SAND2007-7338C

Scalable IO Requirements at Petascale

SC'07

Lee Ward
Principal member of Technical Staff

The acknowledgement statement <u>MUST</u> be used on the title slide of all presentation material distributed outside of Sandia.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Application Example

- Physics and engineering numerical simulation codes
- A problem is partitioned into sub-problems with inter-related parts
- Inter-related parts *must* communicate, frequently
- At certain points, when things reach some sort of equilibrium, an application might defend against a machine interrupt or fault
  - By writing critical information to restart files
  - Ok, by writing ***a lot*** of information to restart files

Feeds

- Amount of data written is some fraction of memory
  - Let's suppose 10% for a ballpark
- Extrapolate from Red Storm; 104 TF, 12,500 nodes, 4 GB per node
  - That's 50 TB of RAM, so a 5 TB restart file
- Apps keep more than the most recent; Maybe a large fraction if they need steering
- 120 TB of disk on the machine (which is tight for us)
  - That is a whopping 24 restart files!
- A 1 PF machine needs 1.2 PB of disk
  - It's 10X faster than Red Storm with 10X the memory
  - Still only 24 restart files
- Which is naïve but gives the right feel

Speeds

- Apps want to spend less than ~10% of their time writing restart dumps
- Same Red Storm, benchmarks say 40 to 50 GB/s
  - A "good" app can actually realize 12 GB/s
- Must scale up to keep the same app spending less than 10% of time writing dumps
  - Pray the app and the file system scale linearly
- We'll need to benchmark 400 to 500 GB/s for our 1 PF machine
  - 10X more data, remember, means we need to supply a 10X faster IO system

Are you Impressed?

- Me, I'm floored
  - Those were optimistic numbers
  - File Systems don't scale linearly
    - It costs to do more coordination and the number of components must increase to supply all of this
  - But apps don't scale linearly but it doesn't help
    - It costs them to coordinate as well
      - That does offset the higher overhead in the FS?
    - Enough? Almost certainly not. Developers work hard to negate the increased cost on a larger machine.

Conclusion

- A simple, napkin, extrapolation of Red Storm to 1.04 PF means
  - 1.2 PB of spinning media
    - Certainly too low; Double it?
  - 400 – 500 GB/s of measured bandwidth, writing
    - Probably high
    - Allowing for 100% error, it’s still hundreds of RAIDs
      - Which gives us an annoying management problem
- But 1.04 PF is only the beginning
  - Peta*scale* is 1 – 1000 PF
  - Many folks are throwing around 10 PF as a starter machine in the range; Uh-oh
- It seems we’re in for some interesting times
  - Which, oddly, doesn’t scare me but I’m not normal ☺