Breakout 3B
I/O Architectures (Intelligence near the storage device (Active Disk), I/O nodes participation in I/O, Network role in I/O, etc.

Session Coordinators: Bancroft

Session Scribes: Miller/Grider

Session Presenter: Bancroft

Session Writeup:

Each Breakout session will provide

1) Current high level topics of File Systems and I/O Research in this area

2) Areas that need to have more research focus

3) Areas that have or will have too much research focus

4) Some rough consensus ranking of areas that need more focus,
less focus,
and overall recommendations including
Short term research needs
Long term research needs

There will be a presentation of this material for each session done by the session leader and a write up for inclusion in the workshop documentation.

Current high level topics of File Systems and I/O Research in this area - from Gary Grider (paper?)

- *Early utilization of intelligence near the disk drive*
  - *Security*
  - *Allocation/mgmt*
  - *Archive*

Current high level topics of File Systems and I/O Research in this area - from Q&A

- David Du: SIMON: Simulation and Modeling for SAN - funded by ONR - This project is to create user-friendly interactive design tools for infrastructure designers to plan and design storage systems in data centers with a guaranteed performance from a set of remote users.

Current high level topics of File Systems and I/O Research in this area - from Q&A

- Bradley Kuzmaul: We are investigating unbounded transactional memory as an alternative to locking protocols to achieve correct concurrent execution of multithreaded programs. In particular, we are investigating a non-preemptive programming model in which the programmer needs not think about the concurrency unless s/he explicitly yields control.

Areas that need to have more research focus (designate short and long term) - from paper

- Work in conjunction with software stack for application
- Export geometry info up to the stack to be exploited
- Take advantage of common placement in the stack
- Archive/Backup participation
- *Take advantage of I/O nodes for more than just connectivity routing and remote execution, perhaps caching, RDMA concepts, etc.*
- *RDMA protocols to allow for scheduled RDMA transfer between client and storage device (AKA ‘‘third party transfer’’)*
- *Help with QOS (end-to-end concept)*

Areas that need to have more research focus (designate short and long term) - from Q&A

- Alok:Runtime optimizations; Architectural acceleration and active storage
- John Howard: Overlay networks based on Distributed Hash Tables
- David Du: OSD Reference Implementation
- David Du: Object Placement for Parallel Tertiary Storage Systems
- David Du: High Performance Tape File System for Backup and Archive
- David Du: QoS Provisioning for OSD-based Storage Systems

Areas that need to have more research focus (designate short and long term) - from Q&A

- David Du: Execution Environment for Active Data Objects
- David Du: Design A Global Storage System based on OSD
- David Du: Realizing Data Provenance from Models to Storage
- David Du: Search and Indexing for Intelligent Storage

Areas that need to have more research focus (designate short and long term) - from Q&A

- Michael Agostino: SAN-wide storage virtualization
- Michael Agostino: SAN-wide shared file system
- Michael Agostino: storage management in the presence of server virtualization

Areas that need to have more research focus (designate short and long term) - from Q&A

- Peter Corbett: pNFS,
- Peter Corbett: Linux NFS client and server
- Peter Corbett: indexing,
- Peter Corbett: high performance file systems
- Peter Corbett: parallel file systems
- Peter Corbett: RDMA,
- Peter Corbett: NFS-RDMA

Areas that need to have more research focus (designate short and long term) - from Q&A

- Peter Corbett: Xen
- Peter Corbett: versioning file systems
- Jim Hughes: 10GB data channel research

Areas that need to have more research focus (designate short and long term) - from Q&A

- Brent Welsh: Comments: Good solutions are ***end-to-end***, yet often individuals (or their projects) focus on particular layers. The result is that they either try to solve too much of the problem within their layer, or they completely ignore some crucial aspect of the problem that is impacted by their layer,yet logically outside their layer.

Areas that need to have more research focus (designate short and long term) - from Q&A

- Maurice Askinazi: proper configuration of network. as i've brought in vendors to evaluate their high performance storage solutions, i increasingly have problems with my net admin. he seems to believe any port on the network is the same as any other. with small interconnects between switches, this obviously isnt true. we need to get unblocked access from the new faster storage to the greedy processor nodes. after observing this problem several times, i think i can make some good suggestions as to changes in network topology that would greatly improve the situation, but maybe this should be locked down with some good rules of thumb that can be pushed to the net admins for official consideration.

Areas that need to have more research focus (designate short and long term) - from Q&A

- Rajeev Thakur: "Robustification" of I/O software. A lot of it is good enough for writing papers, but can't be used by anyone else.
- Rajeev Thakur: End-to-end performance (what applications really see); Performance across all levels of the I/O software stack; Right interfaces at all levels; Educating application programmers and library writers on what they need to do to achieve high performance.

Areas that need to have more research focus (designate short and long term) - from Q&A

- Tyce McLarty: Comments: I think ***caching*** at the file system level across a cluster of machines offers a big potential to reduce the latency seen by applcations, and at the same time greatly increasing efficiency of disk-IO by doing only large, well formed transfers. This is really just a logical extension of the two-phase-IO strategy used in Romio. I still think MEMS data storage would be the ideal hardware to implement this, but flash memory is getting bigger, faster, & cheaper so it might work almost as well.

Areas that need to have more research focus (designate short and long term) - from Q&A

- Mike Folk: Comments: We have learned that the diversity of applications, combined with then complexity of the software stack, make it impossible to create systems that work well in all situations. Systems needs to be tuned at many layers to work well for a given application. R&D is needed that will address this need, both by improving the layers of the software stack and their interoperation, or by improving our ability to tune systems. We have learned that the diversity of applications, combined with then complexity of the software stack, make it impossible to create systems that work well in all situations. Systems needs to be tuned at many layers to work well for a given application. R&D is needed that will address this need, both by improving the layers of the software stack and their interoperation, or by improving our ability to tune systems.
- ***I.e. Research in tuning entire stack? Research in distributing the architecture within the stack layers? Both?***

Areas that need to have more research focus (designate short and long term) - from Q&A

- Pete Wyckoff: Proprietary solutions. They always die and leave us stranded. Don't accept or fund anything that is not an open implementation.
- ***I.e. Research should focus on standards compliance or new standards or open source? All research funding going only to LINUX might be a concern among some mission partners…. But acceptable to others. General open source…***

Areas that need to have more research focus (designate short and long term) - from Q&A

- Pete Wyckoff: ***API elements***. There are quite a number of distinct APIs that span the range between applications and hardware, for transport, block and file data access, application-oriented middleware, etc. None of them is a perfect fit for all situations, but each has its own good aspects: parallelism, locking, object-level access, security, ... Are there ways to think about ***combining chunks of protocols at runtime*** rather than having to switch to a completely different API depending on the feature set? It's hard enough to convince people that IO parallelism is good that I'd prefer not to have to fight SRP vs iSER or DAPL vs RNIC-PI too.

Areas that need to have more research focus (designate short and long term) - from Q&A

- Pete Wyckoff: 1) Transport protocol integration for distributed filesystems: use standards (iSER on iWarp or IB) to move data in the context of a parallel file system (PVFS2). What changes/extensions would be required to the standards to handle the functionality we provide in our home-grown protocol?

Areas that need to have more research focus (designate short and long term) - from Q&A

- Henry Newman: Trying to fix the current state of block devices and shared file systems for large storage systems has received too much attention and has little chance of being affective. The whole concept of I/O and life cycle data management needs to be re-thought with an ***emphasis on looking for techniques used for hardware multithreading that can be applied to I/O***

Areas that need to have more research focus (designate short and long term) - from Q&A

- Henry Newman: 2) Lifecycle data management needs to be integrated into large systems. Along with this new management schemes need to be developed to understand ***who is using what resources when***, and ***metadata management of the data in question*** by both the user and site management.

Areas that need to have more research focus (designate short and long term) - other input

- Better support of metrics (other sessions described this support as needing to be end-to-end in the software stack)
- Role of “intelligent” devices in optimum scheduling across enterprise - an architecture and middleware topic?

Areas that dont have enough research focus (designate short and long term)

1. Research into redistribution of intelligence/responsibilities to various parts of I/O
2. Coordination of information/policies between levels (and caching)
3. QOS in this the world of multiple layers of intelligence (end to end), resource reservation and sharing, dealing with failure, including working with network and other QOS (end to end)
4. Performance diagnosis across layers/modeling
5. Assuming we decide to do this multi-layer intelligence, what abstractions are need to be provided and at what layers, locality, sequentiallity, hints, standards, geometries etc.
6. Exploiting this architecture for Policy functions such as backup/archive/availability
7. A reasoning framework for performance versus portability or other softer topics (perhaps useful in general, not just about this topic)
8. Does it make sense to do research into virtualization that allows for application specific stacks
9. Architectural support for tools that allow debugging
10. Scaling issues for reliability/availability at scale in architectures

Some rough consensus ranking of areas that need more focus, less focus and overall recommendations including Short/Long term research needs

1. Research into redistribution of intelligence/responsibilities to various parts of I/O
    1. Total 73 Government 20 med-long term
2. Coordination of information/policies between levels (includes caching)
    1. Total 37 Government 7 med-long term
3. Assuming we decide to do this multi-layer intelligence, what abstractions are need to be provided and at what layers, locality, sequentiallity, hints, standards, geometries etc.
    1. Total 29 Government 8 long term
4. Performance diagnosis across layers/modeling
    1. Total 29 Government 6 med-long term
5. Scaling issues for reliability/availability at scale in architectures
    1. Total 24 Government 8 med-long term
6. QOS in this the world of multiple layers of intelligence (end to end), resource reservation and sharing, dealing with failure, including working with network and other QOS (end to end)
    1. Total 22 Government 7 med-long term
7. Does it make sense to do research into virtualization that allows for application specific stacks
    1. Total 26 Government 4 med-long term