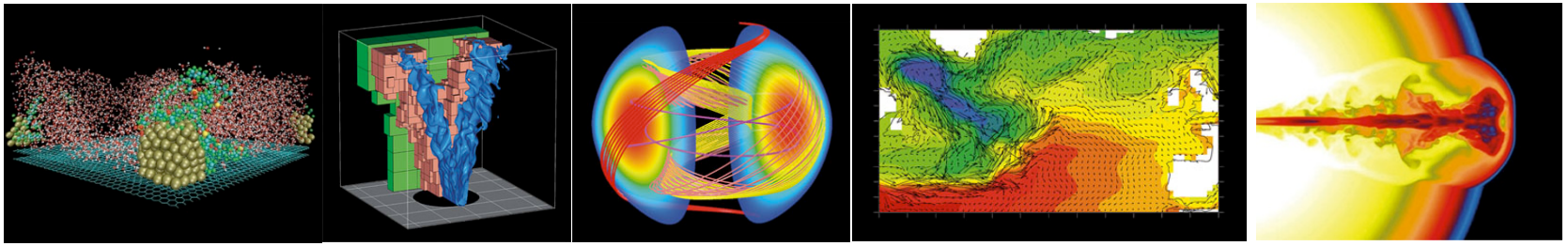

Exa and Yotta Scale Data

William T.C. Kramer

NERSC General Manager

kramer@nersc.gov

510-486-7577

Berkeley National Laboratory

This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE -AC03-76SF00098.

Exa to Yotta Scale Data Panel

- This year we see PetaFlops/s systems in place.
- Enthusiasm for Exascale in 6-10 years – many challenges
- Data scale is 100 to 1,000 times larger than computational scale
  - For Terascale systems many sites have Petascale data stores
- Data is becoming more important
  - More observational data
  - More synergy between observation and simulation
    - Tighter coupling
    - More workflow information
- We will have Exascale data soon and need Zetta scale data for Exascale Flops/s

Exa to Yotta Scale Data Panel

- Garth Gibson – Carnegie Mellon University and and Pansas
- Gary Grider – Los Alamos National Laboratory
- Keith Gray – BP
- Rob Farber – Pacific Northwest National Laboratory

Recent Exascale Report

- DARPA Sponsored
- Exascale Computing Study: Technology Challenges in Achieving Exascale Software
- Many interesting areas - Includes storage and data as well as computation

Extractions

- Studied three sizes of systems
  - Data Center System
  - Departmental Systems
  - Embedded Systems
- Different uses and different ratios
- Three areas of persistent storage
  - Scratch storage
    - Grow 10x to 100x of main memory
  - File storage
    - Grow 1,000x from today's petascale
  - Archival storage
    - > 100x total memory

Extractions

- Four Challenges
  - Energy and Power Challenge
  - Memory and Storage Challenge
  - Concurrency and Locality Challenge
  - Resiliency Challenge
- Cost at exascale may be dominated by data movement
  - $ and Watts

Storage

Today

- Capacity, transfer rate, seek time and power
- Consumer, enterprise, handheld
- Capacity
  - Disk 10x growth over 6 years
  - Archive 1.7-1.9x CAGR per year

Future

- Capacity
  - 10 x growth every 6 years means an Exaflop system needs between 83,000 and 1.3 million drives plus ECC and RAID
- Power
  - 0.8 to 3.8 MW to match exascale (plus RAID and ECC)
- Seek time seems stable

Storage Strawhorse

- Disk 1,000x main memory
- For checkpoint and scratch, drives may be scattered across groups of processors (~16)
  - Minimizes transfer costs
  - Available across interconnect to all processors
    - Implies changing parallel file systems
  - Intermediate levels of storage
    - Latency, cost, capacity between DRAM and disk
- Disk might be 14% of the total power

Exa-Zeta Scale Issues and Question

- Will there need to be PIS (processor in storage) in addition to PIM?
- Will there be any mechanical storage in systems of that time
- Will there be still be the three main classes fo storage - scratch, persistent and archival
- What will file-systems mean in the Yotta Scale time frame
- Will storage be the weakest link of a system (reliability, latency, SW, ...)
- Will Exa-Zetta-Yotta scale data storage SW an HW be evolutionary or does it need to be revolutionary
- Will Raid be enough
- Will the power profile of storage match the power profile of CPUs and memory
- How many levels will there be in the hierarchy of storage
- Will the cloud solve all the Zetta Scale storage issues
- Will tape exist in the storage hierarchy \How much data will we access in order to read a byte? (this is page blocking, etc.)
- Will commercial needs solve the yotta scale storage issues for HPC
- Will finding the data take longer than processing the data?