CRAY®
THE SUPERCOMPUTER COMPANY

The Fusion of Supercomputing and Data Analytics To Drive Scientific Discovery

Bill Blake
SVP & CTO

Big Data and Why We Should Care

Big Data refers to data that is not easily captured, managed and analyzed by traditional tools due to:

- *Volume* (growing > 60%/yr; no sampling/curation)
- *Velocity* (often real time streaming)
- *Variety* (all forms of structured/unstructured data: logs, docs, images)

IDC expects Hadoop to run on over 50% of Big Data Projects over time representing a $8.5B market by 2015

Science will increasingly be data-driven to understand the world

Business will increasingly be data-driven to understand customers

Data-Intensive Processing is driving the need for advanced architectures

Source: Eric Green, Director, National Institute of Health: NextGen 101 Workshop

System Architecture Differences…

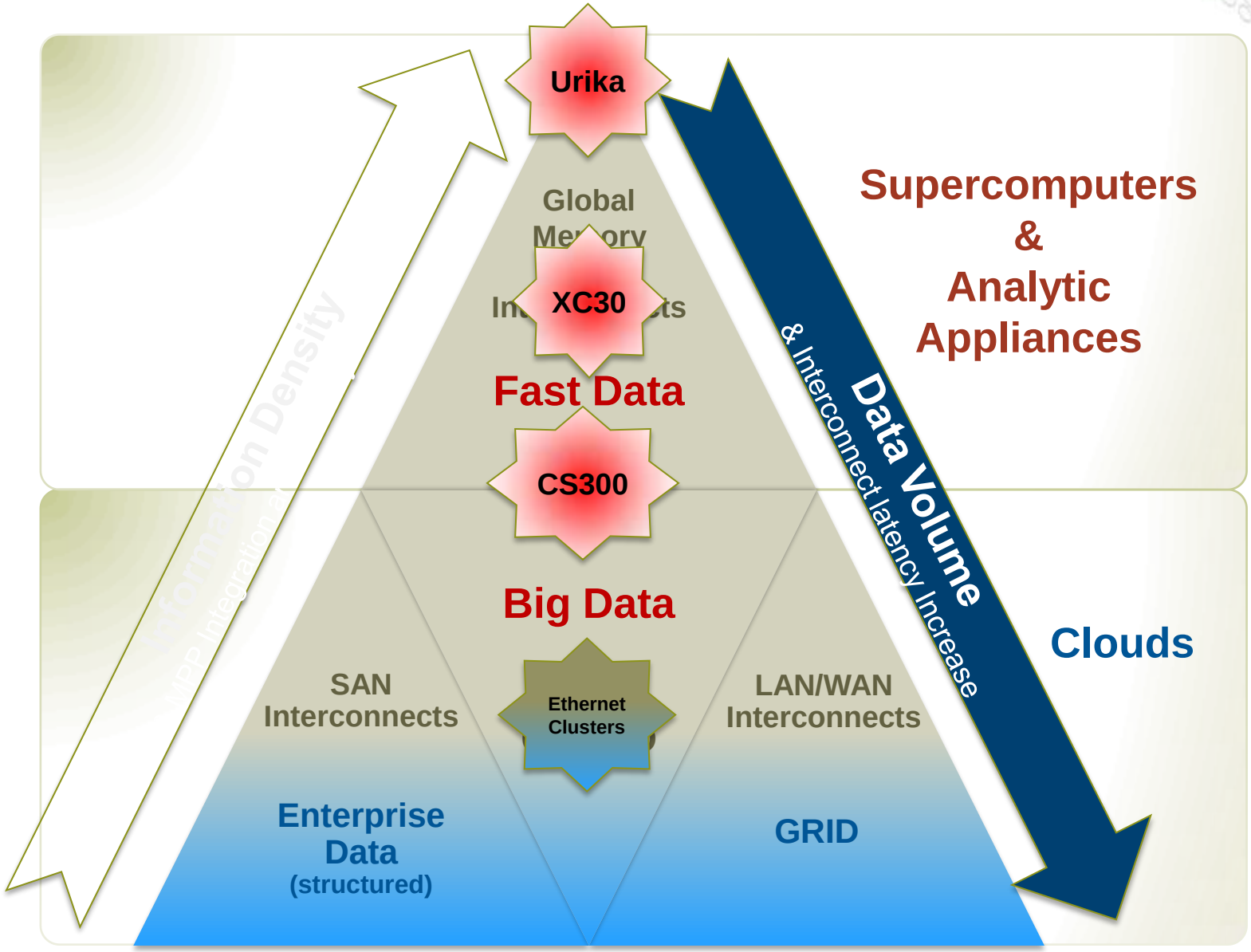
Supercomputing

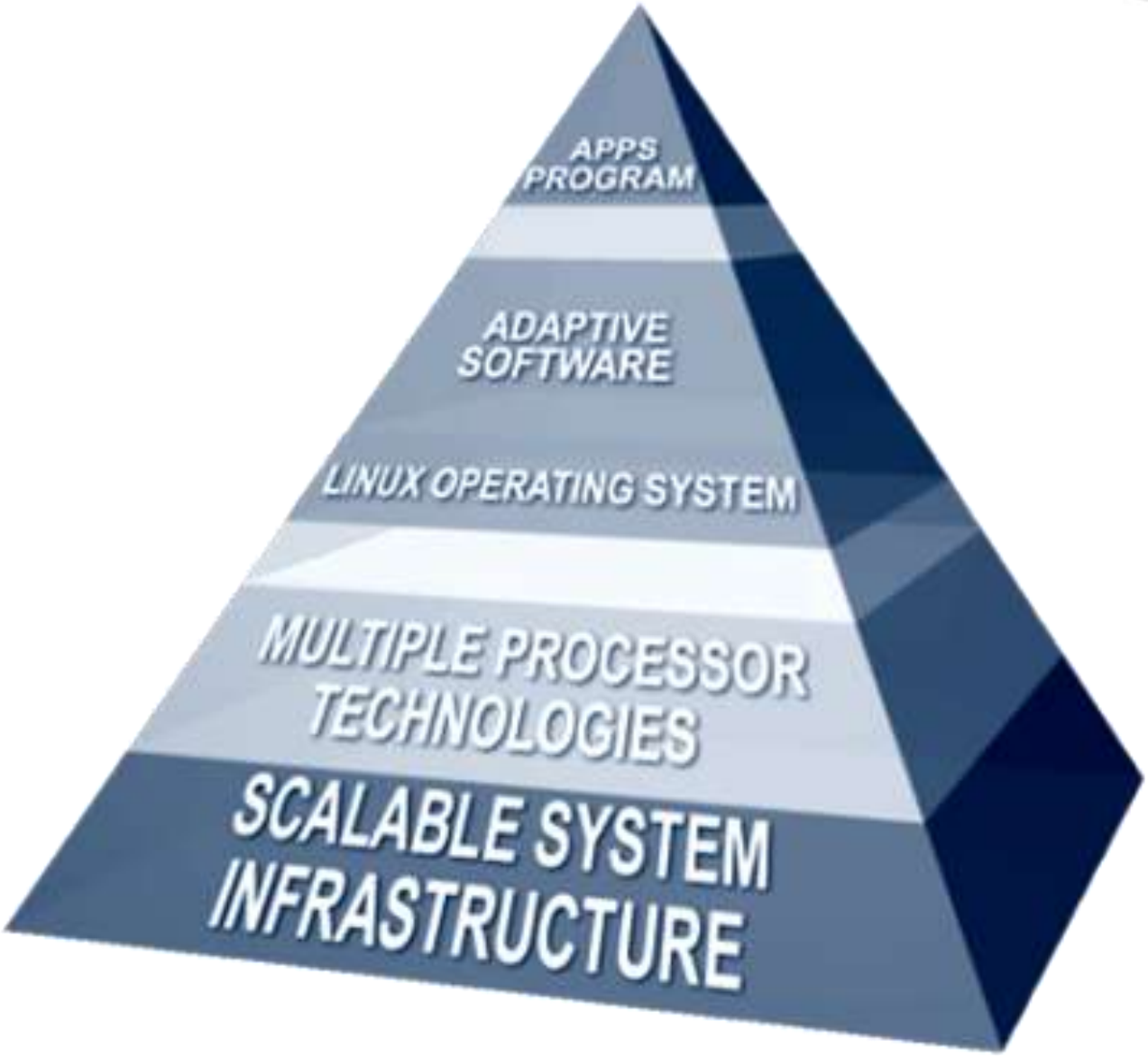
- Scalable computing w/high BW, low-latency, Global Mem Architectures
- Highly integrated processor-memory-interconnect & network storage
- Minimize data movement – load the “mesh” into memory
- Move data for loading, check-pointing or archiving
- “Basketball court sized” systems

Large-scale Data Analytics

- Distributed computing at largest scale
- Divide-and-conquer approaches on Service Orientated Architectures
- Maximize data movement-- Scan/Sort/Stream all the data all the time
- Lowest cost processor-memory-interconnect & local storage
- “Warehouse sized” clouds

Big Data → Fast Data





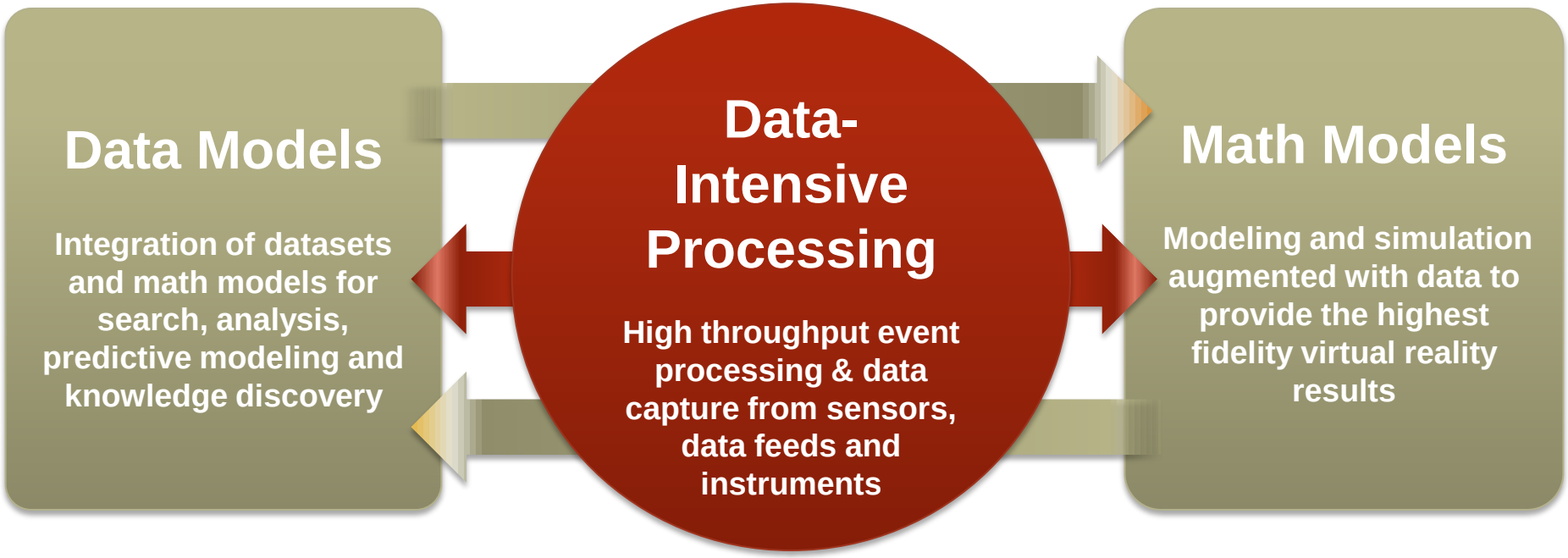
Extending Adaptive Supercomputing to Big Data Workloads

Cray's Vision:

The Fusion of Supercomputing and Big & Fast Data

Modeling The World

Cray Supercomputers solving "grand challenges" in science, engineering and analytics



Advanced Analytic Appliances

Storage & Data Management

Supercomputers

Early MapReduce on Cray Systems

Excellent progress at NERSC providing MapReduce capability on their XE6 systems

- New capabilities for job policies and the run-time environment to support very large numbers of Joint Genome Institute jobs.
- Utilize the Cray Cluster Compatibility Mode (CCM) to support tools like Java and support a throughput oriented scheduling environment
- An excellent real-world example of meeting the needs of the data-intensive community in the world of traditional simulation and modeling

Sandia's Development of MapReduce in MPI (MR-MPI)

- MapReduce functionality implemented in MPI context (no Java)
- MR-MPI library performs data movement between processors and supports and requires local disks for "out of core" large data sets

Active Collaborations underway with DOE labs and NSF centers

Cray Cluster Supercomputers for Hadoop: Purpose-Built, Turnkey Hadoop Solutions

Best Hadoop Distribution

- **Security**– Comprehensive, and fast, encryption
- **Performance** – Faster Hive, Cache acceleration, etc.
- **Management** – Intel Manager for Hadoop Software

Performance of a Cray

- **Proven HPC** – Cray technology & expertise
- **Vast Scale** – Grow to meet any mission requirements
- **Holistic Design** – Balanced compute, networking & storage

Turnkey Solution

- **Reliable** – Rapid ROI… runs as-advertised
- **Support** – One throat to choke, for the whole stack
- **Maintenance** – Update & evolve, without concerns

High Value Hadoop

- **Performance** – Power to accommodate current & future goals
- **Reliability** – Will meet any challenge, without surprises
- **Maintenance** – Easy to maintain & accommodate change

Similar offering in development on XC platform

Thank You

IDC had just finalized a white paper titled ”Cray Cluster Supercomputers Take Aim at Big Computing and Big Data Challenges”

We have a few copies of the paper at the Cray table. Feel free to stop by and get a copy if you are interested.”