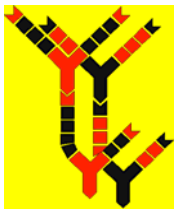


HPC 2014 High Performance Computing

FROM clouds and BIG DATA to EXASCALE AND BEYOND

An International Advanced Workshop
July 7 – 11, 2014, Cetraro, Italy

Session III – Emerging Systems and Solutions

The Fusion of Supercomputing and Big Data: The Role of Global Memory Architectures in Future Large Scale Data Analytics

Bill Blake
Senior VP and CTO, CRAY, USA

COMPUTE | STORE | ANALYZE

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Is Highly Scalable Computing Facing a Branch in the Road Ahead?

With one path leading to Exascale supercomputers delivering billion-way parallel computing for the highest capability running a single application?

And another path leading to Hyperscale with millions of servers and billions of cores in the cloud delivering high capacity for many applications?

This presentation will explore the technology and architectural trends facing system and application developers and portray Cray's system design efforts as the way to deliver the fusion of Supercomputing and High Performance Data Analytics that will achieve a "both/and" capability bringing HPC benefits to a larger community

System Architecture Differences…

Exascale Supercomputing

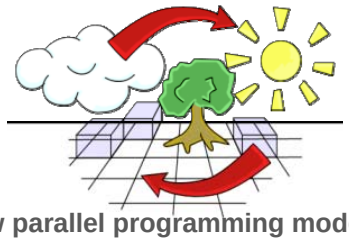
- Scalable computing w/high BW, low-latency, Global Mem Architectures
- Highly integrated processor-memory-interconnect & network storage
- Ability to apply all compute power to one highly parallel application
- Low data movement – load the “mesh” into memory and compute
- Move data for loading, defensive check-pointing or archiving
- “tennis court sized” systems that consume <20 MWatt

Hyperscale Computing – aka the Cloud

- Distributed computing at largest scale
- Divide-and-conquer approaches using Service Oriented Architectures
- Ability to apply compute power to many apps with multi-tenancy
- High data movement-- Scan/Sort/Stream all the data all the time
- Lowest cost processor-memory-interconnect & local storage
- “Warehouse sized” systems that collectively consume >260 MWatt

Need to Create a "Virtuous Cycle"

Cloud provides new distributed programming models that utilize "divide and conquer" approaches with massive scale-out Service Oriented Architectures using local storage and low cost hardware, and new data analytics algorithms where data scientists claim "**the larger the data the simpler the algorithm**"



HPC provides new parallel programming models that utilize highly scalable Global Memory Architectures supported by highest BW, lowest latency interconnects, with powerful algorithms for high fidelity modeling and simulation using highly iterative processing of both capability and capacity workloads that increasingly support data assimilation (from sensors)

COMPUTE STORE ANALYZE

Image courtesy of University of Michigan –Atmospherics Dynamics Modeling Group

©

The Fusion of Supercomputing with Large Scale Data Analytics

Key Trends

- The cost per FLOP is dropping to near zero!
- Analytic application needs for memory size and IO performance are insatiable!
- New levels of memory and storage hierarchy are emerging
- The MPP interconnect must support atomic Global Memory Operations
- Future systems are being influenced by the converged infrastructure of compute & storage
- Converged infrastructure of Hyperscale demands management capability of O(10000) nodes

The Cost per FLOP is Dropping Like a Spent Rocket!

0.000000001 cents per FLOP

TALE OF THE TAPE: SUPERCOMPUTER VS. GAME CONSOLE

	SANDIA LAB'S ASCI RED	SONY PLAYSTATION 3
DATE OF ORIGIN	1997	2006
PEAK PERFORMANCE	1.8 teraflops	1.8 teraflops*
PHYSICAL SIZE	150 square meters	0.08 square meter
POWER CONSUMPTION	800 000 watts	<200 watts

* for GPU; CPU adds another 0.2 teraflops

SANDIA LAB'S RED STORM XT	NVIDIA-powered Game Console
2005	2019
40 teraflops	40 teraflops[2]
280 square meters	0.08 square meter
1 000 000 watts	<200 watts

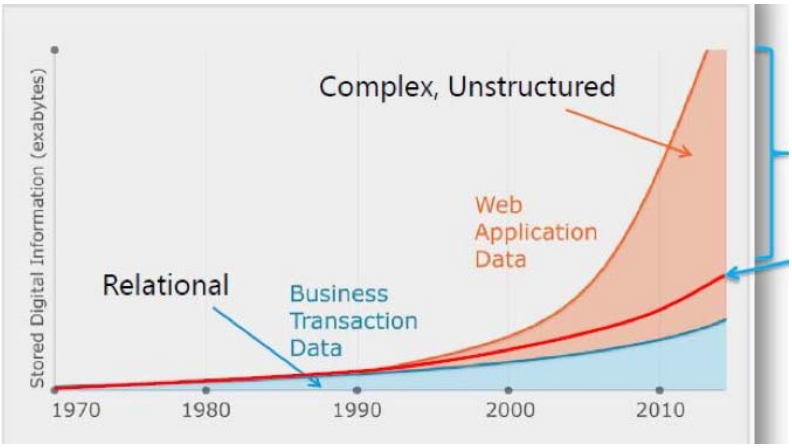
COMPUTE | STORE | ANALYZE

7/8/2014 [2] for GPU; CPU adds another ~10 teraflops (10nm process) 7

this sort of comparison provides a way of illustrating what has changed (peak floating point performance) and what has not changed by nearly as much (memory, interconnect and storage bandwidths). Continue the extrapolation and we get a machine with peak performance of an Exaflop and reasonable power consumption (not 20MW perhaps, but within what can be supplied to a building). The trouble is, it is almost useless. One way of looking at the significance of fusing HPC and fast data is that the reemphasis on data movement is what will broaden the applicability of supercomputers again - perhaps not to their peak of the late 80s, but at least heading back in the right direction.

Big Data Means New Kinds of Data

EMC estimates that by 2020 there will be 40,000 Exabytes of data created, although the majority of that data will not be created by humans but sensors

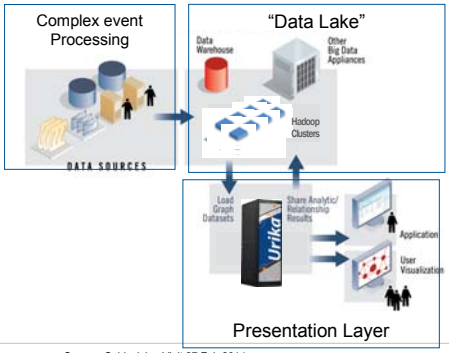


Source: IDC White Paper sponsored by EMC May 2009

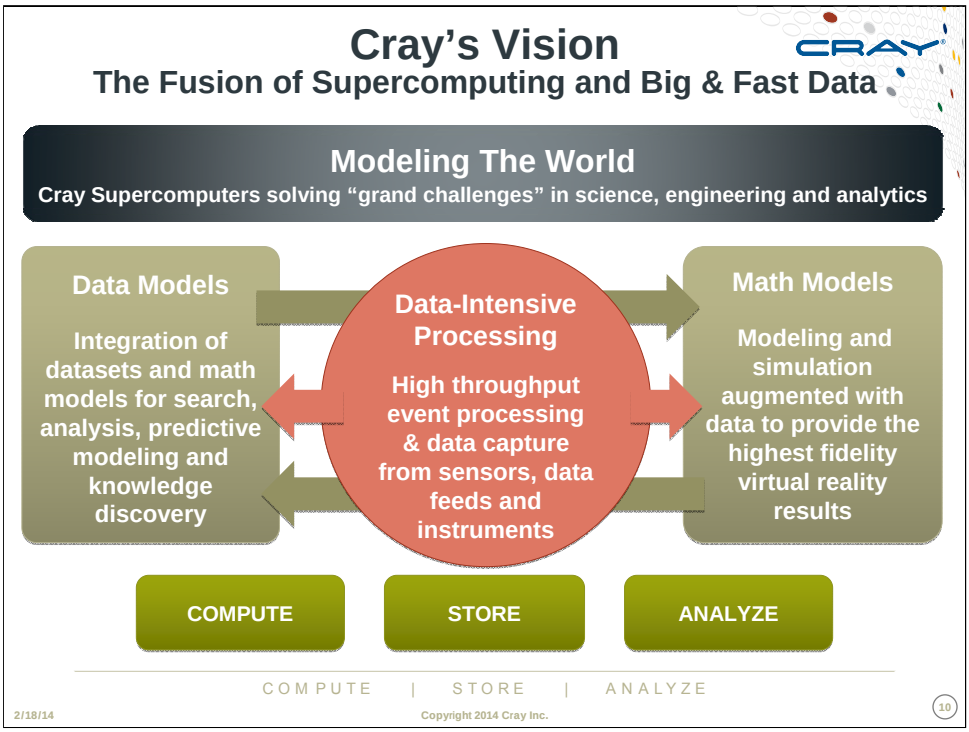
Software-led Infrastructure of the Enterprise

There are substantial changes in the technology used by Datacenters that Cray will adopt to support this business (very different from Exascale!)

- Need to focus on applications and workloads and move to converged infrastructure
- Appliances are first stage of converged infrastructure (compute + storage)



Source: Cablevision Visit 27 Feb 2014


Cray's Vision
The Fusion of Supercomputing and Big & Fast Data
Modeling The World
Cray Supercomputers solving "grand challenges" in science, engineering and analytics
Data Models
Integration of datasets and math models for search, analysis, predictive modeling and knowledge discovery
Data-Intensive Processing
High throughput event processing & data capture from sensors, data feeds and instruments
Math Models
Modeling and simulation augmented with data to provide the highest fidelity virtual reality results
COMPUTE
STORE
ANALYZE
COMPUTE | STORE | ANALYZE
2/18/14
Copyright 2014 Cray Inc.
10

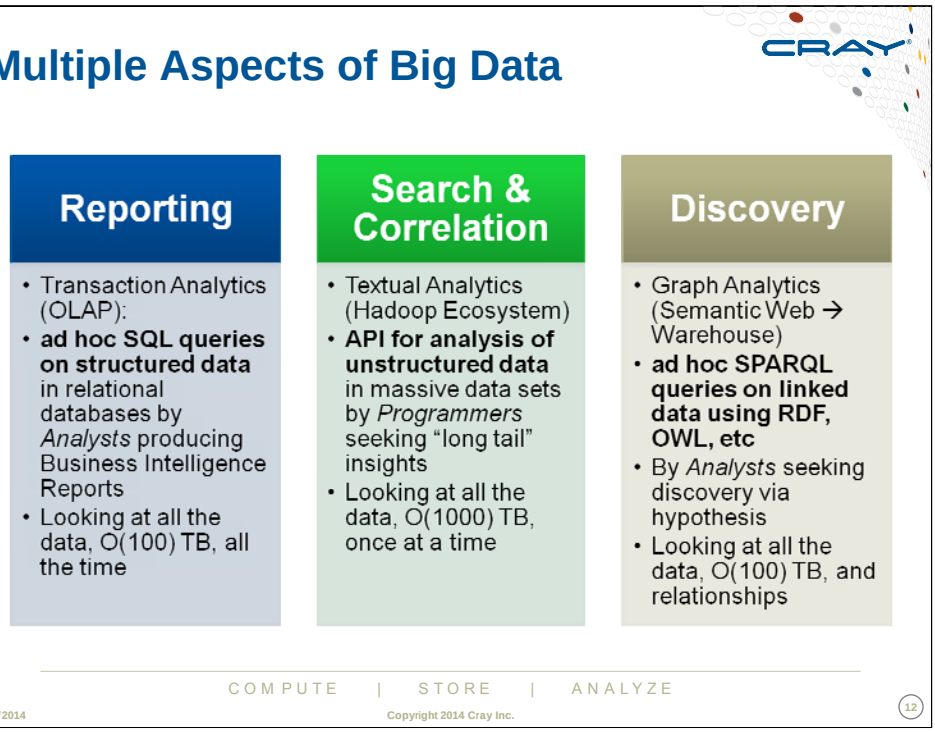
The "Big Data" Challenge

Supercomputing minimizes data movement – "data movement" is highly restricted for defensive or resiliency such as loading, check pointing or archiving.
Programming model is imperative (C++/Fortran + MPI) with focus on the details of **how** parallel programming is done

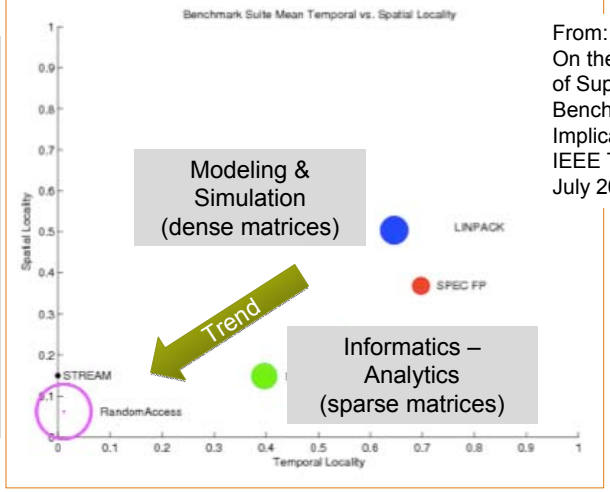
Data-intensive computing is all about data movement - scanning, sorting, streaming and aggregating *all the data all the time* to get the answer or discover new knowledge from unstructured or structured data sources.
Programming model is declarative (query) or functional with emphasis on ***what*** is being computed versus ***how*** it is computed

Cloud Computing is all about virtualization -- Application access to converged infrastructure (Compute/Network/Storage) via IP Stack
Programming Model is Platform as a Service with APIs for ***what*** is being computed rather then ***where*** the computing is done

Multiple Aspects of Big Data



The Future: Global Memory and Latency Hiding



Data Reuse Near Previous Data Access

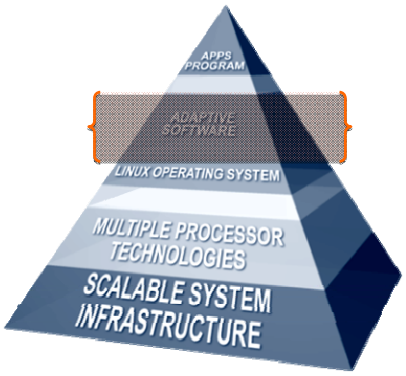
Modeling & Simulation (dense matrices)

Informatics – Analytics (sparse matrices)

Data Reuse over Time

From: Murphy and Kogge, On the Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications, IEEE Trans. On Computers, July 2007

Cray Adaptive Supercomputing
Adapting the system to the application
And not the application to the system



Extending Adaptive Supercomputing to Big Data Workloads

COMPUTE | STORE | ANALYZE

Motivation For "Cascade" XC30

Why are HPC machines unproductive?

- **Difficult to *write* parallel code (for example, MPI)**
  - Major burden for computational languages
- **Lack of programming tools to *understand* program behavior**
  - Conventional models break with scale and complexity
- **Time spent trying to modify code to fit *machine* characteristics**
  - For example, clustered machines have relatively low bandwidth between processors, and cannot directly access global memory
  - Programmers then try hard to reduce communication, resorting to bundling communication up in messages instead of just accessing shared memory

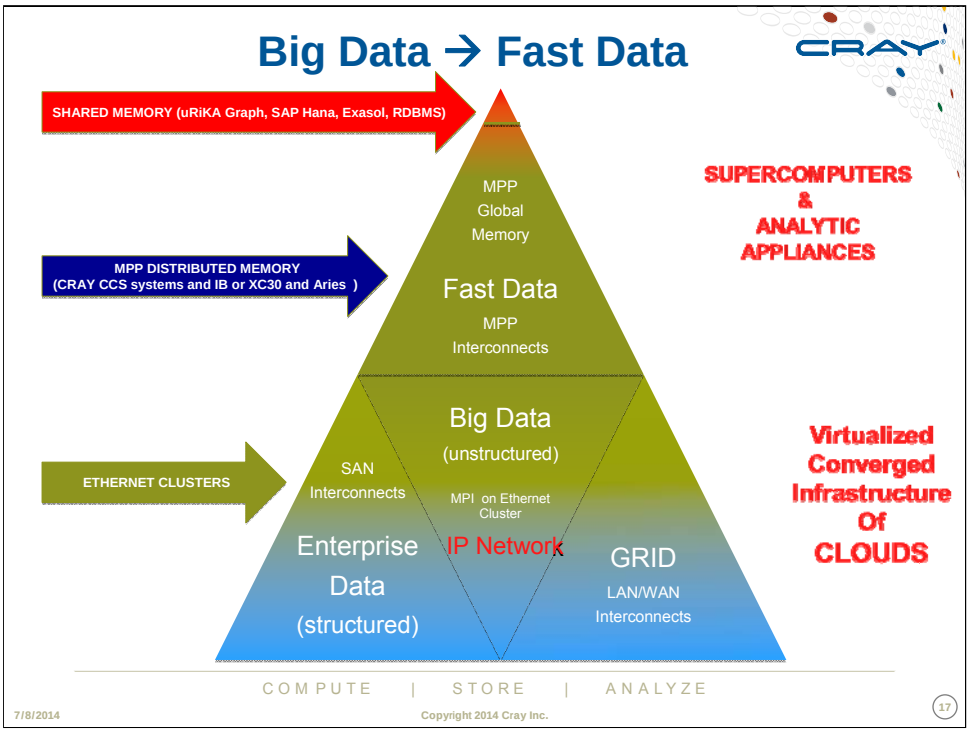
Cray's XC30 system and tools provide the needed help!

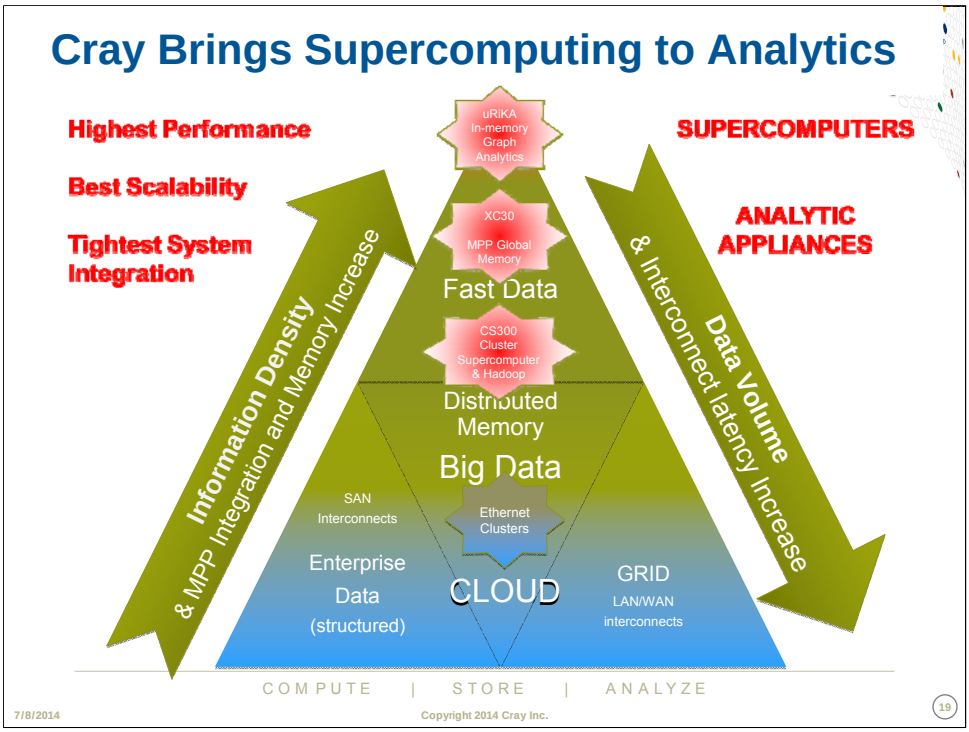
- The Aries network provides hardware assist for MPI operations and atomic Global Memory Operations
- The entire programming tool kit is optimized for parallel programming with runtime analysis allowing best library/kernel to be used dynamically
- Continuing R&D to establish new auto-tuning/optimization approaches

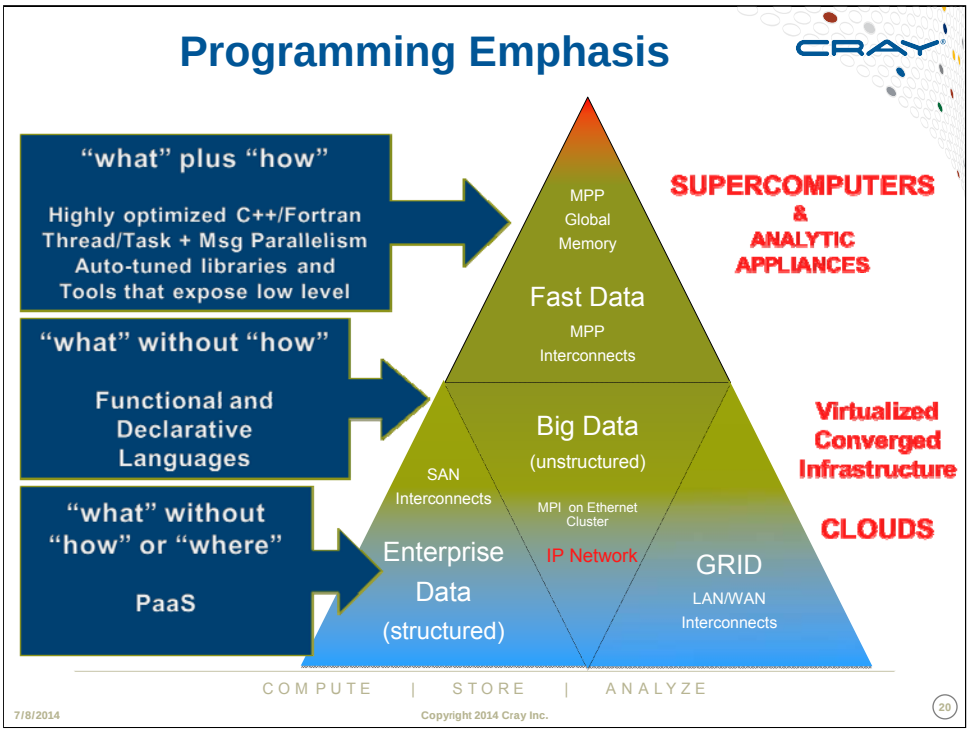
The Cray XC30™ Supercomputer

Continuing the most successful, productive supercomputer product line ever built by Cray

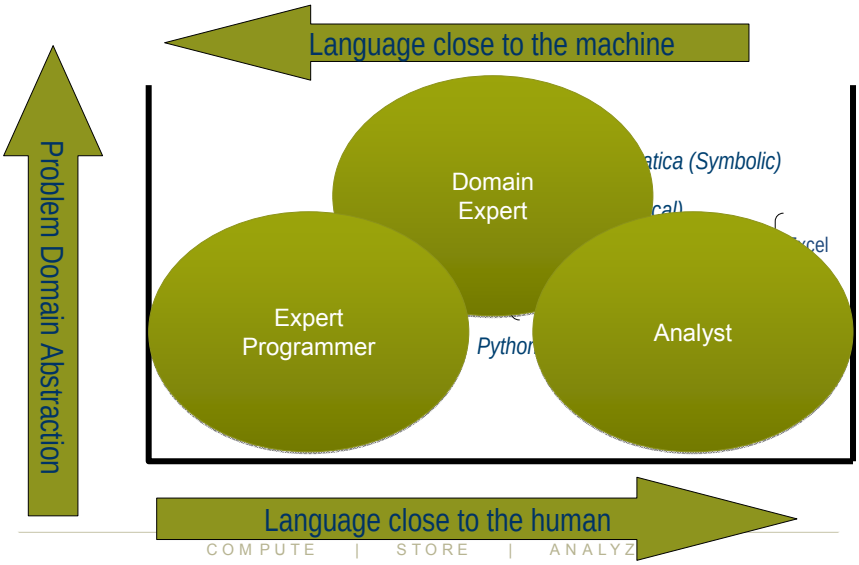
- Previously codenamed “Cascade”
- Completion of six-year U.S. DARPA HPCS contract
- Set the stage for Global Memory Operations at Exascale
- Significant productivity gains for msg passing parallelism
- Productized as the *Cray XC30™ Supercomputer*
- Cray’s latest production supercomputer

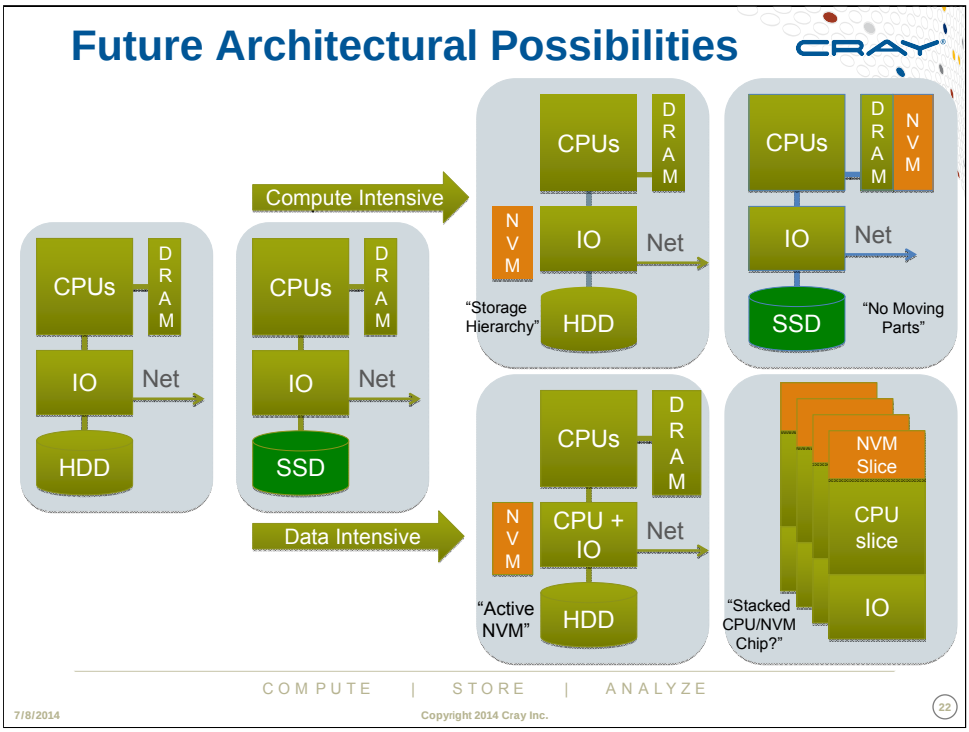

Big Data → Fast Data
SHARED MEMORY (uRiKA Graph, SAP Hana, Exasol, RDBMS)
MPP DISTRIBUTED MEMORY
(CRAY CCS systems and IB or XC30 and Aries)
ETHERNET CLUSTERS
MPP
Global
Memory
Fast Data
MPP
Interconnects
Big Data
(unstructured)
SAN
Interconnects
MPI on Ethernet
Cluster
Enterprise
Data
(structured)
IP Network
GRID
LAN/WAN
Interconnects
SUPERCOMPUTERS
&
ANALYTIC
APPLIANCES
Virtualized
Converged
Infrastructure
Of
CLOUDS
COMPUTE | STORE | ANALYZE
7/8/2014
Copyright 2014 Cray Inc.
17


Cray Brings Supercomputing to Analytics
Highest Performance
Best Scalability
Tightest System Integration
SUPERCOMPUTERS
ANALYTIC APPLIANCES
uRiKA In-memory Graph Analytics
XC30 MPP Global Memory
Fast Data
CS300 Cluster Supercomputer & Hadoop
Distributed Memory
Big Data
Ethernet Clusters
SAN Interconnects
Enterprise Data (structured)
CLOUD
GRID
LAN/WAN interconnects
Information Density & MPP Integration and Memory Increase
Data Volume & Interconnect latency Increase
COMPUTE | STORE | ANALYZE
7/8/2014
Copyright 2014 Cray Inc.
19


Programming Emphasis
"what" plus "how"
Highly optimized C++/Fortran
Thread/Task + Msg Parallelism
Auto-tuned libraries and
Tools that expose low level
"what" without "how"
Functional and
Declarative
Languages
"what" without
"how" or "where"
PaaS
MPP
Global
Memory
Fast Data
MPP
Interconnects
Big Data
(unstructured)
MPI on Ethernet
Cluster
IP Network
SAN
Interconnects
Enterprise
Data
(structured)
GRID
LAN/WAN
Interconnects
SUPERCOMPUTERS
&
ANALYTIC
APPLIANCES
Virtualized
Converged
Infrastructure
CLOUDS
COMPUTE | STORE | ANALYZE
7/8/2014
Copyright 2014 Cray Inc.
20

The Challenge of Expressing Analytics?
Queries and Program Code Do Not Mix Well

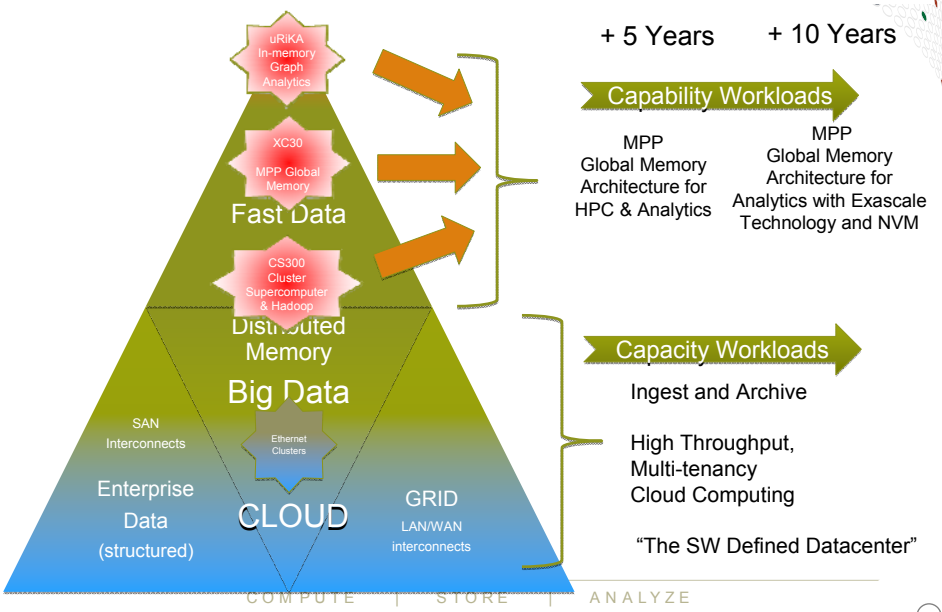



Future Architectural Possibilities
CRAY
CPUs
DRAM
IO
Net
HDD
Compute Intensive
CPUs
DRAM
IO
Net
SSD
Data Intensive
CPUs
DRAM
NVM
IO
Net
HDD
"Storage Hierarchy"
CPUs
DRAM
NVM
IO
Net
SSD
"No Moving Parts"
CPUs
DRAM
NVM
CPU + IO
Net
HDD
"Active NVM"
NVM Slice
CPU slice
IO
"Stacked CPU/NVM Chip?"
COMPUTE | STORE | ANALYZE
7/8/2014
Copyright 2014 Cray Inc.
22

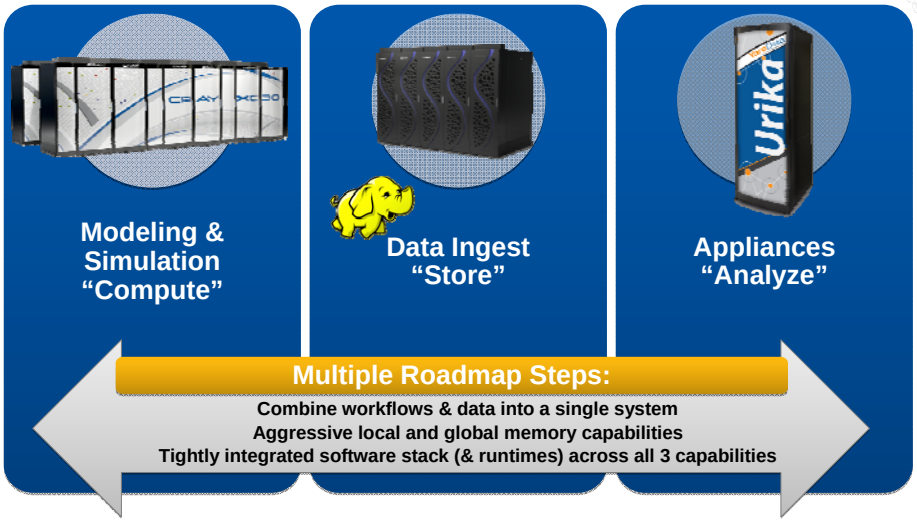
System Implications of Fast, Cheap, Non-Volatile Memory

- **Operating System**
  - **Evolutionary changes** integrated into existing architectures (same file system semantics but with buffering, persistent objects)
  - **Revolutionary changes** with whole-system persistence will effect memory management, I/O, fault management, etc.
  - Relegate magnetic storage to an archival function
- **Power**
  - Today 30% to 40% of system power is DRAM (HDD is 10%)
- **Applications**
  - Manipulate data 100x – 1000x faster (both throughput and latency improve)

The Next Generation of Analytics Will Need Very High Performance Global Memory Operations



Cray's Roadmap "Fusion"



COMPUTE | STORE | ANALYZE

Concluding Comments

- Warehouse scale distributed computing, aka Cloud, provides an excellent multi-tenancy resource for high throughput capacity computing especially where virtualization of converged compute/network/storage affords the use of resources in various locations
- But highly parallel **analytic** workloads, especially those that require low latency messaging and/or global memory operations that benefit greatly from the high performance interconnects and the tight integration of MPP machines, will not migrate from MPP to Cloud
- Many Cloud developments will "condense" into future big memory MPP systems, including programming models, RDF and NoSQL databases, software defined networks and storage, and hypervisors that combined with the high performance message passing and global atomic memory support in MPP networks (e.g., Cray Aries) will best support **the fusion of HPC and large-scale analytics**

Thank You!

bill.blake@cray.com

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products 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.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.