Petaflops IO

Aug. 15 2005

Lee Ward

Architecture

- MPP machine
- 50,000+ compute nodes at least
  - Maybe 500,000 for a dense implementation
- 1,000+ IO nodes at least
- Memory sizes vary
  - 1 GB on PIM or System on Chip
  - 64 GB for more classic nodes
- High speed network
  - .25 – 2 µs latency
  - 15 – 40 GB/s bandwidth
  - Clos or fat-tree

Implications

- **Architectural performance disconnects increase**
  - **Disk is the same old technology**
  - **CPU-memory is a little worse**
  - **Network-Storage is a lot worse**
    - **Storage latency is nearly the same as today**
- **Light weight operating systems**
  - **IBM BG and SUNMOS/Puma/Catamount supplanted by Linux**
    - **Linux real-time support has improved**
    - **Linux has robust deadline schedulers**
  - *Device interrupts are still not well tolerated*
- *Code base is same but application differs*

Sandia National Laboratories

User Interface

- **More naturally supports efficient parallel IO**
  - **Reference to Tom Ruwart's report on POSIX efforts**
- **Heavy leverage of single-sided comms in infrastructure software**
  - **Leased locks are impossible**
    - **Reliance on the timely reception and action based on callback software architecture is a non-starter**

The Tri-lab

- **Our problems remain the same**
  - **Energy, shock, stress, flow**
  - **All requiring the same tightly coupled solutions**
- **We, and industry, deploy highly integrated file system solutions**
  - **A common storage system from the desktop to the premier supercomputer**
  - **With fast store, backup, HSM, and archive serviced**
    - **But it's young and we'll have operational difficulties**

Enabling R&D Thoughts

- Active disk
  - With sandboxes and well separated and defined protection domains
  - On disk μ-schedulers
  - A standard interface for depositing applets on the disk and ties to the OS for managing same
- A new, persistent, storage technology
  - For the file system journal at least
- MPI middleware w/o collective IO ops
  - Too many interfaces and caveats for efficient exploitation by mortals
  - Can indpendent ops do double duty?

Sandia National Laboratories