"FAST" Futures

HEC-IWG File Systems & I/O Research Guidance Workshop

ps. FAST == File and Storage Technologies

FAST05 has 125 submissions for 24 slots :-) but not many about HPC :-(

Garth Gibson
Carnegie Mellon University & Panasas CTO
garth@{cs.cmu.edu, panasas.com}

August 16, 2005

What research is Panasas supporting?

- Two forms of Panasas support for research: time and money
- Time
  - Object Storage Device (OSD) interface (version 2: recovery, mgmt, RAID)
  - Parallel NFS extension to NFSv4 protocol, pNFS client driver SW structure
    - Vehicle for getting a high performance parallel file system standard
  - Well defined POSIX extensions for high performance file systems
- Money
  - Broad support for CMU PDL agenda, especially pNFS client APIs & better understanding of what is hard about configuring & managing systems
- Adequacy of existing support
  - OSD v2 is mostly adequate, except for using “collections” in FSs and apps
  - The rest is underfunded, esp. choices in pNFS protocols & implementations

Areas needing more attention

- SCALE -- how does scaling impact FAST HPC
  - Reliability/availability degrades with scale
    - Performance during repair becomes a much more common case
  - Good metadata performance gets more complex with scale
    - Namespace implementation techniques increasingly deferred/reordered/predicted
  - Configuration mistakes/mistuning bigger impact with scale
- SHARING -- how can one size fit all
  - Multi-cluster & fault tolerance means routing failover, which tends to not give sufficiently continuous high bisection bandwidth
  - Security infrastructure configuration is currently a black art at best
  - Performance impact of one cluster's workload on another cluster's workload is not always linear, not adequately predictable
  - Concurrent write sharing without serializability overhead

Areas needing more attention II

- SCALE II -- increasing scale works against rapid deployment
  - Protocol corner cases get heavily exercised (callback/recovery storms)
  - Installations much larger than testable in lab, frequently behind the wall
  - More inflight state means more complex inflight interactions
  - Requirement for zero loss not being relaxed
    - It'd be nice of FS could issue signal to cluster: "Just dumped last hour's changes, please compensate", but its probably not going to happen
  - Need tools and implementation techniques to
    - Verify protocols statically; and/or verify state transitions against models dynamically
    - Simulate larger systems; fault inject into more complex corner cases
    - Diagnose rare problems on the first occurrence with little access to system
- SCALE III -- navigating namespace scaling in size
  - Simple "ls" file tree insufficient & generalized search is only part of answer
  - Need richer schemes to answer "so what have I got & where is what I want"

What research areas are overworked?

- Peer to peer filesystem
  - Nearly every networking academic has invented a P2P file system as the “killer app” for user level distributed systems toolkits
    - Napster and desktop backup
- Fixed content indexing tricks
  - Nearly every search-head has invented a crawler, attributes database and duplicates elimination system for write once data
  - I’m probably wrong on this, in that search is far from a mature topic area, but it sure feels like we don’t know the right questions to ask

What research areas are overlooked?

- Big move to server virtualization just starting
  - Higher HW utilization for multi-CPU servers
  - Deferring OS choice to user
  - “Process migration” via virtual machines for load balancing
  - VMWare/EMC
  - Xen open source and support company
  - Intel/AMD HW mods
- Server virtualization impact on file and storage?
  - VMWare Vmotion cache coherent block storage to fault VM’s VM to new node

Basic Challenge for HPC

- HPC requirements always extreme
  - Basic mission in HPC is bigger, faster, safer large scale computers
  - Too often beyond the 3-sigma point on the normal curve
  - Too costly to develop & maintain competitive offerings for 3-sigma market
  - Markets big enough to fund complete solutions need much less today
  - This is the HPC Market Gap
- Constant need to attend to HPC market gap -- two broad/different paths
  - Acknowledge and live beyond the pale
    - Simplify! Creating and enlarging HPC-only solutions -- open source
  - Walk the tightroad of guessing right (defining) the big markets' roadmaps
    - Pull standards & compromise: POSIX, MPI-IO, OSD, NFSv4, pNFS, iSER, ….

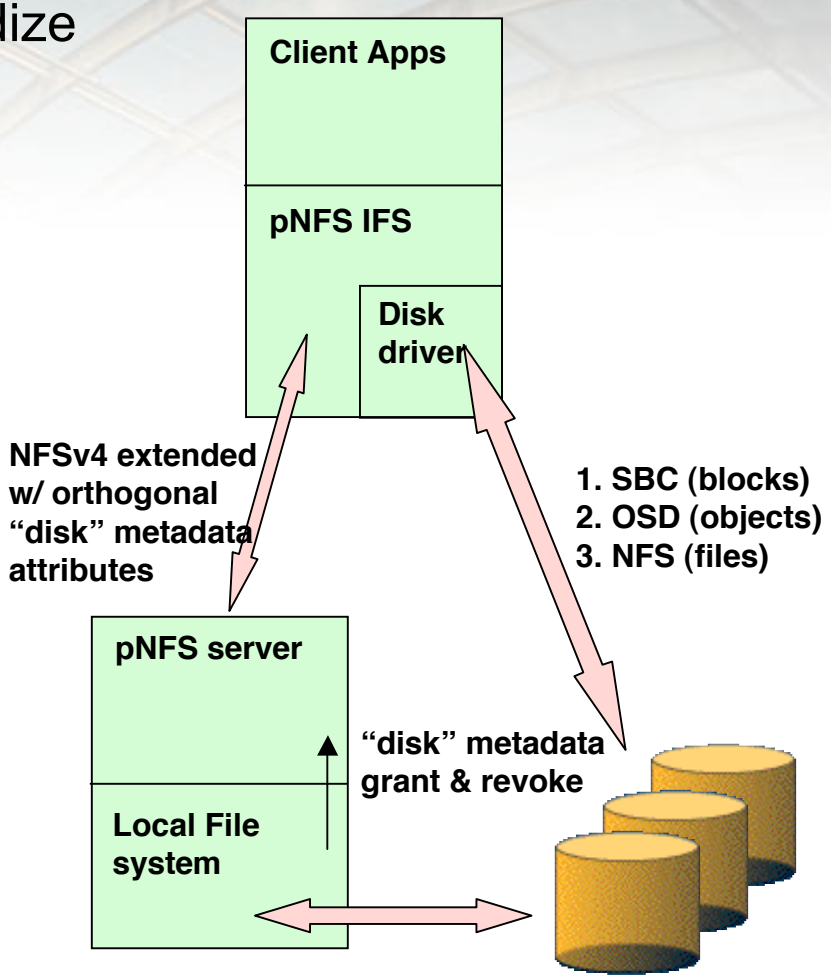
Our Future is FAST!

Garth Gibson
garth@{cs.cmu.edu,panasas.com}

August 16, 2005

pNFS Extensions to NFSv4

- Endow NFS with LAYOUTs to standardize high-bandwidth/parallel file systems
- Inclusiveness favors success
- 3+ flavors of out-of-band LAYOUTs:
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI
  - OBJECTS: OSD/iSCSI/TCP/IP/GE
  - FILES: NFS/ONCRPC/TCP/IP/GE



IETF pNFS Documents:

draft-gibson-pnfs-problem-statement-01.txt

draft-welch-pnfs-ops-03.txt

draft-zelenka-pnfs-obj-01.txt

draft-black-pnfs-block-01.txt