Big Data and SGI

Greg Thorson
Chief Engineer
3-27-2013

Concept of Big Data is Not New

- John Mashey is credited with coining the term "Big Data" in 90's at SGI
  - It is older than that, though
- What is new?
  - All the hype is not coming from status quo

128 Processor Origin at SC96

©2012 SGI

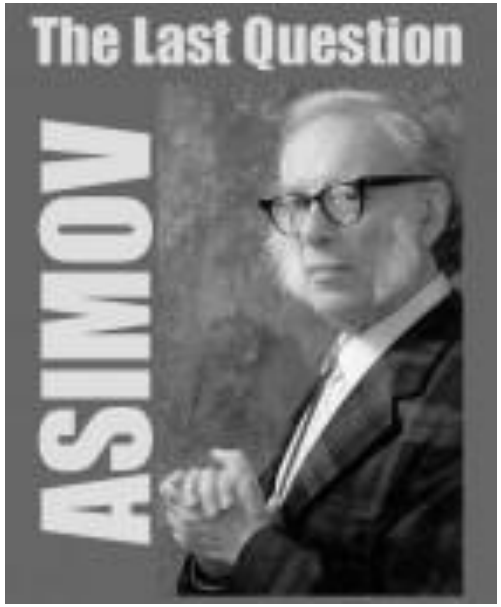
Big Data: What is new?

- Cloud
- Need for next layer of fruit in scientific discovery
- Sources of data
- New technological balance points
- Mobile

©2012 SGI

Big Data: What is new?

- Cloud
  - Even the concept of cloud is not new
    - Isaac Asimov: “The Last Question”, Science Fiction Quarterly, November 1956
  - Cloud is a delivery mechanism, not an architecture
    - Private
    - Virtual Private
    - Public
- Need for next layer of fruit in scientific discovery
- Sources of data
- New technological balance points
- Mobile



©2012 SGI

Big Data: What is new?

- Cloud
- Need for next layer of fruit in scientific discovery
  - Reached limits of I/O
  - Need to make correlations across data
- Sources of data
- New technological balance points
- Mobile

©2012 SGI

Big Data: What is new?

- Cloud
- Need for next layer of fruit in scientific discovery
- Sources of data
  - Computation
  - Sensors
    - Square Kilometer Array
      - O(100) PB/day of data
  - Transactions
  - Archive
- New technological balance points
- Mobile

©2012 SGI

Big Data: What is new?

- Cloud
- Need for next layer of fruit in scientific discovery
- Sources of data
- New technological balance points
  - NVRAM
    - Lower latency
  - GPU and Xeon Phi
    - High mem bw on small set
  - Optical Interconnects
    - Farther reach
- Mobile

©2012 SGI

Big Data: What is new?

- Cloud
- Need for next layer of fruit in scientific discovery
- Sources of data
- New technological balance points
- Mobile
  - Incredible power that is “apparently” in the palm of your hand

©2012 SGI

What Tools are Useful for Big Data?

- Big Data Covers a Spectrum
  - Search
    - MapReduce
    - SGI shipped 10,000s of Hadoop nodes in 2012
  - HPC Coupled
  - Relationships Between Data
  - High Performance Storage
  - Archive

CloudRack™
Tray Architecture

©2012 SGI – Company Confidential

What Tools are Useful for Big Data?

- Big Data Computation Covers a Spectrum
  - Search
  - HPC Coupled
    - Conditioning required before I/O
    - High-performance interconnect
    - Efficient power and cooling
  - Relationships Between Data
  - High Performance Storage
  - Archive

SGI ICE-X
Blade Architecture

©2012 SGI – Company Confidential

What Tools are Useful for Big Data?

- Big Data Covers a Spectrum
  - Search
  - HPC Coupled
  - Relationships Between Data
    - Genomics; de novo was found to be important
    - Python, Java, etc. on big memory or high thread count
    - I don't like the use of the "compute' here as it belittles the dominant value of the connection between the data for many graph problems
    - Latency hiding helps throughput, but has benefit for interactivity
    - UV: 64 TB of coherent shared memory
    - In-memory database
- High Performance Storage
- Archive

SGI UV
Shared-Memory Systems

©2012 SGI – Company Confidential

Finding Threats, Trends Quickly

Culturomics Example from Wikipedia

- Kalev Leetaru
- 42 Million Dates in the Past Millenium
- 80 Million Locations
- ***24 Hours Development Time***

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

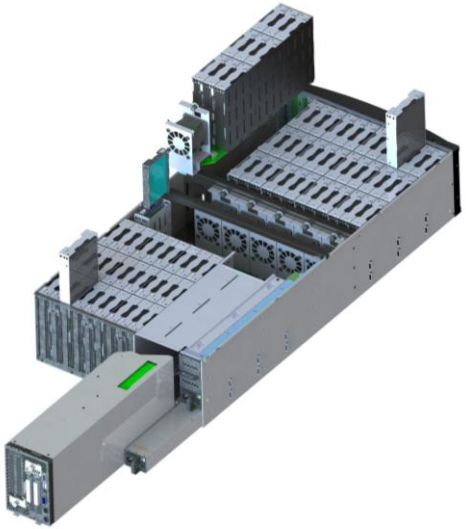
©2012 SGI

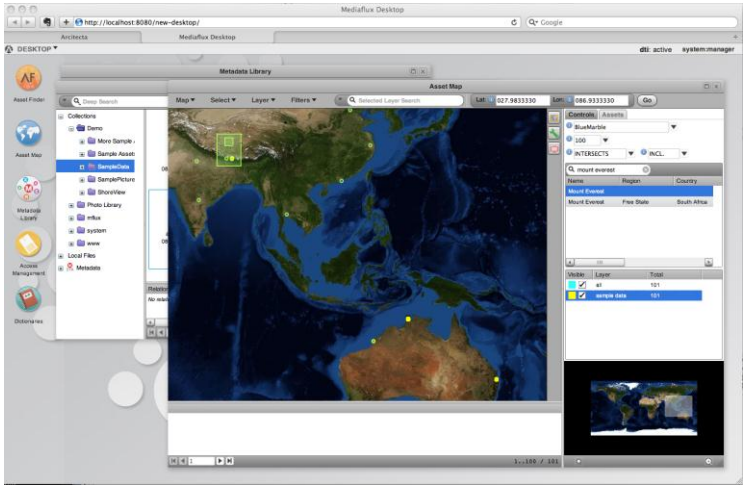
sgi.com/go/wikipedia

sgi

What Tools are Useful for Big Data?

- Big Data Covers a Spectrum
  - Search
  - HPC Coupled
  - Relationships Between Data
  - High Performance Storage
    - SGI shipped 600 PB in 2012
  - Archive
    - LiveArc: Digital Asset Management and Collaborative Knowledge

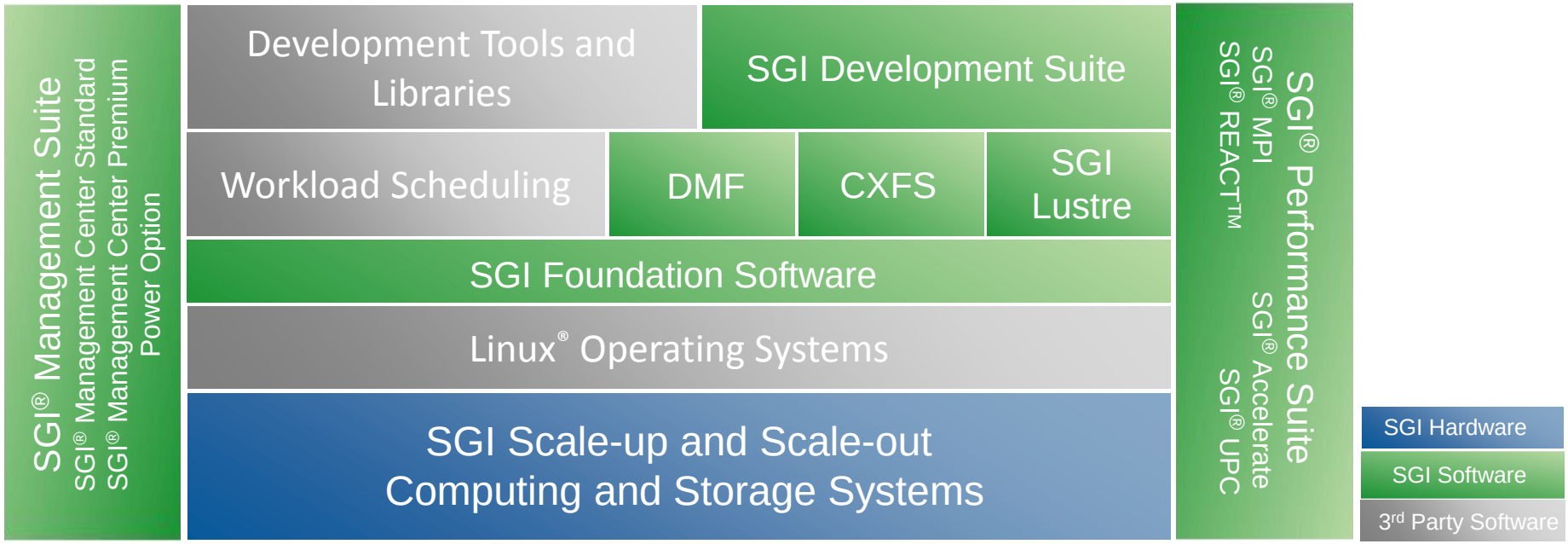




©2012 SGI – Company Confidential

Where Does SGI want to Go?

- We have a great platform stack with a roadmap to make it even better
- We want to collaborate more heavily to get the upper layers added to solve your problems



©2012 SGI – Company Confidential

SGI UV

sgi.com