File System Monitoring Task Group Status and Discussion

Andrew Uselton, Intel

April 10, 2014

Members

- Liam Forbes, UA Fairbanks
- Andrew Uselton, Intel
- Jeff Layton, Intel
- Ben Evans, Terascala
- Mark Nelson, Inktank
- Alan Wild, Exxon-Mobil
- Jeff Garlough, Cray Inc.
- Cheng Shao, Xyratex

Committee Tasks

- Develop a list of existing parallel filesystem monitoring tools
- Identify tools' capabilities, identify others we think should exist
- Compare and contrast the tools to each other and the capabilities we think should exist.
- See http://wiki.opensfs.org/BWG_File_System_Monitoring for up to date progress and participate.

The Problem

- Instrumenting a new file system
  - to detect if a component has failed
  - to ensure you are meeting your target performance numbers
  - to determine what future improvements can be made.
- Collect what information? How often?
- What tools/utilities/commands?

The Features – what to monitor.

- Per-LUN metrics
  - read/write IOPS
  - read/write rate
  - bandwidth (peak rate)
  - rebuild/verify statistics
  - device utilization (%)
  - average utilization
  - aggregate data rates
- Metadata
  - ops
  - queue depth
  - latency
- RPCs
- Aggregate metrics
- Component status
  - disk
  - controller
  - switch
  - servers – CPU sys, io_wait
- Network status (link health)
  - server to client
  - Server to controllers
  - Controller to disk

This is not an exhaustive list

The Tools– how to monitor.

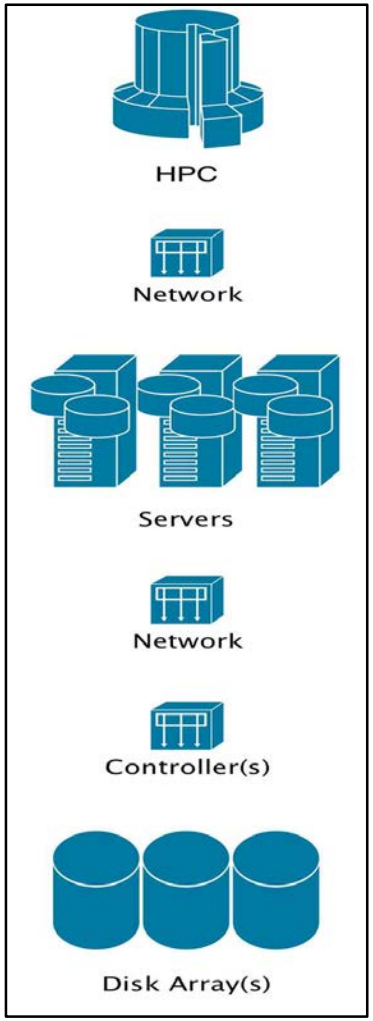
- Vendor tools
- LMT/Cerebro
- Chroma
- collectl and ganglia
- collectd and graphite
- blktrace
- perf
- sysprof
- lltop and xltop
- iostat
- sar
- Alan Wild's perl script
- Nick Cardo's MPI utilities

This is not an exhaustive list

An I/O Process Model

- There are many tools.
- We solicit more.
- There are many metrics.

A model of the I/O process would help identify the sources and meaning of any such data.



Thank you

Questions?