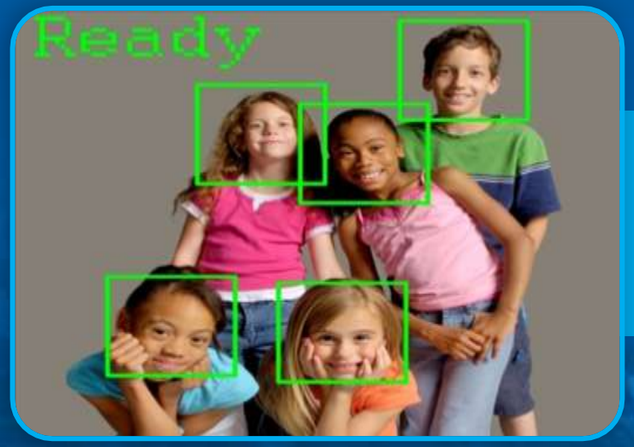
Streamlined debug process through analysis of large volumes of historical test data

HPC Insights from Big Data

Better Medical Therapies

Fastest Route Through Traffic

Urban Planning and Simulation



Improved Security (Scary guys)

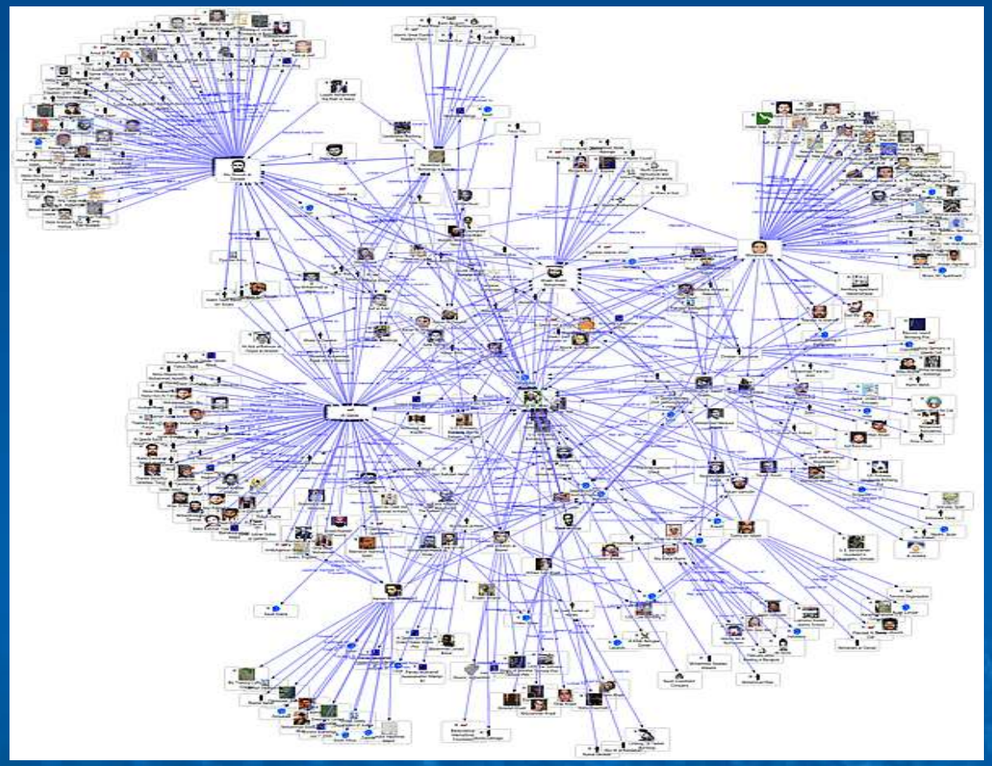
intel

Graph Data

Analytic problems with Data with a graph structure:

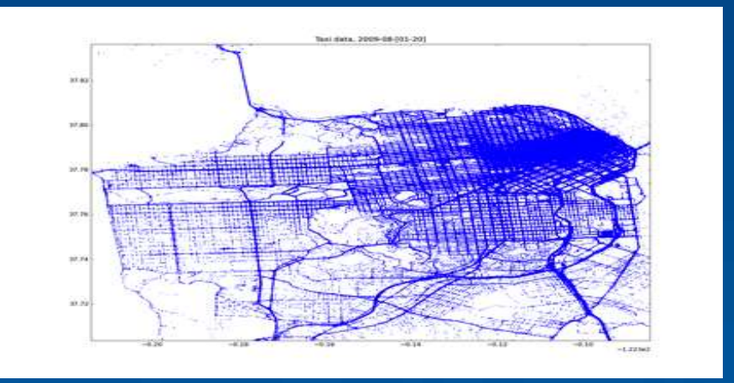
- Match Genomes of tumors
- Find the bomber in the haystack?

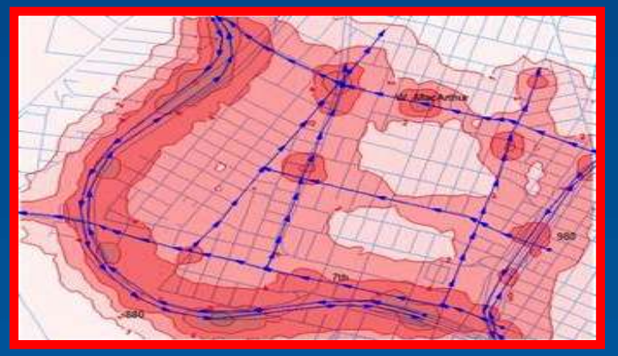
Some of these graphs are..
HUGE.



Data for the compute is ~never where you want it from a computation efficiency point of view.
Rarely here, rarely low latency to access.

What can you do with the data?





Crowdsourcing + physical modeling + sensing + data assimilation
to produce:

From Alex Bayen, UCB

Intel's view of the world - Enabling Discovery through Big Data

Example: Life Sciences Applications

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

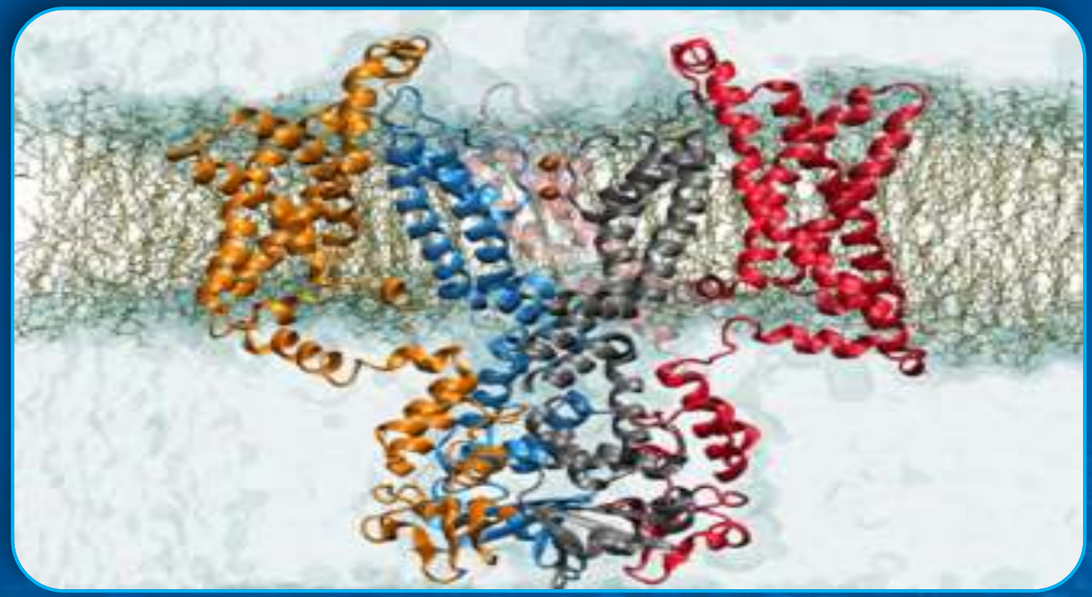
Intel is working with app/code developer, upstream code changes

Some algorithms have quadratic scaling with the size of the problem, so are the needs for larger working data sets resulting in compute capacity constraints

Technology innovations in compute, storage, and data fluidity

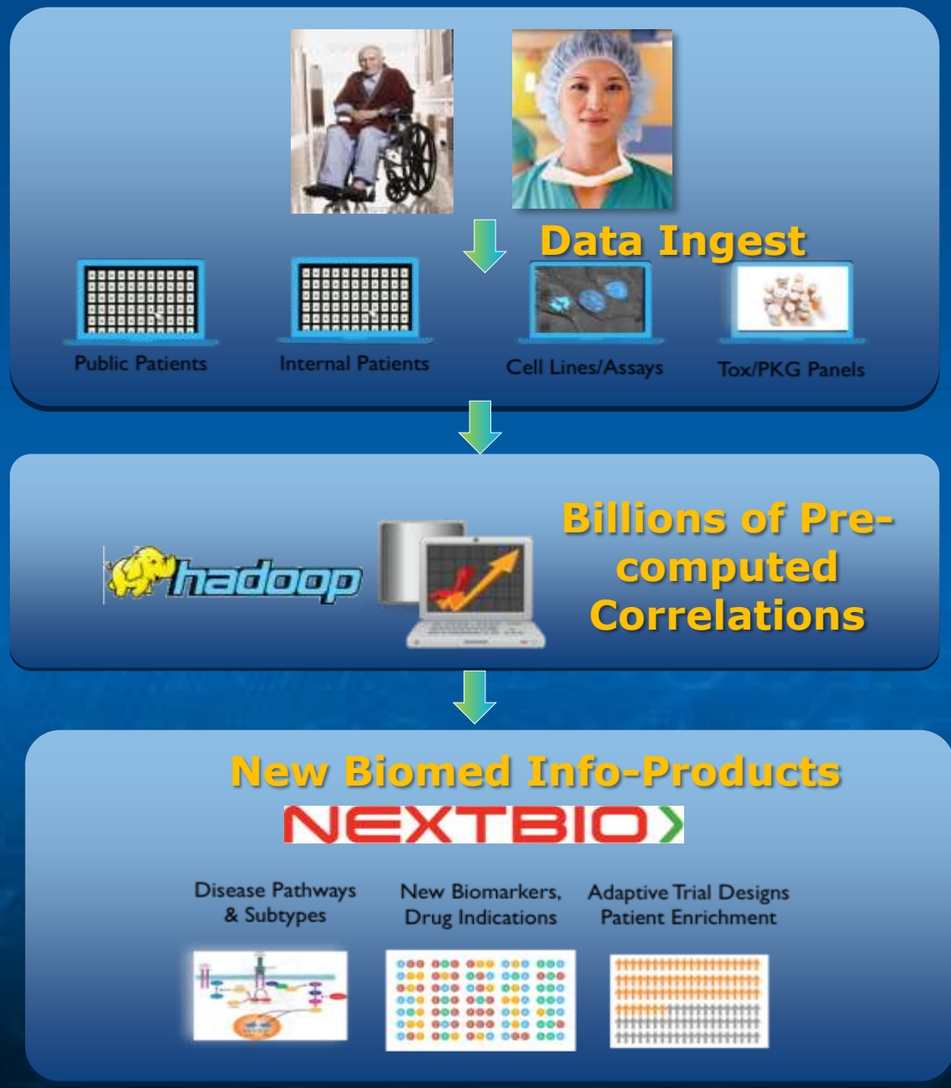
Databases are distributed, data is siloed and will likely stay that way

Tools like hadoop, lustre, graphlab, in-memory etc.



Big Data needs Balanced Infrastructure

Genomics for Translational Medicine using Intel® Distribution for Apache Hadoop - NextBio



- **Challenge:** Derive new value added patient discovery services while bringing down genome processing costs
- **Solution:** Dynamically partition & scale correlation of patient data to all public data using Hadoop and Hbase
- **Benefits: Contributes to 800x reduction in cost to process 4 Million genome variants**

HPC Compute Technology and Big Data

- HPC transforms Big Data From Batch orientation to Mission critical and real time orientation
- Data moves from target of static search to constructed in a pipeline of intelligent agents
- This requires shift form simple storage architecture to Greater Intelligence in the Storage System
- Cluster computation for rapid/accurate real time answers to real time data
- Fabric Performance becomes Essential to Insight delivery

INSIGHT

Big Data Architecture Framework changing over next 2 years

Data Access
User Authentication

Data Aggregation

Data Analytics

Modeling and Simulation

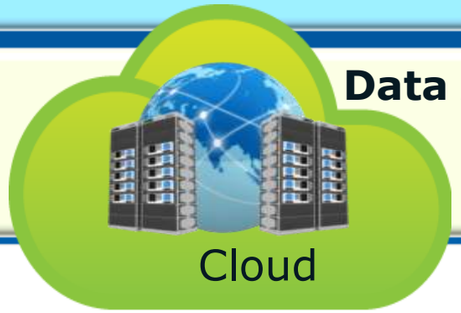
Visualization

Data as a Services

Distributed
Event, Message

Virtual
Real-Time, Cached, Federated

Persistence
EDW, Marts



Data Aggregation, Ingestion, Integration and Processing Services

Distributed High Performance Data Processing
Hadoop* MapReduce

Integration Tools

Data Location, Access, Storage and Distribution

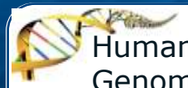
NAS - SAS and Distributed Storage

10GBe Fast Fabric

Databases DBMS / NoSQL

MPP Databases
DW Appliances

Data Sources


Human Genome & Drug Discovery

Diagnostic Images

Surveillance and Medical Device Streaming Data

Log Files

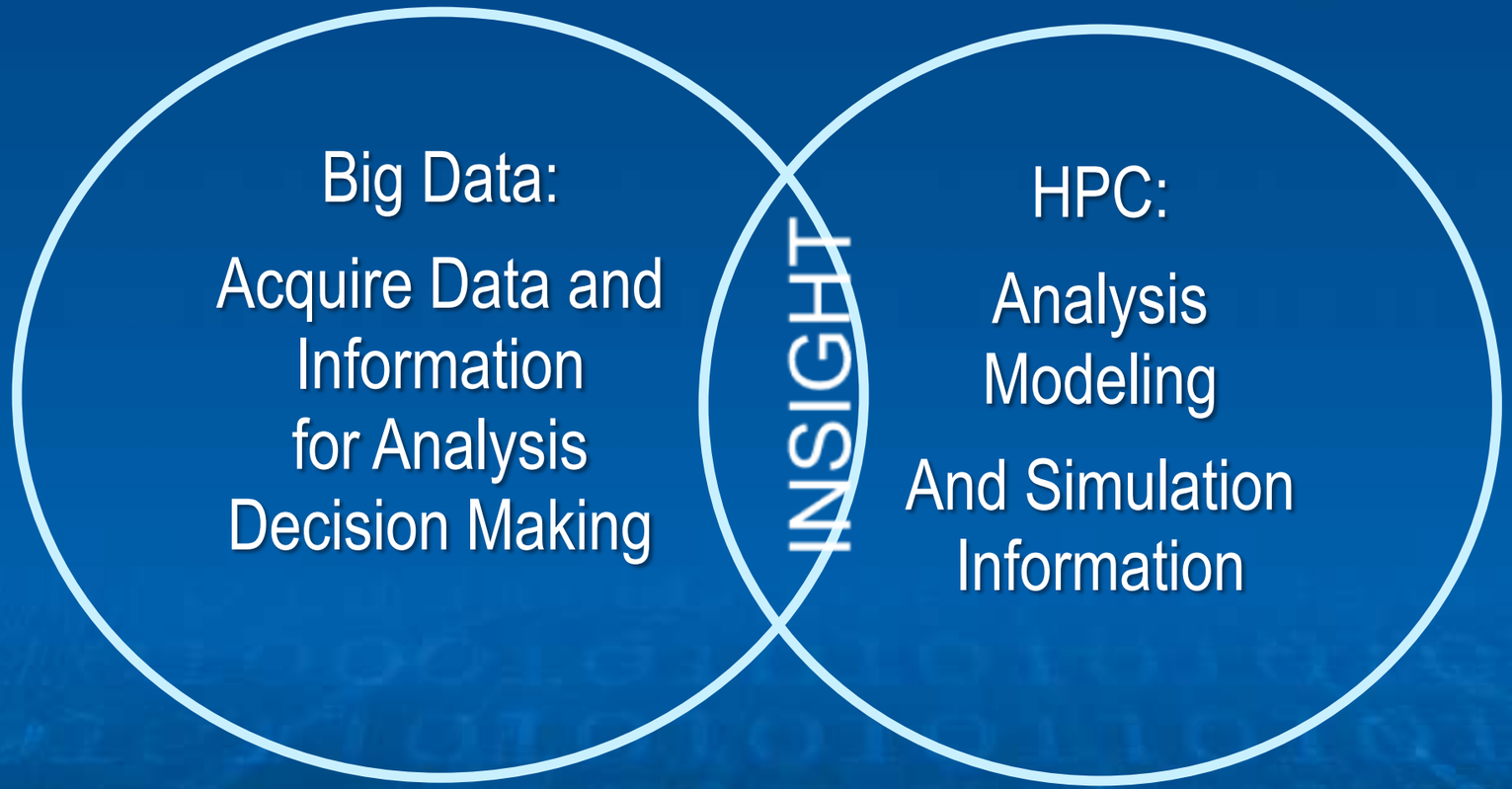
Social Media

Medical Records

GIS

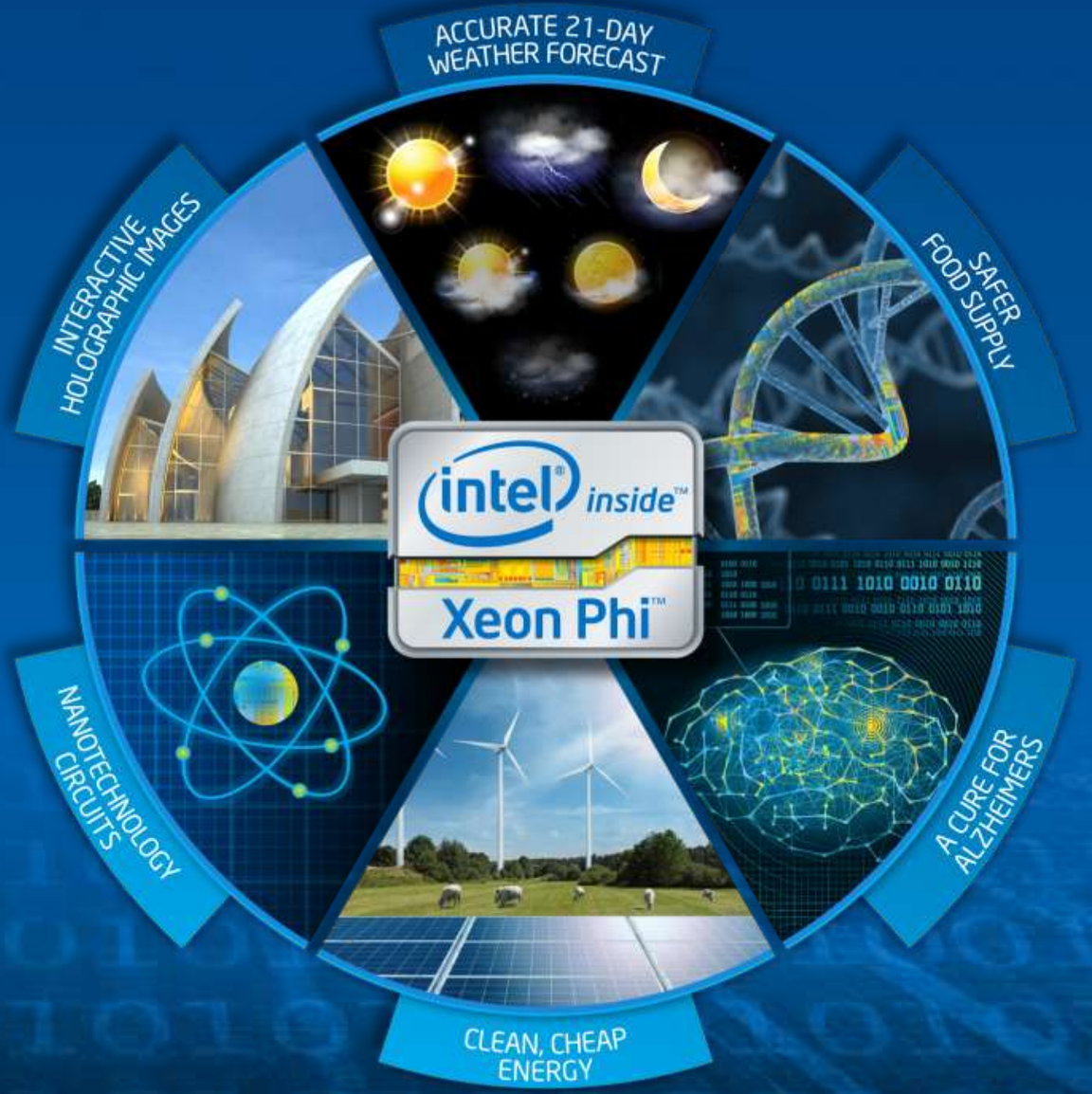
IS360.. Internally generated data.

Computation of Insight from Big Data



Parallelism is Your
Path to the Future

Intel is ..more than ever.. Your Roadmap



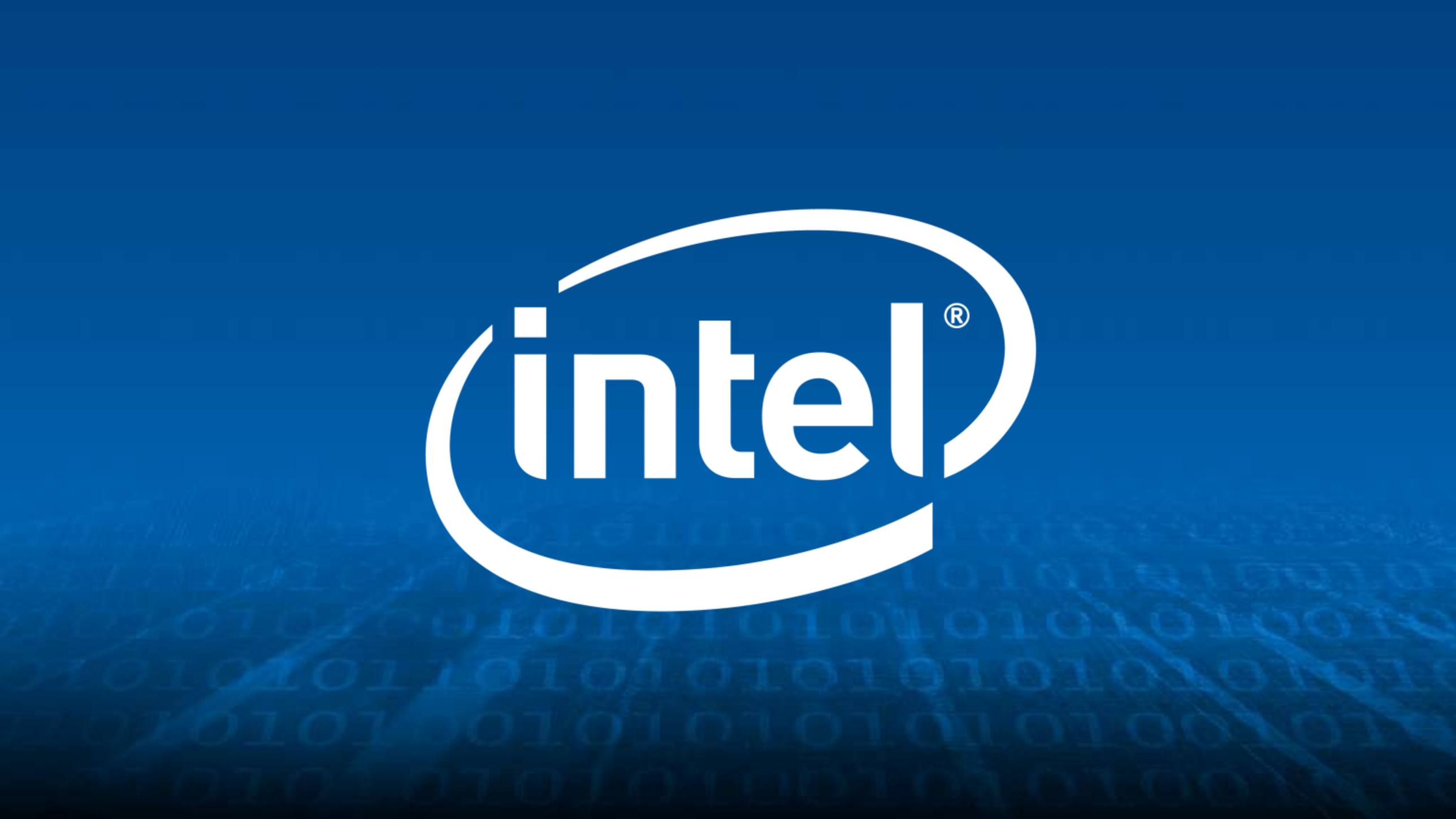
Model courtesy of Martin Lubich
www.loramel.net

HPC's Value for Big Data

"HPC is key for extracting value from Big Data...HPC will move Big Data from the target of static searches to a pipeline of intelligent agents."

John Hengeveld, Intel
(Intel Developer Forum, September 2012)

© 2012 IDC


intel®

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the fourth quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “believes,” “seeks,” “estimates,” “may,” “will,” “should” and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel’s actual results, and variances from Intel’s current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company’s expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions, including supply constraints and other disruptions affecting customers; customer acceptance of Intel’s and competitors’ products; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel’s response to such actions; and Intel’s ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel’s results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel’s ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel’s results is included in Intel’s SEC filings, including the company’s most recent Form 10-Q, Form 10-K and earnings release.

Rev. 10/16/12

Legal Information

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent Earnings Release and our most recent Form 10-Q or 10-K filing for more information on the risk factors that could cause actual results to differ.

If we use any non-GAAP financial measures during the presentations, you will find on our website, intc.com, the required reconciliation to the most directly comparable GAAP financial measure.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to:
http://www.intel.com/products/processor_number

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit
http://www.intel.com/go/virtualization

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security

Intel, Intel Xeon, Intel Atom, Intel Xeon Phi, Intel Itanium, the Intel Itanium logo, the Intel Xeon Phi logo, the Intel Xeon logo and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other names and brands may be claimed as the property of others.
Copyright © 2013, Intel Corporation. All rights reserved.

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to:

Copyright © 2012 , Intel Corporation. All rights reserved. Intel, the Intel logo, Xeon, Xeon Phi, Xeon Phi logo, Core, VTune, and Cilk are trademarks of Intel Corporation in the U.S. and other countries. *Other names and brands may be claimed as the property of others.

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Copyright© 2012, Intel Corporation. All rights reserved.
*Other brands and names are the property of their respective owners.

intel.com/software/products