LAYOUTGET at flush
IETF 81 NFSv4 WG Meeting, July 25, 2011

Sorin Faibish – EMC (sfaibish@emc.com)

Problem Statement

- Move the layoutget for write from open to flush time will be a performance bottleneck
- Give the MDS server to optimize by over-allocation unilaterally
- Protocol allows a maximum layout size but it is not used
- Will affect all layouts not only block; there is no way for the MDS to know the intention of application and it cannot deliver entire layout for new files.
- Current Linux cache writeback not optimal; cannot depend on it
- Not clear the impact on sparse files

Affect All Layouts

- pNFS server has no information on application
- Allocating more than clients ask will impact performance
- Will affect multiple clients scalability and performance
- Flushing large amounts of data is a problem for performance
- Server has no idea of the amount to be flushed as there is no application data;
- Even file layout cannot send the entire layout as it doesn't know how much to allocate
- Ask for a larger layout than needed - wasteful

Solution?

- Is this a protocol issue we need to resolve or just optimization?
- What action to take:
  - Add some text in 5661 (errata)
  - Use recommendations for implementation
  - Leave this to the servers to optimize/decide
  - Add errata to 5663 only